New Journal of Physics
The open access journal at the forefront of physics
Deutsche Physikalische Gesellschaft DPG | IOP Institute of Physics

A stochastic approach for quantifying immigrant integration: the Spanish test case

Elena Agliari[1,2], Adriano Barra[1,2], Pierluigi Contucci[3], Richard Sandell[4] and Cecilia Vernia[5]

[1] Dipartimento di Fisica, Sapienza Università di Roma, Piazzale Aldo Moro 2, I-00185, Roma, Italy
[2] INdAM, Gruppo Collegato dell'Università di Roma 'Tor Vergata', Dipartimento di Matematica, via della Ricerca Scientifica 1, I-00133, Roma, Italy
[3] Dipartimento di Matematica, Università degli Studi di Bologna, piazza Porta San Donato 2, I-00124, Bologna, Italy
[4] Departamento de Ciencias Sociales, Universidad de Carlos III de Madrid, Avenida de la Universidad 30, E-28911, Madrid, Spain
[5] Dipartimento di scienze fisiche informatiche e matematiche, Università di Modena e Reggio Emilia, Via G. Campi 213/B, I-41125, Modena, Italy
E-mail: adriano.barra@roma1.infn.it

Received 28 May 2014, revised 14 July 2014
Accepted for publication 21 August 2014
Published 24 October 2014

New Journal of Physics **16** (2014) 103034

doi:10.1088/1367-2630/16/10/103034

Abstract

We apply stochastic process theory to the analysis of immigrant integration. Using a unique and detailed data set from Spain, we study the relationship between local immigrant density and two social and two economic immigration quantifiers for the period 1999–2010. As opposed to the classic time-series approach, by letting immigrant density play the role of 'time' and the quantifier the role of 'space,' it becomes possible to analyse the behavior of the quantifiers by means of continuous time random walks. Two classes of results are then obtained. First, we show that social integration quantifiers evolve following diffusion law, while the evolution of economic quantifiers exhibits ballistic dynamics. Second, we make predictions of best- and worst-case scenarios taking into account large local fluctuations. Our stochastic process approach to integration lends itself to interesting forecasting scenarios which, in the hands of policy makers, have the potential to improve political responses to integration problems. For instance, estimating the standard first-passage time and

maximum-span walk reveals local differences in integration performance for different immigration scenarios. Thus, by recognizing the importance of local fluctuations around national means, this research constitutes an important tool to assess the impact of immigration phenomena on municipal budgets and to set up solid multi-ethnic plans at the municipal level as immigration pressures build.

Keywords: continuous time random walks, quantitative sociology, immigration theories

1. Introduction

A particular political challenge of growing immigration is immigrant integration. It is considered a necessity for minimizing frictions and confrontation between immigrants and natives in the host community, as well as a precondition for a competitive and sustainable economy [1]. In response to the recent rapid growth in the number of immigrants throughout many major regions in the world, the need for political intervention targeting integration has become increasingly urgent [2]. Still, effective policymaking in this area is obstructed by the lack of rudimentary knowledge about how immigrant integration responds to an increase in immigration.

To this end, in a recent work [3] a new approach for studying key-integration quantifiers, based on methods, models, and ideas from statistical physics, was proposed. The theory describes and predicts how typical integration quantifiers change when the density of migrants increases. The results predicted a linear growth for the averages of economic quantifiers like permanent and temporary jobs given to immigrants, and a square root growth for the averages of social quantifiers like mixed marriages and infants born to mixed couples. This framework is a powerful tool for policy makers interested in assessing and evaluating integration progresses at the national level.

To deal with the phenomena at the municipality level we use here a different theoretical framework based on the theory and techniques of continuous random walks [4, 5]. The approach developed in [3], based on a full micro-macro statistical mechanics theory, revealed in fact a high efficacy to forecast average values, but, since the developed model does not yet have an exact solution, its related phase space picture is not fully disclosed and does not yet cover the structure of the fluctuations around the mean values. The random walk approach that we follow here instead, which is based on a meso-macro stochastic process, has the advantage of allowing for full analytical control of both mean values and fluctuations.

We consider classical quantifiers of integration such as the fraction of all temporary and permanent labor contracts given to immigrants, the fraction of marriages with spouses of mixed origin (native and immigrant), and the fraction of newborns with parents of mixed origin. The evolution of these quantifiers versus the percentage of migrants inside the host country is 'locally erratic'; that is, when considered at a fine level of resolution such as the municipality, it can be thought of as a *random walk* where the time change is represented by the change of migrant density in the municipality, and the integration quantifier—playing the role of the space variable—changes according to suitable probability distributions defining the stochastic process. Instead of obtaining the evolution of averages via statistical mechanics, with this approach the evolution of averages is here the result of averaging over the whole ensemble of municipalities, i.e., averaging over all the random walks.

From a sociological perspective, the evolution of the quantifiers with respect to the density of immigrants is, in fact, a random process stochasticity which may depend on several exogenous factors driving immigration: fluctuations in the ratio between work demand and work request in the host country [2] or 'biases' resulting from (for example) push–pull factors [2] or different types of network-induced migration outcomes [6–8]. However our aim here is not to explain or disentangle these mechanisms, but rather to look at the evolution of quantifiers as a combined effect of a 'drift' in the presence of some 'noise' regardless of its source/origin. For this task we use random walk theory, which the latter constitutes the prototype of a stochastic process, and, at the same time, the basic model of diffusion phenomena and non-deterministic motion. Indeed, applications can be found in the study of, for example, transport in disordered media (e.g., [9]), anomalous relaxation in polymer chains (see e.g., [10]), financial markets (see e.g., [11]), and quantitative analysis in sports (see e.g., [12]).

Using stochastic process theory enables us to develop a mesoscopic description of the behavior of the integration quantifiers and also to address questions such as whether these socio-economic metrics are determined by memoryless stochastic processes or by processes with long-time correlations. Moreover, this framework enables us to analyze rare events and non-Markovian quantities which are important determinants for planning when they are used as key tools for quantifying fluctuations. That is, we aim to provide efficient tools to help assess the progress (or deficit) in integration as well as to generate strong predictions for extreme-case scenarios at lower administrative levels such as municipalities, and thereby, through an interplay between statistical mechanics and stochastic processes, we broaden the scope of practical applications of the quantitative theory of immigrant integration as a whole. Example of typical questions begging an answer are: what is the worst/best case scenario in the two integration branches—social and economic integration—in a particular municipality if immigrant density changes from say 5 to 7%? And how does the effect magnitude of this change compare with the effect magnitude of an equivalent change at the national level, i.e., average change, or in a similar/dissimilar municipality? In other words, through first-passage-time and maximum-span techniques, we obtain estimates for the expected value of immigrant density for which a particular integration quantifier—say, the share of immigrant workers or the number of mixed marriages—reaches a given threshold above which new policies, structures, services, facilities, etc., have to be made available.

The work is organized as follows: first we describe the database and the procedures for data extraction (section 2), then we explain in detail the mapping between the evolution of a social quantifier and of a random walk (sections 3 and 4), then we report the related results (section 5). Finally, we discuss how such outcomes may be exploited to more effectively set up multiethnic plans and immigration policies in general (section 6). In the appendix we provide technical checks of the robustness of our approach.

2. Data description, analysis and elaboration

Data considered here refer to quarterly observations during the period 1999 to 2010. It is drawn from Spain's Continuous Sample of Employment Histories (the so-called Muestra Continua de

Vidas Laborales or MCVL)[6] and from the local offices of Vital Records and Statistics across Spain (Registro civil)[7]. The former provides detailed data on labor contracts, and the latter provides detailed data on spouses and parents of newborns. Information on the municipalities' immigration density are drawn from the Municipal population registers[8]. A unique feature of the Spanish data is that data sources include so-called 'undocumented immigrants', that is, immigrants who lack a residence permit. Undocumented immigrants are usually not included in official statistical sources. However, their assimilation within the immigrant population is often significant, and excluding them would underestimate the true size of the immigrant population as well as the frequency of the socio-economic events used to measure integration.

Because 'municipality' is the lowest administrative level for which data on density is available, the individual data on mixed events is aggregated to the level of the municipality. From these datasets, for each municipality[9] we obtain quarterly time series for the following quantities:

$$J_p = \frac{\#\text{permanent contracts to immigrant}}{\#\text{permanent contracts}}, \tag{1}$$

$$J_t = \frac{\#\text{temporary contracts to immigrant}}{\#\text{temporary contracts}}, \tag{2}$$

$$M_m = \frac{\#\text{mixed marriages}}{\#\text{marriages}}, \tag{3}$$

$$B_m = \frac{\#\text{newborns with mixed parents}}{\#\text{newborns}}. \tag{4}$$

Notice that the contracts counted in equations (1) and (2) are given to immigrants by native employers.

As explained below, by studying how the quantities in equations (1)–(4) vary with the overall fraction of immigrants, we can unveil the growth law determining their evolution and use this information to provide for them.

To assess the evolution of the Immigrants–Natives system, a convenient quantity to use as a control parameter is

[6] It is an administrative data set with longitudinal information for a 4% non-stratified random sample of the population who are affiliated with Spain's Social Security. We use data from the waves 2005 to 2010. The residence municipality is only disclosed if the population is larger than 40000.

[7] These data are compounded by the 'National Statistical Agency' (INE). The residence municipality is only disclosed if the population is larger than 10000.

[8] More precisely, we use the size of the immigrant population and the native population in each municipality as reported in the 2001 Census as our baseline. Thereafter, based on the information contained in the 'Statistics over residential variation in Spanish municipalities' and statistics on vital events (births and deaths) as elaborated by Spain's 'National Statistical Agency' (INE), we estimate local immigrant densities for different points in time between 1999 and 2010.

