OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Inferring the Structure of Social Contacts from Demographic Data in the Analysis of Infectious Diseases Spread

Laura Fumanelli[1]*, Marco Ajelli[1], Piero Manfredi[2], Alessandro Vespignani[3,4,5], Stefano Merler[1]

1 Bruno Kessler Foundation, Trento, Italy, **2** Department of Statistics and Mathematics Applied to Economics, University of Pisa, Pisa, Italy, **3** Department of Health Sciences and College of Computer and Information Sciences, Northeastern University, Boston, Massachusetts, United States of America, **4** Institute for Quantitative Social Sciences at Harvard University, Cambridge, Massachusetts, United States of America, **5** Institute for Scientific Interchange Foundation, Turin, Italy

Abstract

Social contact patterns among individuals encode the transmission route of infectious diseases and are a key ingredient in the realistic characterization and modeling of epidemics. Unfortunately, the gathering of high quality experimental data on contact patterns in human populations is a very difficult task even at the coarse level of mixing patterns among age groups. Here we propose an alternative route to the estimation of mixing patterns that relies on the construction of virtual populations parametrized with highly detailed census and demographic data. We present the modeling of the population of 26 European countries and the generation of the corresponding synthetic contact matrices among the population age groups. The method is validated by a detailed comparison with the matrices obtained in six European countries by the most extensive survey study on mixing patterns. The methodology presented here allows a large scale comparison of mixing patterns in Europe, highlighting general common features as well as country-specific differences. We find clear relations between epidemiologically relevant quantities (reproduction number and attack rate) and socio-demographic characteristics of the populations, such as the average age of the population and the duration of primary school cycle. This study provides a numerical approach for the generation of human mixing patterns that can be used to improve the accuracy of mathematical models in the absence of specific experimental data.

Citation: Fumanelli L, Ajelli M, Manfredi P, Vespignani A, Merler S (2012) Inferring the Structure of Social Contacts from Demographic Data in the Analysis of Infectious Diseases Spread. PLoS Comput Biol 8(9): e1002673. doi:10.1371/journal.pcbi.1002673

Editor: Marcel Salathé, Pennsylvania State University, United States of America

Received March 21, 2012; **Accepted** July 22, 2012; **Published** September 13, 2012

Funding: This work has been partially funded by the EC-ICT contract no. 231807 (EPIWORK) to LF, MA, SM and AV, the National Institutes of Health R21-DA024259 award to AV and the European Centre for Disease Prevention and Control (ECDC) grant 2009/002 to MA and PM. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lfumanelli@fbk.eu

Introduction

The accurate characterization of the structure of social contacts in mathematical and computational models of infectious disease transmission is a key element in the assessment of the impact of epidemic outbreaks and in the evaluation of effective control measures. For instance, the transmissibility potential of a disease and the final epidemic size strongly depend on mixing patterns between individuals of the population, which in turn depend on socio-demographic parameters (e.g. household size, fraction of workers and students in the population) [1–5]. For this reason, several efforts have been recently carried out in order to obtain contact data with the aim of quantifying "who meets whom (where, when, how long and how often)" [6–12], possibly also over time [13,14]. Empirical data collection on a large scale is however extremely difficult and although several models tackling both new emerging epidemics and endemic diseases have introduced a significant amount of information on contact patterns [3,5,15–31], it is clear that the increasing use of data-driven models in the support of public health decisions is calling for novel approaches to the estimation of mixing patterns in human populations.

In this study we propose to overcome the above challenges by developing a general computational approach to derive mixing patterns from routinely collected socio-demographic data. In particular we focus on contact matrices by age of 26 European countries for which we are in the position to construct a synthetic society in the computer by integrating available social and census data. The use of contact matrices is the simplest way to improve on the homogeneous mixing assumption while at the same time preserving the analytical transparency of the model. The proposed approach is based on the simulation of a virtual society of agents that allows the estimate of contact matrices by age in different social settings: household, school, workplace and general community. Unlike classical agent based approaches of epidemic transmission [3,5,15,17,19,32] and network models [33,34], which are aimed at characterizing the spatio-temporal spread of epidemics tagging each individual in the population with a set of social attributes, we use the same detailed information on social contacts to construct contact matrices by age in the different settings to be used in compartmental models. This approach integrates population details, providing an effective description of the population structure to be used in computational models relying on compartmental schemes both at the continuous and individual based scale. Such a strategy might be very convenient to reduce the computational time demand in the analysis of large scale geographical models [21,35–39].

Author Summary

The dynamics of infectious diseases caused by pathogens transmissible from human to human strongly depends on contact patterns between individuals. High quality observational data on contact patterns, usually presented in the form of age-specific contact matrices, are difficult to gather and are currently available only for few countries worldwide. Here we propose a computational approach, based on the simulation of a virtual society of agents, allowing the estimation of contact patterns by age for 26 European countries. We validate the estimated contact matrices against those obtained by the most extensive field study on contact patterns, with data collected in eight European countries. We show that our contact matrices share some common features, e.g. individuals tend to mix preferentially with individuals their own age, and country-specific differences, which can be partly explained by differences in population structures due to different demographic trajectories followed after WWII. Our analysis highlights well defined correlations between epidemiological parameters and socio-demographic features of the populations. This study provides the first estimates of contact matrices for many European countries where specific experimental data are still not available.