[9] Due to data protection, data on mixed marriages and newborns with mixed parents are only available for municipalities with a population larger than 10 000. In addition, and due to data protection, municipality coding for the labor contract data is only available if the municipality's population exceeds 40 000. However, about 85% of Spain's immigrants reside in the included municipalities.

$$\Gamma = N_{imm}N_{nat}/N^2 = \gamma(1-\gamma), \tag{5}$$

where $\gamma = N_{imm}/N$ is the ratio between the number of immigrants N_{imm} and the overall population N, in such a way that its complementary $1-\gamma$ is the number of native people N_{nat} over N ($N = N_{imm} + N_{nat}$). Indeed, Γ provides an intensive measure of the cross-links existing among the communities of natives and of immigrants (however, for small values of γ, $\Gamma \sim \gamma$, hence we can roughly map the percentage of immigrants with the time in our bridge). Moreover, unlike other possible choices such as time, using Γ avoids any inaccuracy due to seasonality and allows direct comparison of municipalities of different sizes (see also [3]). We also stress that Γ nicely captures the 'mixed' nature of the relationships described by the quantifiers in equations (1)–(4) as, whenever $\Gamma = 0$, that is $\gamma = 0$ (i.e., $N_{imm} = 0$) or $\gamma = 1$ (i.e., $N_{imm} = N$), $J_t = J_p = M_m = B_m = 0$.

Complete time series for data on labor contracts involve a number $\mathcal{M}_J$ of municipalities, with $\mathcal{M}_J = 124$, and consist of 2976 data entries over the period 2005–10, which is sampled quarterly (i.e. 24 trimesters overall). Complete series for data on marriages and newborns involve a number $\mathcal{M}_F$ of municipalities, with $\mathcal{M}_F = 581$, and consist of 23 240 data entries spanning the period 1999–2008 which is sampled quarterly (i.e. 40 trimesters overall).

Thus, for any municipality i, we consider five time series: one for $\Gamma^{(i)}$ and one for each observable in equations (1)–(4), hereafter denoted generically as $X^{(i)}$, with $i = 1,...,4$.

As Γ varies, each series $X^{(i)}$ determines a 'path' in the related space, and this point process can be looked at as a continuous-time random walk (CTRW)[10], where the time variable is given by Γ and while the space variable is given by $X^{(i)}$; see figure 1. This mapping is fully described in the next section.

Finally, in figure 2 we show the time series for $X^{(i)}$ and $\Gamma^{(i)}$ versus time (in units of trimesters) to highlight the different shapes of paths.

2.1. Telegraphic introduction on CTRWs

A CTRW process can be depicted as a dynamical point (to fix ideas embedded in a one-dimensional space, since here we need only such a case), which occupies a position $r(t)$ at time t (see also figure 3). Let us suppose that the point starts on the origin, that is $r(0) = 0$. It then stays fixed to its position until time t_1, when it jumps to $\boldsymbol{\xi}_1$, where it waits until time $t_2 > t_1$, when it jumps to a new location $\boldsymbol{\xi}_1 + \boldsymbol{\xi}_2$, and so on. The series $\{t_1, t_2,...\}$ defines the times of jumping events. The times $\tau_1 = t_1 - 0$, $\tau_2 = t_2 - t_1$, etc are called waiting times.

The waiting times $\{\tau_i\}$ and the width of the instantaneous jumps $\{\xi_i\}$ are continuous random variables extracted from the distribution $\psi(\xi, \tau)$. The latter determines the long-time properties of the walk: a diverging average waiting time typically corresponds to sub-diffusive behaviors, whereas a diverging variance for jump widths typically corresponds to super-diffusive behaviors.

In particular, for the so-called decoupled continuous random walk (namely where the distribution $\psi(\boldsymbol{\xi}, \tau)$ factorizes into $\psi(\boldsymbol{\xi}, \tau) = f(\boldsymbol{\xi})\psi(\tau)$), the waiting times and the instantaneous displacements are mutually independent (identically) distributed random variables.

[10] The continuous time random walk (CTRW) was introduced by Montroll and Weiss [18]; see also [4, 5] for recent reviews and SI for a deeper description.

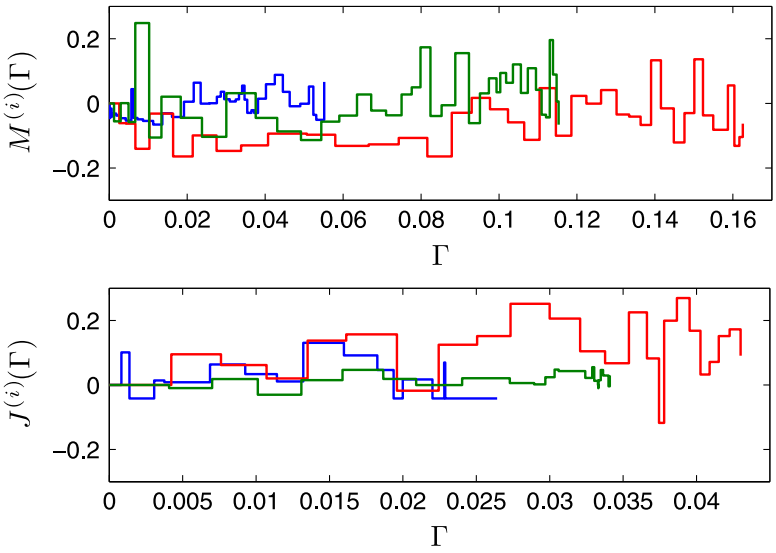

Figure 1. Examples of paths for the quantifiers M_m (upper panel) and J_p (lower panel) shown as a function of Γ. Three different municipalities are depicted in different colors. These paths can be compared with a theoretical one depicted in figure 3 and related to a CTRW. In this figure the time series $\{X^{(i)}\}$ and $\{\Gamma^{(i)}\}$ have been properly initialized to allow an effective comparison; more precisely values are shifted as $X_j^{(i)} \to X_j^{(i)} - X_1^{(i)}$ and $\Gamma_j^{(i)} \to \Gamma_j^{(i)} - \Gamma_1^{(i)}$, for any j.

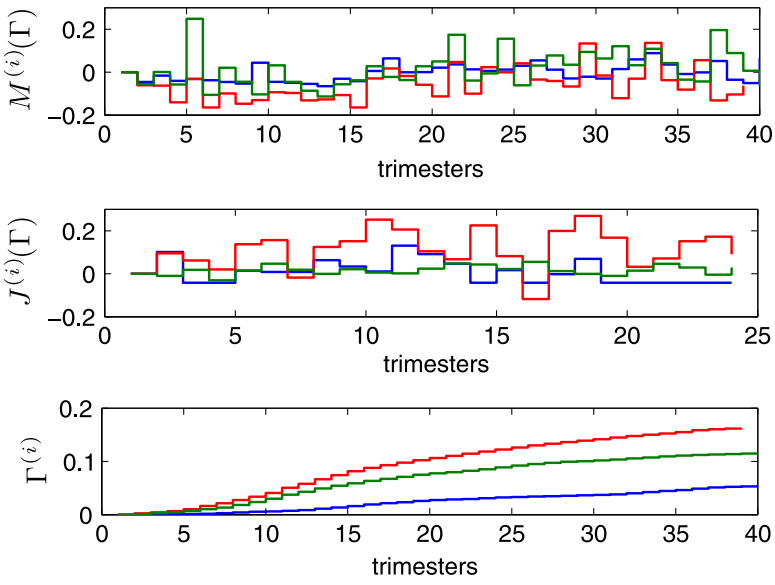

Figure 2. Examples of paths for the quantifiers M_m (upper panel) and J_p (lower panel) shown as a function of time (1 unit = 1 trimester). Three different municipalities (the same as in figure 1) are depicted in different colors. Notice that seasonality effects emerge for marriages: during summer months marriages are more frequent.

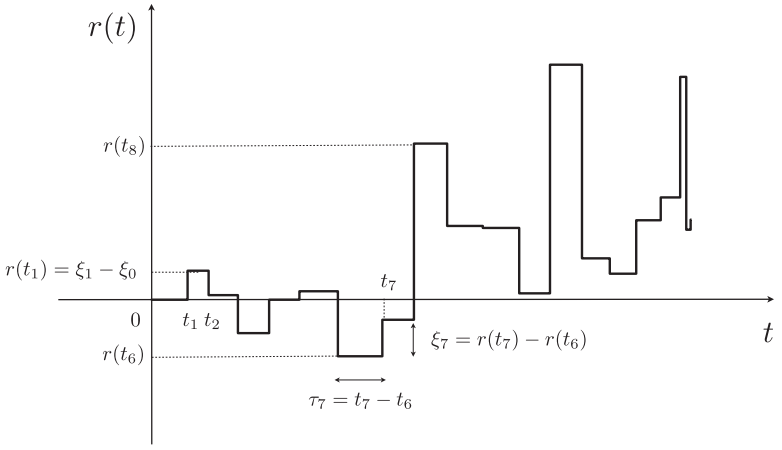

Figure 3. Example of path realized by a CTRW for which step widths and waiting times are extracted from the distributions given by equations (28) and (30), respectively, and with parameters consistent with those found experimentally (see table 2).

The position r of the particle at the k-th jump, that is, at time t_k, is given by the sum $r(t_k) = \sum_{i=1}^{k} \xi_i$. Getting $r(t)$, namely a direct dependence on t, requires the introduction of the random variable $n(t)$, representing the number of steps m performed up to time t and defined by $n(t) = \max\{m: t_m \leqslant t\}$, in such a way that

$$r(t) = \sum_{i=1}^{n(t)} \xi_i. \tag{6}$$

The expected value $\bar{r}(t)$ of the displacement can be derived from the probability distributions for the waiting time and for the step length. In fact, focusing on the decoupled case[11], we can define $\bar{\xi} = \int \xi f(\xi) d\xi$ and $\bar{\tau} = \int \tau \psi(\tau) d\tau$, whereby, as long as $\bar{\tau}$ is finite, one can show that, in the limit of large t [13]

$$\overline{r(t)} \sim \bar{\xi} \frac{t}{\bar{\tau}}. \tag{7}$$

Thus, if there is no net drift ($\bar{\xi} = 0$), the average displacement is zero and one usually looks at the mean square displacement, which turns out to scale as $\overline{r^2(t)} \sim \bar{\xi}^2 t/\bar{\tau}$, and the purely diffusive limit can be recovered.

On the other hand, in the presence of a net drift ($\bar{\xi} \neq 0$), the mean displacement can also be expressed in terms of the mean number of steps $\overline{n(t)}$ performed up to time t as (see e.g., [13, 14])

$$\overline{r(t)} = \overline{n(t)} \cdot \bar{\xi}, \tag{8}$$

and, accordingly, $\overline{r^2(t)} \sim \overline{r(t)}^2$ [13, 14]. From equation (8), one can see that if the average time diverges or displays any anomalous behavior, the biased motion turns out to be anomalous as well.

Of course, the definitions given here can be extended to a geometrical space with arbitrary topology [4].

[11] As we will show, this is the case recovered by our experimental data

Despite the fact that this random walk process is, by definition, Markovian, one can also introduce non-Markovian related quantities such as the mean-first passage time $\tilde{t}$ and the maximum span $\tilde{r}$, [15].

The mean-first passage time represents the mean time taken by a random walk to first reach a (fixed) point placed at a given initial distance r. Its dependence on r qualitatively depends on the kind of diffusion realized, in particular:

$$\tilde{t} \sim r^2, \text{ for pure diffusion} \tag{9}$$

$$\tilde{t} \sim r, \text{ for biased diffusion.} \tag{10}$$

The maximum span represents the farthest distance ever reached by a random walk up to time t. Again, the functional form of $\tilde{r}$ as a function of t depends on the kind of diffusion realized:

$$\tilde{r} \sim \sqrt{t} \text{ for pure diffusion} \tag{11}$$

$$\tilde{r} \sim t, \text{ for biased diffusion.} \tag{12}$$

These relatively simple laws stem from the peculiarity of the one-dimensional structure. In general, the behavior of $\tilde{t}$ and $\tilde{r}$ functionally depends on the underlying topology.

Indeed, due to their non-Markovian nature, estimating such quantities may be rather tricky, yet they are intensively studied because they provide useful information and play an important role in many real situations (e.g. transport in disordered media, neuron firing, spread of diseases, and target search processes [4, 16, 17]).

To summarize, the CTRW is a stochastic model for which $\psi(\tau)$ and $f(\xi)$ serve as input functions. The output is provided by the temporal series $\{t_1, t_2, ...\}$ and $\{r_1, r_2, ...\}$ from which quantities such as mean squared displacement, mean first-passage time, etc can be calculated.

In the next section, the jump widths ξ_i's as well as the positions $r(t)$ will assume different meanings (i.e. number of mixed marriages, of infants born to mixed couples, and of temporary/permanent contracts to immigrants) according to the specific quantifier addressed.

3. The mapping in a nutshell

Let us denote with $X^{(i)}$ a generic quantifier (i.e. the number of mixed marriages, of newborns from mixed couples, and of temporary/permanent contracts to immigrants), where i specifies the municipality. According to the quantifier considered i is bounded by $\mathcal{M}_J$ or by $\mathcal{M}_F$.

Therefore, we have the time series

$$\left\{X_1^{(i)}, X_2^{(i)}, ..., X_{\mathcal{T}}^{(i)}\right\}, \tag{13}$$

$$\left\{\Gamma_1^{(i)}, \Gamma_2^{(i)}, ..., \Gamma_{\mathcal{T}}^{(i)}\right\}, \tag{14}$$

where $X_n^{(i)}$ and $\Gamma_n^{(i)}$ are the values of the quantifier and of the number of cross-links at the nth trimester and $\mathcal{T}$ is bounded by the overall number of trimesters over which measures have been taken (i.e., 24 for job quantifiers and 40 for family quantifiers).

For a (one-dimensional) CTRW of $\mathcal{T}$ steps, defined by the two series

$$\left\{\xi_1, \xi_2, ..., \xi_{\mathcal{T}}\right\}, \tag{15}$$

$$\{t_1, t_2, ..., t_{\mathcal{T}}\}, \tag{16}$$

where ξ_n is the jump width and t_n is time when the n-th step occurs, we recall that the position $r(t)$ of a walker at time t is obtained by $r(t) = \sum_{j=1}^{n(t)} \xi_j$, where $n(t)$ is the number of steps performed up to time t.

Analogously, we can state that, for the i-th municipality, the value of the quantifier $X^{(i)}(\Gamma)$ corresponding to fraction of cross-link Γ is

$$X^{(i)}(\Gamma) = \sum_{j=1}^{n^{(i)}(\Gamma)} \Delta X_j^{(i)}, \tag{17}$$

where $\Delta X_j^{(i)} = X_{j+1}^{(i)} - X_j^{(i)}$ and $n^{(i)}(\Gamma)$ is the latest trimester for which $\Gamma_j^{(i)} < \Gamma$.

Therefore, we can look at the set of $\mathcal{M}$ municipalities as a set of $\mathcal{M}$ random walks. Actually, before proceeding, a couple of remarks are in order.

In principle, Γ and X are bounded by 1, yet, the number of immigrants corresponds to a small fraction of the overall population in such a way that $\Gamma, X << 1$ and we can neglect boundaries[12].

Moreover, Γ and X are not continuous variables as there exists an intrinsic unit given by 1/#*number of marriages*, 1/#*number of newborns* and 1/#*number of contracts*, representing our experimental sensitivity. However, such a unit is in general much smaller than the quantities measured, which can therefore be considered as continuous.

Therefore, we can treat the set of $\mathcal{M}$ municipalities as a set of $\mathcal{M}$ random walks, for which we can build the following ensemble average:

$$\langle X(\Gamma) \rangle \equiv \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} X^{(i)}(\Gamma). \tag{18}$$

Similarly, for the average square distance covered

$$\langle X^2(\Gamma) \rangle \equiv \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \left[X^{(i)}(\Gamma) \right]^2. \tag{19}$$

The progression of the quantifiers $\langle X(\Gamma) \rangle$ averaged over the whole set of municipalities, that is to say, the average displacement of the related CTRW, is shown in figure 4, where fits evidence the following behaviors

$$\left\langle J_t(\Gamma) \right\rangle \sim \Gamma, \tag{20}$$

$$\left\langle J_p(\Gamma) \right\rangle \sim \Gamma, \tag{21}$$

$$\left\langle M_m(\Gamma) \right\rangle \sim \sqrt{\Gamma}, \tag{22}$$

$$\left\langle B_m(\Gamma) \right\rangle \sim \sqrt{\Gamma}. \tag{23}$$

perfectly consistent with those outlined in [3], despite the fact that procedure for their derivation is conceptually different; this confers robustness to the above results.

[12] Conversely, if boundaries can not be neglected the mapping could still be feasible but we should refer to the theory of random walks on finite chains

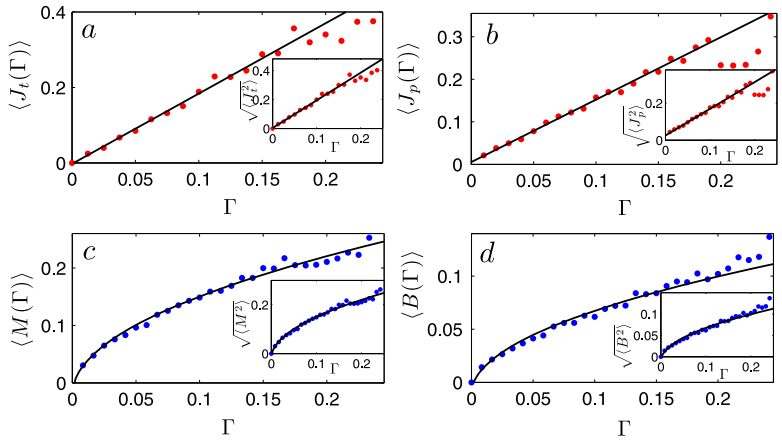

Figure 4. 'Mean displacement' (main figures) and 'mean square displacement' (insets) versus 'time' for the CTRWs associated with J_t (panel a), J_p (panel b), M_m (panel c) and B_m (panel d). Data available were binned over Γ and averaged over the set of $\mathcal{M}$ municipalities; the resulting values (•) and the related best fit (solid line) are shown. In particular, for family quantifiers we fitted by the law $r = p_1\sqrt{t} + p_2$, whereas for job quantifiers we used the law $r = p_3 t + p_4$; best fit coefficients are summarized in table 1. In general, the goodness-of-fit R^2 ranges between 0.97 and 0.99. Notice that $\sqrt{\langle X^2(\Gamma)\rangle} \sim \langle X(\Gamma)\rangle$ suggests the presence of a drift [13].

To summarize, in our random-walk picture for the time evolution of the social quantifier X, in each municipality the quantifier starts from zero and, for a given variation of the related immigrant percentage Γ, the quantifier increases or decreases until the path ends. The trajectory of X versus Γ qualitatively resembles the position of a CTRW as a function of time (see figures 1 and 3).

In the next section we analyze the CTRWs associated with the quantifiers and try to get a microscopic perspective for the origin of these laws. Such a perspective will enable us to speculate about possible effects and to make crucial forecasts.

4. Formalizing the mapping

We first check that the CTRWs corresponding to J_p, J_t, M_m and B_m are decoupled, that is, the related probability distributions $\psi(\Delta X, \Delta\Gamma)$ for the generic increments ΔX and $\Delta\Gamma$ can be factorized into $f(\Delta X)\psi(\Delta\Gamma)$: this is achieved through direct inspection of the scatter plots reported in figure 5. This point is further explored in the appendix, where we also show that $\Delta X_n^{(i)}$ and $\Delta\Gamma_n^{(i)}$ turn out to be uncorrelated with respect to the 'instantaneous values' $X_n^{(i)}$ and $\Gamma_n^{(i)}$.

Thus, we can proceed by studying separately $f(\Delta X)$ and $\psi(\Delta\Gamma)$. We recall that such distributions provide qualitative information about the diffusive behaviors of the walks associated with our quantifiers, that is, on their time progress. Moreover, from $f(\Delta X)$ and $\psi(\Delta\Gamma)$, we are able to derive the expectation values

$$\overline{\Delta X} = \int \Delta X f(\Delta X) d\Delta X, \tag{24}$$

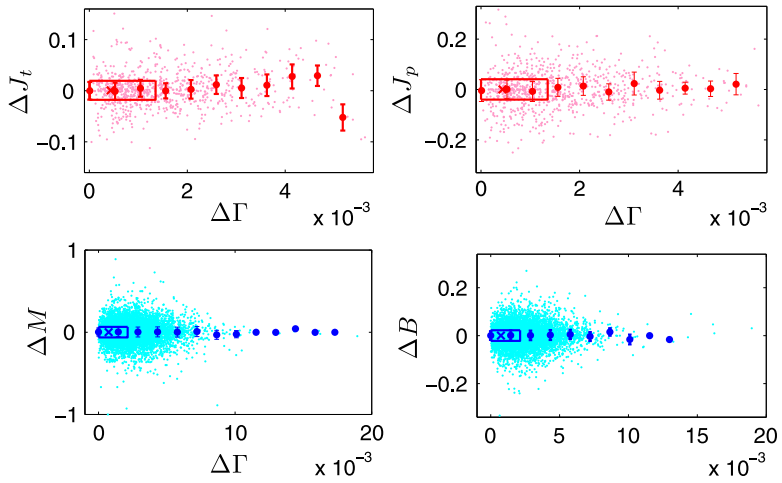

Figure 5. These scatter plots evidence the existence of any correlation between the 'waiting times' $\Delta\Gamma$ and the 'jump width' ΔJ_t, ΔJ_p, ΔM_m, and ΔB_m: each point represents the increments ΔX_n versus $\Delta\Gamma_n$; all $\mathcal{T}$ steps and the whole set of municipalities are considered. The clouds of data are uniform and do not reveal any special trend. Binned spots evidence the possible values of increments $\Delta\Gamma_n$, and for each bin we calculated the average of the related increments ΔX_n; the related standard deviations are also depicted. Notice that such averages are basically constant (at least within the error) with respect to $\Delta\Gamma_n$, and this allows us to conclude that no clear correlation emerges.

$$\overline{\Delta\Gamma} = \int \Delta\Gamma \psi(\Delta\Gamma) d\Delta\Gamma, \tag{25}$$

which act as the expected jump length and as the expected waiting time, respectively. Analogously, we can derive $\overline{n(\Gamma)}$, which acts as the expected number of steps performed up to 'time' Γ, that is

$$\overline{n(\Gamma)} = \sum_n n\, Q(n|\Gamma), \tag{26}$$

where $Q(n|\Gamma)$ is the probability that $\sum_j^n \Delta\Gamma_j$ is smaller than Γ, but $\sum_j^{n+1} \Delta\Gamma_j$ is larger that Γ.

From these quantities, one finally has (see e.g. [13, 14])

$$\overline{X(\Gamma)} = \overline{n(\Gamma)} \cdot \overline{\Delta X}. \tag{27}$$

Of course, the expectation $\overline{X(\Gamma)}$ and the ensemble average $\langle X(\Gamma)\rangle$ ought to be consistent (as checked in the next section). This ensures the ergodicity of the system and will enable us to exploit the analytical results derived starting from the probability distribution functions also for our 'time' series.

4.1. Step width and waiting time distributions

Let us start with the distribution for the 'step lengths' $f(\Delta X)$. In figure 6 we show the histogram for the increments ΔJ_t, ΔM, ΔJ_p, and ΔB obtained from experimental data. In all cases the symmetric, centered exponential distribution

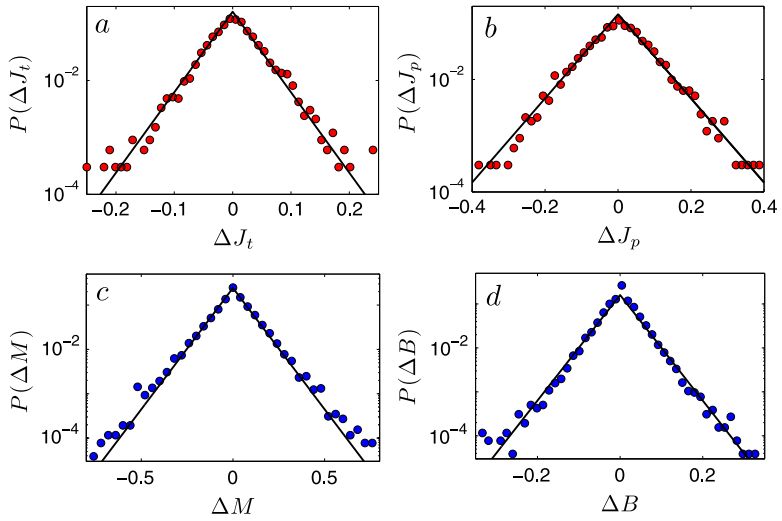

Figure 6. Distributions $f(\Delta J_t)$ (panel a), $f(\Delta J_p)$ (panel b), $f(\Delta M_m)$ (panel c) and $f(\Delta B_m)$ (panel d) measured from experimental data, without distinguishing between municipalities; that is, we merged the increments pertaining to the whole ensemble of walks and built a unique histogram. Notice the semi-logarithmic scale plot. Data (•) are fitted by using equation (28) (solid line); best-fit coefficients and averages on raw data are presented in table 2

Table 1. Best-fit coefficients related to plots shown in figure 4.

Quantifier X	p_1	p_2
$\langle M_m \rangle$	0.54 ± 0.02	−0.019 ± 0.009
$\sqrt{\langle M_m^2 \rangle}$	0.57 ± 0.03	0.007 ± 0.06
$\langle B_m \rangle$	0.25 ± 0.01	−0.010 ± 0.009
$\sqrt{\langle B_m^2 \rangle}$	0.287 ± 0.002	−0.007 ± 0.004
Quantifier X	**p_3**	**p_4**
$\langle J_t \rangle$	1.9 ± 0.1	−0.003 ± 0.001
$\sqrt{\langle J_t^2 \rangle}$	1.9 ± 0.1	0.003 ± 0.001
$\langle J_p \rangle$	1.47 ± 0.06	0.005 ± 0.003
$\sqrt{\langle J_p^2 \rangle}$	1.45 ± 0.07	0.025 ± 0.008

$$f(\Delta X) = \lambda e^{-\lambda|\Delta X|}, \tag{28}$$

provides an excellent fit. An exponential distribution for step lengths ensures that the related CTRW does not exhibit any super-diffusive feature as the central limit theorem is fulfilled.

Now, the fit coefficient λ depends on the quantifier considered and is directly related to the expected value by $\lambda_X^{-1} = \overline{\Delta X}$. Results are presented in table 2, where a comparison with the experimental average values $\langle |\Delta X| \rangle$ and $\langle \Delta X \rangle$ is also provided.

The goodness of the fit is corroborated by the fact that λ_X^{-1} and $\langle |\Delta X| \rangle$ coincide within the error. However, looking at $\langle \Delta X \rangle$, we report a slight deviation: although one would expect a null average value due to the centrality of the distribution, the average is systematically positive for

Table 2. The second column contains the best-fit coefficients obtained by fitting, according to equation (28), the probability distribution function of the displacements ΔX shown in figure 6; the third and fourth columns contain the related average values, where the average is performed on raw data over all municipalities. Because it is the support of the exponential distribution positive, λ_X^{-1} has to be compared with $\langle|\Delta X|\rangle$. Moreover, we checked that the absolute error on $\langle|\Delta X|\rangle$ is approximately equal to $\langle|\Delta X|\rangle$ itself, as expected from an exponentially-distributed variable. Notice that the average displacement $\langle\Delta X\rangle$ in a single step is positive for any quantifier (the standard deviation $\sqrt{\langle\Delta X^2\rangle - \langle\Delta X\rangle^2}$ calculated on raw data is comparable with $\langle\Delta X\rangle$).

Quantifier X	λ_X^{-1}	$\langle\|\Delta X\|\rangle$	$\langle\Delta X\rangle$
J_t	0.031 ± 0.002	0.03	0.003
J_p	0.058 ± 0.003	0.06	0.003
M_m	0.079 ± 0.002	0.08	0.003
B_m	0.035 ± 0.001	0.03	0.001

all quantifiers, which implies that, as Γ increases, X is more likely to grow rather than to decrease. In the random-walk picture, this can be interpreted as the presence of a drift which biases the motion of the walker. Let us now move to the distribution for the 'waiting times' $\Delta\Gamma$. In figure 7 we show the histogram for the increments $\Delta\Gamma$ obtained from experimental data related to the time period and to the municipalities considered. Interestingly, here qualitative differences emerge between the job quantifiers, i.e. J_t and J_p, and the family quantifiers, i.e. M_m and B_m.

Before proceeding it is worth stressing that for job quantifiers and family quantifiers the time along which sampling has been performed is not exactly the same, being, respectively, 2005–10 and 1999–2008 (of course, the consistency between the related time series has been checked for the overlapping period [3]). Now, to ensure that the qualitative differences reported do not stem from different time intervals, but instead are intrinsic, we repeated the analysis shown in figure 7 by restricting the calculations to only to the common time lapse 2005–08 and, indeed, we checked the robustness of the result.

In fact, calling ψ_F and ψ_J the distributions for family and job quantifiers, respectively, the reason for their intrinsic difference can be explained by the way mapping between quantifier evolution and random-walks has been fixed. In particular, there exist trimesters i for which a growth in the number of immigrants is reported; i.e. $\Gamma_i - \Gamma_{i-1} > 0$, but no change in the quantifier X considered occurs, i.e. $\Delta X_i = 0$. In such cases the two trimesters behave as practically merged as the overall waiting time approaches $\Gamma_{i+1} - \Gamma_{i-1}$. This concept can be repeated iteratively until each step of the walk actually corresponds to a true displacement. Thus, as one can see from figure 7, such merging is more frequent for family quantifiers in such a way that the related waiting times display a larger range; or, to put it another way, the integration of immigrants within the market is more direct: as long as new immigrants arrive, a fraction of them get a job, either permanent or temporary. Conversely, the integration of immigrants from a familiar perspective is more complex and does not follow a prescribed pattern: not surprisingly, the arrival of new immigrants does not necessarily correspond to integration when considering these quantifiers. This is consistent with the results in [3], where from a different perspective, it is shown that the qualitative difference between the laws

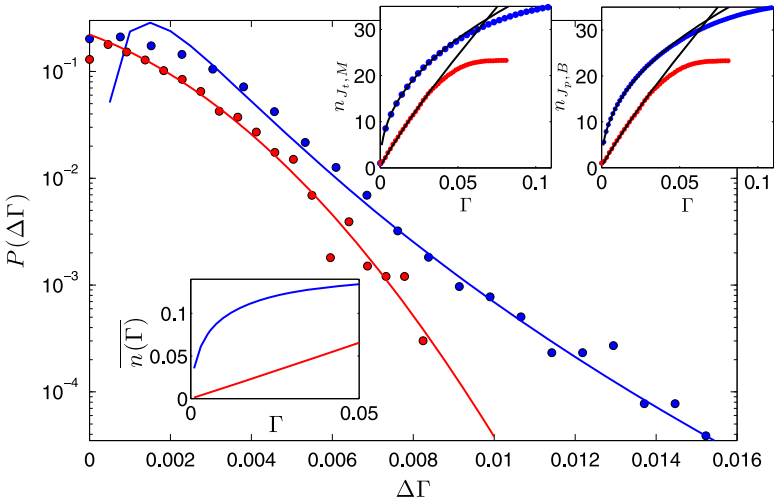

Figure 7. Main plot: histograms for $\Delta\Gamma$, derived from experimental data concerning marriages (blue symbols) and permanent jobs (red symbols), are shown and compared. Solid lines represent the best fit according to a lognormal distribution (see equation (29)) and a half-Gaussian distribution (see equation (30)), respectively. Fitting coefficients and related errors are reported in table 3. Notice that such histograms were derived without distinguishing between municipalities. Lower inset: average number of steps performed up to time Γ, calculated numerically from equation (29) (red line) and equation (30) (blue line), respectively. Upper insets: average number $\langle n(\Gamma)\rangle$ of steps performed by the related random walker up to the fraction of immigrants Γ. Solid lines correspond to the law $y \sim x$ and $y \sim \sqrt{x}$, respectively and evidence qualitatively different behaviors for marriages and jobs. This picture corroborates the validity of equation (8) with the ensemble average: $\langle r\rangle \sim \langle n\rangle\langle \Delta r\rangle$, which bridges the picture itself with figure 1. The fit is robust only up to relatively small values of Γ; then experimental averages are underestimated. This is due to the fact that the statistics are robust only for values of Γ which are reached by (almost) all walks. For larger values our averages are only an underestimate of the expected, effective mean value of n.

$M_m(\Gamma)$, $B_m(\Gamma)$, and $J_t(\Gamma)$, $J_p(\Gamma)$ is due to a different degree of interaction among agents in the two different scenarios (families and jobs).

It is worth stressing that such an effect is not directly imputable to the seasonality of marriages; this can be seen, for instance, from the fact that for newborns the same effect emerges as well, but their time series do not display any seasonality.

Let us now analyze in more detail the waiting time distributions.

For family quantifiers the distribution $\psi_F(\Delta\Gamma)$ fitting the experimental histogram is a log-normal distribution

$$\psi_F(\Delta\Gamma) = \frac{1}{\Delta\Gamma\sqrt{2\pi}\,\sigma}\exp -\frac{(\log\Delta\Gamma - \mu)^2}{2\sigma^2}, \tag{29}$$

for which the average value is expected to be $\overline{\Delta X} = e^{\mu+\sigma^2}$. As for jobs, the best fit is provided by a half-normal distribution

Table 3. Best-fit coefficients obtained by fitting the probability distribution function of the 'waiting time' $\Delta\Gamma$ shown in figure 7 according to equations (29) and (30). The relative error on fit coefficients ranges between 10% and 20%. Within the error there is perfect consistency between the average values $\overline{\Delta X}$ and $\langle \Delta X \rangle$, as well as between the variance of such distributions and the variance on the related raw data. Here we report only data for marriages and permanent jobs; for newborns and temporary jobs, analogous analysis shows only slight quantitative changes.

Γ	μ	σ^2	$\overline{\Delta\Gamma}$	$\langle \Delta\Gamma \rangle$
Job	$(1.2 \pm 0.2) \cdot 10^{-3}$	$(6.7 \pm 0.6) \cdot 10^{-6}$	$(2.0 \pm 0.2) \cdot 10^{-3}$	$(1.7 \pm 0.2) \cdot 10^{-3}$
Family	-6.6 ± 0.9	0.32 ± 0.04	$(1.7 \pm 0.3) \cdot 10^{-3}$	$(1.9 \pm 0.3) \cdot 10^{-3}$

$$\psi_J(\Delta\Gamma) = \frac{\sqrt{2}}{\sqrt{\pi}\sigma} \exp - \frac{(\Delta\Gamma - \mu)^2}{2\sigma^2}, \tag{30}$$

for which the average value is expected to be $\overline{\Delta X} = \mu$. Details on fitting coefficients and average values are all presented in table 3; notice that, in both cases, $\overline{\Delta\Gamma}$ turns out to be comparable with the ensemble average $\langle \Delta\Gamma \rangle$.

Thus, although both ψ_J and ψ_F fulfill the central limit theorem and display a finite mean, the latter displays a long tail so that we expect that the growth for family quantifiers may be slowed down.

Now, given ψ_J and ψ_F, we can derive the number of steps performed up to time Γ, exploiting the properties of Laplace transforms (see e.g. [13, 14]). Examples of numerical results of these calculations are shown in the lower inset of figure 7; the difference between the two cases is striking.

To check this point, we measure directly on raw data the average number $\langle n(\Gamma) \rangle$ of steps performed before reaching the time Γ (see figure 7). Indeed, for jobs we find a roughly linear growth, i.e. $\langle n(\Gamma) \rangle \sim \Gamma$, whereas for marriages and births we find a slower growth, i.e. $\langle n(\Gamma) \rangle \sim \sqrt{\Gamma}$.

Such a qualitative difference, together with equation (8), immediately explains the results of equations (20)–(23).

Summarizing, both processes display a non-null positive drift, i.e. $\langle \Delta X \rangle > 0$, yet the resulting behaviors are qualitatively different over the time window considered. Such a difference ultimately stems from deep differences in the waiting times: a broader distribution for $\Delta\Gamma$ occurs in the case of family quantifiers and the related random walks may experience rather long waiting times, although the jump widths remain narrowly distributed. The net result is just a slowing down in the progress of the quantifier.

Conversely, as for jobs, both ΔX and $\Delta\Gamma$ are narrowly distributed so that at each trimester we do not expect strong variations in the fraction of new immigrants getting a job.

Such a difference suggests an intuitive motivation, namely that the mechanisms underlying the emergence of mixed marriages are more complex and may be subjected to mutual interaction among individuals. This is perfectly consistent with the statistical–mechanics description of the phenomenon provided in [3].

5. First predictive outcomes for social planners

We now turn to the theory's predictive capacity. The aim is to present concrete instruments that can aid policy makers at the municipal level in their work to accommodate and plan for further immigration. We focus on two well-known observables: the (mean) first passage time, and the (mean) maximum walk span.

5.1. Mean first-passage time

Mean first-passage-time quantities have been extensively investigated in a number of different fields, ranging from chemical kinetics to finance, because they provide an estimate for the average time at which a given stochastic event is triggered [16, 17].

Given the process $X(\Gamma)$, we calculate the value $\tilde{\Gamma}(x)$ at which the quantifier reaches a certain threshold x. To evaluate the typical value of $\tilde{\Gamma}(x)$ we perform an average over the ensemble of walks; that is

$$\langle \tilde{\Gamma}(x) \rangle = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \tilde{\Gamma}^{(i)}(x). \tag{31}$$

The quantity $\langle \tilde{\Gamma}(x) \rangle$ allows predictions on the consequences of additional immigrants on integration and when an integration threshold is likely to be reached. For instance, let us say that when an integration quantifier reaches the threshold x, some integration policies, activities, or services must be activated (e.g. concerning public education, public health, etc). Then, as Γ approaches $\langle \tilde{\Gamma}(x) \rangle$, local projects and plans need to be activated.

In figure 8 we show the mean-first passage time for the quantifiers considered in this work as a function of X.

The mean first-passage time is especially useful for policies, plans, and services that are coupled with a concrete 'discrete' integration target and when we need to know the expected time when the politically defined threshold is reached and plans must be activated.

For example, we could ask at which value of Γ (which is related to the percentage of migrants) we expect the number of newborns born to mixed parents reaches the threshold of 10%. By simply looking at the behavior of $\langle X(\Gamma) \rangle$, by inverting, we would get $\Gamma \sim 0.2$. However, due to huge fluctuations (in some peculiar municipalities), the threshold of 10% can be reached much earlier, since the first passage time returns a value $\Gamma \sim 0.04$. Hence, planning based on average evolutions only may underestimate reality by a factor thereby rendering planning and resource allocation extremely ineffective.

5.2. Walk span

The walk span represents the largest point reached by the walker up to a given time; that is, the largest value $\tilde{X}$ reached by X up to Γ. More precisely, we say that for the i-th walk, at the k-th step, the span is $\tilde{X}^{(i)}$ if $X(n)^{(i)} < \tilde{X}^{(i)}(k)$, $\forall n \leqslant k$. Again, to evaluate the typical value of $\tilde{X}(k)$, we perform an average over the ensemble of walks; that is

$$\langle \tilde{X}(k) \rangle = \frac{1}{\mathcal{M}} \sum_{i=1}^{\mathcal{M}} \tilde{X}^{(i)}(k). \tag{32}$$

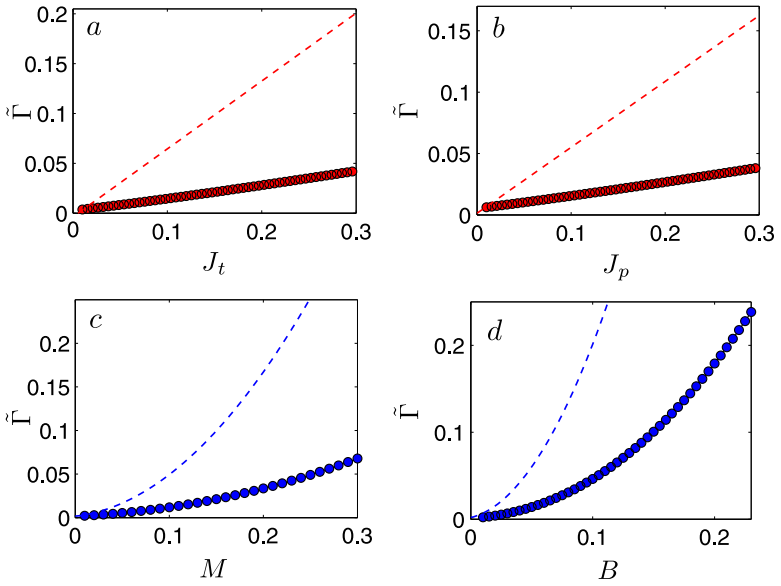

Figure 8. Mean time $\tilde{\Gamma}$ to first reach a given value of J_t (panel a), of J_p (panel b), of M (panel c), of B (panel d). Experimental data (•) are obtained by first getting the mean number of steps to first reach the distance X and then by inverting through $n(\Gamma)$ (see figure 7); this procedure improves the stability of results. Errors on data points are $\lesssim$15%. Solid lines are best fits given by $y = p_1''\Gamma + p_2''$ (upper panels) and by $y = p_3''\Gamma^2 + p_4''$ (lower panels), being $p_1'' = 0.13 \pm 0.02$, $p_2'' = 0.0014 \pm 0.0005$ for J_p, $p_1'' = 0.11 \pm 0.02$, $p_2'' = 0.0045 \pm 0.0001$ for J_t, and $p_3'' = 0.70 \pm 0.01$, $p_4'' = 0.0047 \pm 0.0003$ for M_m, $p_3'' = 4.54 \pm 0.03$, $p_4'' = 0.000\,44 \pm 0.0002$ for B_m. These results are compared with the related $\Gamma(X)$ (dashed line) derived from results shown in figure 4; see also data in table 1 for comparison.

The average walk span provides information on the capacity to integrate further immigration. In fact, in organizing local integration policies and making appropriate priority decisions among different integration initiatives, one is interested in the span of, say, the number of children, or the number of immigrants with permanent jobs, rather than in their average number, since the latter may lead to dramatic over- and underestimations.

In figure 9 we show the span of the quantifiers considered in this work as a function of Γ. We notice that the qualitative differences already evidenced for $\langle X(\Gamma)\rangle$ are robust and the span for marriages and births grows like $\sqrt{\Gamma}$, whereas the span for temporary and permanent jobs grows like Γ. The persistence of such behaviors is consistent with the fact that such random walks display distributions for waiting time and step width having finite average and variance. For instance, for a simple random walk on a line the span grows in time like $\sqrt{t}$, whereas in the presence of a drift one has a linear law t [4].

6. Conclusions

Theoretical models, originally developed to solve physical problems, are increasingly being used to study social phenomena. Statistical mechanics and stochastic process theory are particularly well suited for this task, and have generated a novel quantitative understanding of the underlying complexity of social interactions. In this paper we focused on stochastic

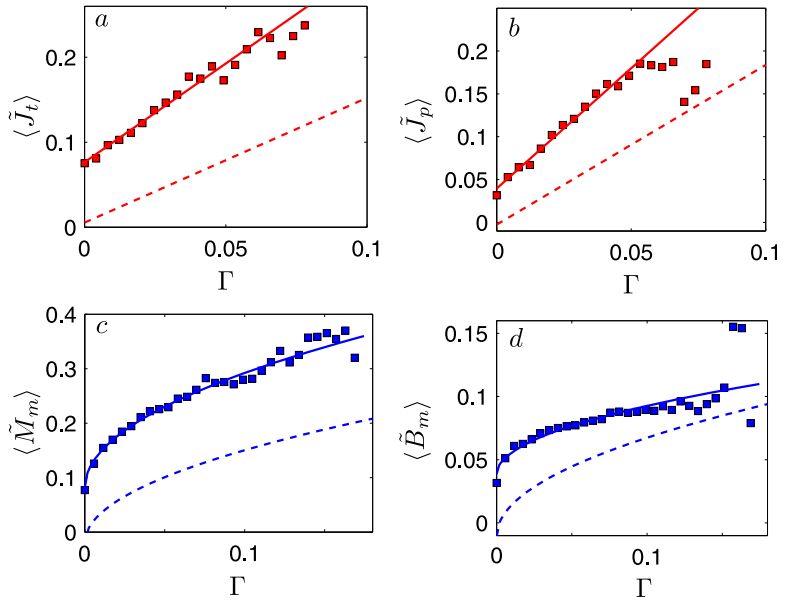

Figure 9. Span of the walk for permanent jobs (panel *a*), for temporary jobs (panel *b*), for marriages (panel *c*), and for newborns (panel *d*) versus Γ. Errors on data points are $\lesssim$10%. Solid lines are best fits given by $y = p_1'\Gamma + p_2'$ (upper panels) and by $y = p_3'\sqrt{\Gamma} + p_4'$ (lower panels), being $p_1' = 2.3 \pm 0.2$, $p_2' = 0.08 \pm 0.01$ for J_p, $p_1' = 2.8 \pm 0.2$, $p_2' = 0.04 \pm 0.01$ for J_t, and $p_3' = 0.8 \pm 0.1$, $p_4' = 0.04 \pm 0.01$ for M_m, $p_3' = 0.17 \pm 0.02$, $p_4' = 0.04 \pm 0.01$ for B_m. These results are also compared with the curves $X(\Gamma)$ from figure 4 (dashed line); see also data in table 1 for comparison.

processes. We identified the random behavior of the four integration quantifiers with random walkers: each municipality draws a random walk in the quantifier–migrant's density plane. Averaging over all the municipalities then allowed us to investigate the evolution of the quantifier averages, which are found to scale with the square root of the percentage of migrants for familiar quantifiers and linearly with the percentage of migrants for job quantifiers, in complete agreement with previous findings obtained through the statistical–mechanical route [2]. We inferred the distributions of jumps and waiting times (which are found to be decoupled). Whereas jump distributions are exponentially distributed for all the quantifiers, waiting-time distributions depend on the context: social quantifiers have log-normal distributions, whereas economic quantifiers display Gaussian distributions.

This difference has a simple explanation. Although there is a correlation, even on a short timescale, between the last-arrived migrant and that immigrant's incorporation into the labor market (to sustain himself or herself), the same is not true for marriages or newborns. Clearly, the correlation is likely to be negligible between the last arrived immigrant and a mixed marriage or birth event (i.e., it is unlikely that the arriving immigrant and the one, say, marrying a native are the same person). This results in a stronger noise affecting social quantifiers, which destroys the net drift, leaving simple diffusion as the only survivor. On the contrary, driven by the migrant's necessity to work, economic quantifiers display ballistic motion. Another element that contributes to the macroscopic differences resides in the much broader distribution of jumps for the working quantifiers: The fat tail encoding for the long jumps in the working

quantifiers implies a larger value of drift, that, coupled with much less noise – for the reasons just mentioned – results in ballistic motion.

From a practical perspective, no power-law distributions are found. Hence, the central limit theorem holds, which implies that the theory is suitable for generating predictions. To this end, we introduced two predictive non-Markovian tools: the 'mean first passage time' and the 'maximum span walk'. Using these tools we were able to tackle in a more scientific way questions that traditionally have been answered by using guesstimates. For example, our predictive framework can easily produce forecasts of the share of newborns with mixed parents following an increase in the share of immigrants from, say, 3 to 5%. We make two types of forecasts: first, we assess the evolution of the mean of this quantifier.

The evolution is obtained by evaluating from figure 4 the average increment, which is roughly from $B(\Gamma) = 0.04$ to $B(\Gamma) = 0.05$. Second, we assess the mean worst case by dealing with fluctuations. These fluctuations are obtained by extrapolating data from figure 8, which gives a $\tilde{B}(\Gamma) \sim 0.08$, i.e. more than 50% higher than its average value.

Although the investigated quantities are non-Markovian ($\langle \tilde{X}(\Gamma) \rangle$ and $\langle \tilde{\Gamma}(X) \rangle$), their behavior is still treatable: each of them can indeed be studied separately as a one-dimensional random walk also concerning the first passage time and the maximum span walk.

On a broader level, this work provides a concrete rigorous method for quantitative studies of social-science problems. The choice of immigrant integration is motivated by its prominent place in both the EU and the US political agendas. By uncovering the local variation patterns in the quantifiers, we produced a scientific tool for anticipating the consequences of further immigration on local integration processes. Information of this type has not been available in the past and is of great value for the development of immigration policies and multi-ethnic planning at the local level. However, while this work advances our knowledge of integration phenomena, other effects, like segregation phenomena, that may spontaneously develop in the host country have yet to be considered and incorporated into the theoretical framework developed here.

Acknowledgements

This work is supported by the FIRB grants RBFR08EKEV and RBFR10N90W. EA and AB acknowledge also partial financial support by GNFM-(INdAM) via Progetto Giovani 2013 (AB) and 2014 (EA). RS is grateful to the project Competition, Adaptation and Labour-Market Attainment of International Migrants in Europe (CALMA)? granted by the VI National Plan for Scientific Research, Spanish Ministry of Economy and Competitiveness (CSO2012-38521), for partial financial support.

Appendix. Robustness checks

In this appendix we present additional measures performed on the raw data described in section 2 to further check the robustness and soundness of our approach.

First, we consider the time evolution of the ensemble average of $\Delta\Gamma$ and of ΔX; namely, at any steps n we measure the average (over all the municipalities available) of 'waiting times' and of 'step lengths', as given by

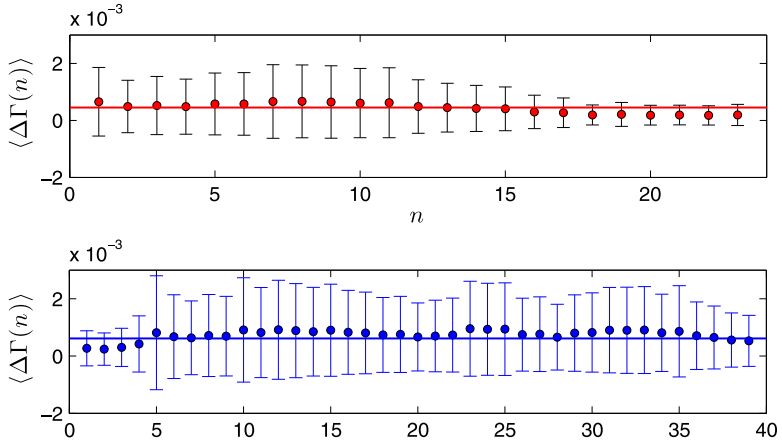

Figure 10. Ensemble averages $\langle \Delta\Gamma(n)\rangle$ obtained for time series pertaining to job (upper panel) and to family (lower panel) quantifiers, as defined in equation (A.1). In the two cases, the available trimesters (indexed by n and corresponding to the number of steps performed by the walker) are 24 and 40, respectively. Error bars correspond to standard deviations.

$$\langle \Delta\Gamma(n)\rangle = \frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}} \Delta\Gamma_n^{(i)}, \tag{A.1}$$

$$\langle \Delta X(n)\rangle = \frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}} \Delta X_n^{(i)}, \tag{A.2}$$

where X denotes any of the quantifiers J_t, J_p, M_m, and B_m, defined in equations (1)–(4), and $\mathcal{M}$ equals $\mathcal{M}_J$ or $\mathcal{M}_F$ according to the kind of quantifier (job or family, respectively) considered. In figures 10 and 11 these averages are plotted as a function of n. Apart from the case of marriages, discussed separately below, these averages are flat and scattered around the overall mean values $\mathbb{E}[\langle \Delta X\rangle]$ and $\mathbb{E}[\langle \Delta\Gamma\rangle]$, obtained by averaging (over n) $\langle \Delta X(n)\rangle$ and $\langle \Delta\Gamma(n)\rangle$, respectively. More precisely,

$$\mathbb{E}\big[\langle \Delta X\rangle\big] = \frac{1}{\mathcal{T}-1}\sum_{n=1}^{\mathcal{T}-1} \langle \Delta X(n)\rangle, \tag{A.3}$$

and similarly for $\mathbb{E}[\langle \Delta\Gamma\rangle]$. The flatness of $\langle \Delta\Gamma(n)\rangle$ and of $\langle \Delta X(n)\rangle$ suggests that, step by step, the stochastic process is homogeneous.

As for marriages, we notice that $\langle \Delta M_m(n)\rangle$ displays some degree of periodicity due to the seasonality characterising the date of celebration of marriages, as already discussed in section 2. However, if we perform a partial average on these data in such a way that we retain only one value per year, hence removing seasonality effects, we recover a flat distribution also for marriages.

The next step concerns correlations between the increments ΔX and $\Delta\Gamma$.

The possible presence of long time correlations is analyzed by measuring

$$\Xi_X(\delta) = \mathbb{E}\Big[\Big\langle \Delta X_n^{(i)} \Delta\Gamma_{n-\delta}^{(i)}\Big\rangle\Big], \tag{A.4}$$

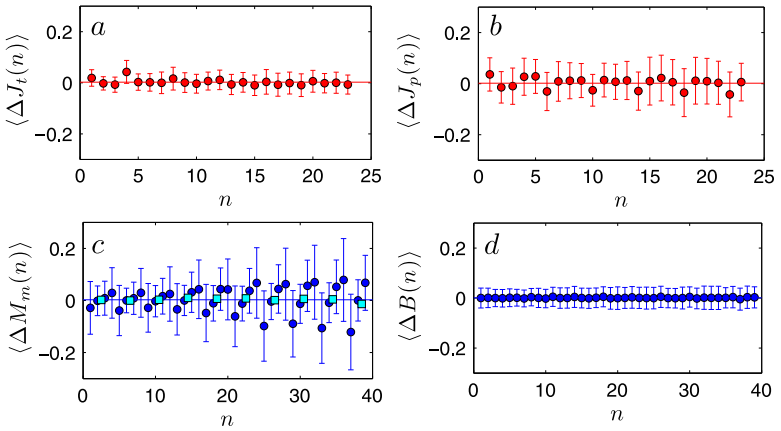

Figure 11. Ensemble averages $\langle \Delta J_t(n) \rangle$ (panel *a*), $\langle \Delta J_p(n) \rangle$ (panel *b*), $\langle \Delta M_m(n) \rangle$ (panel *c*), and $\langle \Delta B_m(n) \rangle$ (panel *d*), as defined in equation (A.2). For job quantifiers (upper panels) the available trimesters (corresponding to the number of steps performed by the walker) are 24; for family quantifiers (lower panels) they are 40. Notice that for marriages, seasonality effects emerge because marriages are more likely to be celebrated during the last two trimesters. However, if we replace the four data corresponding to the same year, we recover a flat distribution (brighter squares). Error bars correspond to standard deviations.

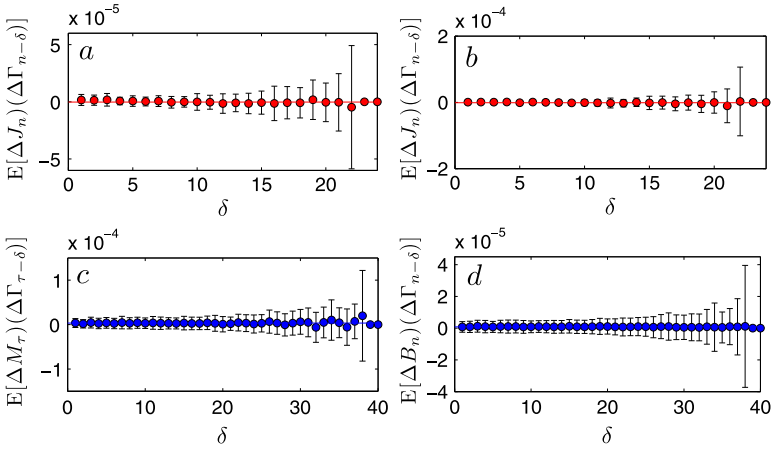

Figure 12. Long time correlations Ξ_{J_t} (panel *a*), Ξ_{J_p} (panel *b*), Ξ_{M_m} (panel *d*), and Ξ_{B_m} (panel *d*) as a function of δ.

where, again, the bracket $\langle \cdot \rangle$ denotes the average over municipalities (indexed by $i = 1, ..., \mathcal{M}$) and $\mathbb{E}$ denotes the average over steps (indexed by $n = 1, ..., \mathcal{T}$). The results obtained for all the quantifiers are presented in figure 12, showing that, in general, $\Xi_X(\delta)$ fluctuates around zero, suggesting that there is no correlation between the increments ΔX and $\Delta \Gamma$.

Moreover, we verified that two successive increments for ΔX and for $\Delta \Gamma$ are uncorrelated. This is done by considering the covariances

$$\Omega_\Gamma^{(i)} = \mathbb{E}\left\{ \left[\Delta \Gamma_n^{(i)} - \mathbb{E}\left(\Delta \Gamma^{(i)}\right) \right] \left[\Delta \Gamma_{n+1}^{(i)} - \mathbb{E}\left(\Delta \Gamma^{(i)}\right] \right\}, \tag{A.5}$$

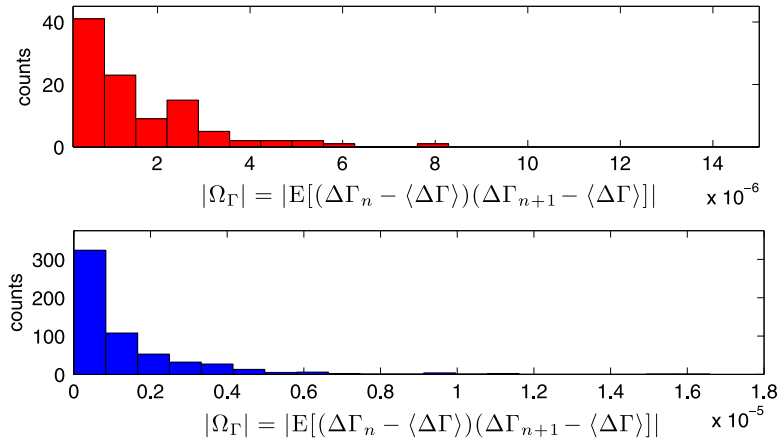

Figure 13. Histograms for short time correlations between waiting times referring to job quantifiers (upper panel) and to family quantifiers (lower panel), calculated according to equation (A.5), with $i = 1,..., \mathcal{M}_{J,F}$, where $\mathcal{M}_J = 124$ and $\mathcal{M}_F = 581$. Notice that we considered the absolute value $|\Omega_\Gamma^{(i)}|$ to better highlight the peak at zero. In any case the average correlation is, within the error, equal to zero.

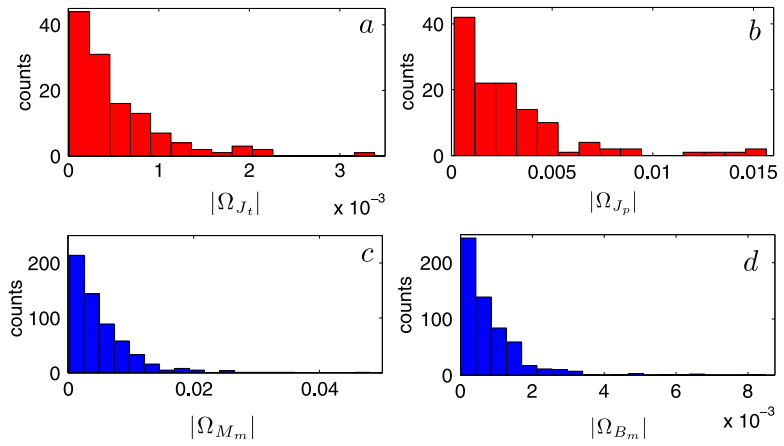

Figure 14. Histograms for short time correlations between step lengths: $\Omega_{J_t}^{(i)}$ (panel *a*), $\Omega_{J_p}^{(i)}$ (panel *b*), $\Omega_{M_m}^{(i)}$ (panel *d*), and $\Omega_{B_m}^{(i)}$ (panel *d*), calculated according to equation (A.6), with $i = 1,..., \mathcal{M}_{J,F}$, where $\mathcal{M}_J = 124$ and $\mathcal{M}_F = 581$. Notice that we considered the absolute value $|\Omega_X^{(i)}|$ to better highlight the peak at zero. In any case the average correlation is, within the error, equal to zero.

$$\Omega_X^{(i)} = \mathbb{E}\left\{\left[\Delta X_n^{(i)} - \mathbb{E}\left(\Delta X^{(i)}\right)\right]\left[\Delta X_{n+1}^{(i)} - \mathbb{E}\left(\Delta X^{(i)}\right)\right]\right\}, \tag{A.6}$$

where $\mathbb{E}(\Delta X^{(i)})$ represents the average step length along the *i*th path. Notice that in equations (A.5)–(A.6) the average is performed only over steps, hence retaining the dependence on the municipality indexed by *i*. The histograms of $\Omega_X^{(i)}$ for all the quantifiers are depicted in figures 13 and 14. Notice that, in any case, $\Omega_X^{(i)}$ is mostly zero and its mean over municipalities is, within the error, equal to zero, suggesting that there is no significant correlation.

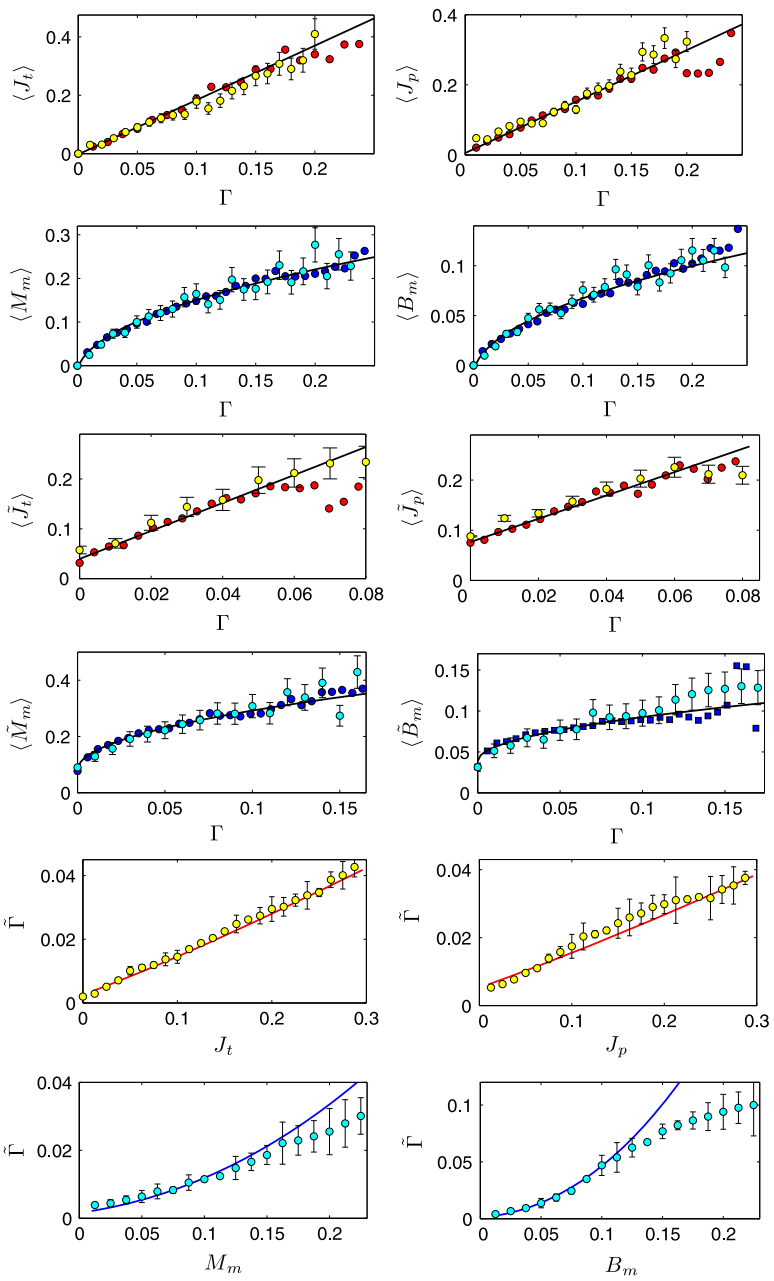

Figure 15. Comparison between simulated quantities (symbols in brighter colors) and quantities obtained from real data (symbols in darker colors); the latter are the same already reported in figures 4, 8, and 9, respectively. The real data for displacements (panels a–d) were used for the calibration of the distribution parameters (in particular σ^2), which are then kept fixed to recover the non-Markovian quantities (panels e–n). In general the comparison is very good, especially for small values of Γ and social quantifiers, where, indeed, the statistics are more sound.

The final step of this analysis aims to check the robustness of the bridge between the mesoscopic scale and the macroscopic scale: we simulated continuous-time random walks with step lengths and waiting times drawn from equation (28) and equations (29) and (30), respectively, and we measured the resulting displacement $X(\Gamma)$, the maximum span $\tilde{X}(\Gamma)$, and the first-passage time $\tilde{\Gamma}(X)$. Such values were then compared with those obtained from real data of social quantifiers and previously shown in figures 4, 8, and 9, respectively. As shown in figure 15, the comparison is, in general, very good. However, some remarks are in order.

In the simulations meant to recover the behavior of job quantifiers we realized $\mathcal{M}_J = 124$ random walks made of $\mathcal{T} = 24$ steps, whereas in the simulations meant to recover the behavior of family quantifiers we realized $\mathcal{M}_F = 581$ random walks made of $\mathcal{T} = 40$ steps, consistently with raw data available (see section 2).

The choice of the parameters λ_X, μ, and σ depends on the particular quantifier considered. For a given quantifier, the parameters are fixed and the simulation provides a measure for the average value, span, and fist passage time.

The parameters λ_X used for the distribution $f(\Delta X)$ (see equation (28)) are those given in table 2. The parameters μ and σ used for the distribution $\psi(\Delta\Gamma)$ (see equations (29)–(30)) are only close to those given in table 3. More precisely, for family quantifiers we used $\mu = -5.9$ and $\sigma^2 = 11.2$ (to be compared with $\mu = -6.6 \pm 0.9$ and $\sigma^2 = 0.32 \pm 0.04$) and for job quantifiers we used $\mu = 7.9 \times 10^{-4}$ and $\sigma^2 = 9.\,810^{-6}$ (to be compared with $\mu = (1.2 \pm 0.2)10^{-3}$ and $\sigma^2 = (6.7 \pm 0.6)10^{-6}$). The discrepancy is more pronounced as far as the variance of waiting times for social quantifiers is concerned. We argue that this is due to the fact that the log-normal fit nicely captures the long tail of the distribution, yet the head of the distribution as well as any cut-off effects are not accounted for. Admittedly, the fits provided in section 4.1 are meant to highlight a qualitative difference in the two classes of quantities. Indeed, despite a quantitative adjustment in these parameters, the distributions outlined turn out to successfully mimic the true behavior. Moreover, we stress that such an adjustment does not impair the predictive power of the theory. In fact, the parameters μ and σ were calibrated over the average displacements (which are assumed to be available), whereas the maximum span and the mean first-passage time were derived without any further parameter tuning.

References

[1] European Commission 2010 *Handbook on Integration for Policy-makers and Practitioners* (Luxembourg: Publications Office of the European Union).

[2] Castles S and Miller M J 2009 *The Age of Migration–International Population Movements in the Modern World* (New York: Pallgrave McMillian)

[3] Barra A, Contucci P, Sandell R and Vernia C 2014 *Sci. Rep. Nature* **4** 4174

[4] Weiss G 1994 *Aspects and Applications of the Random Walk* (Amsterdam: North-Holland)

[5] Klafter J and Sokolov I 2011 *First Steps in Random Walks* (Oxford: Oxford University Press)

[6] Massey D and Zenteno R 1999 *Proc. Natl Acad. Sci.* **96** 5328

[7] Wilson K L and Portes A 1980 *Am. J. Sociol.* **86** 295

[8] Sandell R 2012 *Int. Migrat. Rev.* **49** 971

[9] Montroll E W and Schlesinger M F 1984 *Nonequilibrium Phenomena II: From Stochastics to Hydrodynamics* (Amsterdam: North-Holland)

[10] Hughes B, Montroll E and Schlesinger M 1982 *J. Stat. Phys.* **28** 111

[11] Bouchaud J and Potters M 2003 *Theory of Financial Risk and Derivative Pricing* (New York: Cambridge University Press)
[12] Gabel A and Redner S 2012 *J. Quant. Anal. Sports* **8** 1416
[13] Klages R, Radons G and Sokolov I M (ed) 2007 *Anomalous Transport: Foundations and Applications* (Berlin: Wiley)
[14] Metzler E B R and Klafter J 1999 *Phys.* A **266** 343
[15] Majumdar S N 2010 *Phys.* A **389** 4299
[16] Redner S 2001 *A Guide to First-Passage Processes* (London: Cambridge University Press)
[17] Metzler S R R and Oshanin G (ed) 2013 *First Passage Phenomena and Their Applications* (Singapore: World Scientific)
[18] Montroll E and Weiss G 1965 *J. Math. Phys.* **6** 167