Those matrices are appropriately combined in order to obtain the overall "adequate" total contact matrix for influenza-like-illness. In order to validate the proposed approach we compare the obtained contact matrices by age with the results of the Polymod study [9], the first large-scale survey on social mixing patterns relevant to infectious disease transmission. We show that the synthetically generated matrices share several common features with the Polymod matrices, e.g. strong assortativeness and the presence of similar secondary diagonal contact patterns. We integrate the synthetic contact matrices in a simple model for acute infectious diseases and highlight the role played by social and demographic factors in determining the different epidemic patterns in different countries. Further analysis and validation on the derived contact matrices is provided by investigating seroprevalence data for the 2009 H1N1 influenza pandemic in the UK.

The proposed method is extremely general and can be readily exported to other countries in the world for which the necessary social and demographic data can be gathered. We consider this approach an important step in order to overcome the current difficulties in real data gathering. Furthermore the computational path to the estimate of contact matrices represents a convenient scheme for the introduction of detailed individual based information in a wide range of modeling approaches working at the population level. For this reason we publicly release the entire collection of contact matrices to the scientific community (see Table S1 and S2; public download is also available at http://www.epiwork.eu/resources/matrices).

Materials and Methods

Socio-demographic data

In order to provide a quantitative estimate of contact matrices for 26 European countries we used highly detailed data on the country-specific socio-demographic structures (e.g., household size and composition, age structure, rates of school attendance, etc.) available at the Statistical Office of the European Commission [40]. These data were used to generate highly detailed synthetic populations for all countries of the study area. More specifically, census data on frequencies of household size and type, age of household components by size were used to group individuals into households. Data on rates of employment/inactivity and school attendance by age, structure of educational systems, school and workplace size allowed us to either assign individuals to schools and workplaces or tag them as inactive, according to their age.

The procedure to generate the synthetic populations is quite standard in the context of individual based models and is therefore discussed in detail in Text S1. In the following paragraph we present the approach used in the computation of the synthetic contact matrices.

Computing contact matrices

The mathematical representation of epidemics relies on the description of the transmission process which is usually modeled through the force of infection, that is the rate at which a susceptible individual acquires the infection because of the interactions with infectious individuals. This quantity is proportional to the number of infectious individuals, the specific transmission probability of the infection during a contact and the overall rate of contacts of each individual with other individuals in the population. Although a vast majority of studies assumes the population as homogeneous –all individuals are equal with same average contact rate– the social and demographic structure of the population is generally reflected in heterogeneous contact patterns among individuals. Age is obviously one of the main determinants of the mixing pattern of individuals. Children tend to spend more time with children and members of their household, active adults mix with individuals in their workplace etc. Mixing patterns by age are generally defined by a contact matrix whose elements M_{ij} represent the average frequency of "adequate" contacts that an individual of age i has with individuals aged j. We define a contact as "having shared the same physical environment" [11] (e.g., the same household, or the same school or workplace). To compute age-specific contacts we postulate that at the finest scale of the single units, e.g. single households or schools, mixing is homogeneous. This hypothesis is necessary given the lack of information on contacts at this fine scale. By aggregating units we then compute "setting-specific" contact matrices, represented by the four matrices accounting for contacts within household members (matrix H), within schoolmates/teachers (matrix S), within workplace colleagues (matrix W) and in the general community (matrix R). Finally, the overall age-specific matrix M is computed as a linear combination of the four matrices H, S, W, R.

To give an example, let us see in more detail the computation of the matrix of contacts within households, H. For each individual k of age i, living in household $h^{(k)}$ of size $v_H^{(k)}>1$, the household contacts with individuals of age j are defined as the set of individuals of age j living in household $h^{(k)}$. We denote the number of individuals in this set by $h_j^{(k)}$. Then, once $h_j^{(k)}$ is computed, we obtain the probability that individual k has contacts with individuals of age j by dividing the number of contacts with members of age j by $v_H^{(k)}-1$, which represents the number of individuals living in the same household as k. The expression for the frequency of contacts within households H between individuals of ages i and j is then

$$f_{ij}^H=\begin{cases}\dfrac{1}{n_i^H}\displaystyle\sum_{\substack{1\le k\le N_i\\ v_H^{(k)}>1}}\dfrac{h_j^{(k)}-\delta_{ij}}{v_H^{(k)}-1} & \text{if } n_i^H>0\\ 0 & \text{otherwise,}\end{cases}$$

where N_i is the total number of individuals of age i in the population, n_i^H is the number of individuals of age i with at least one contact in the household (note that some individuals may have zero contacts) and δ_{ij} is the Kronecker delta function that allows excluding individual k from the set of her/his own contacts. A straightforward example of the computation of household contacts frequencies is provided in Figure 1.

In order to transform the frequency of contacts into contact matrices relevant for infectious disease spreading, we need to consider the following quantities:

- $p_i^H = \frac{n_i^H}{N_i}$ which is the probability for age bracket i to have at least one contact in the household. This probability is less than one as individuals may live alone;
- N_i^H which is the rate of total effective contacts in the household for individuals of age i. This number depends on the time scale, the type of disease etc.

We can therefore define the synthetic contact matrix for households

$$M_{ij}^H = p_i^H f_{ij}^H,$$

which provides the relative frequency of contact among age classes. In order to obtain the rate of effective contacts between classes i and j we need then to consider the product $N_i^H M_{ij}^H$. In the following, in the lack of a better knowledge, we will assume that the effective contact rate is age independent: $N_i^H = N^H$. Expressions analogous to the previous one can be used to compute the frequency of contacts by age within schools (superscript S) and within workplaces (superscript W); the matrix for schools M^S is given by the sum of the matrices for each school level, from pre-primary to higher education. As regards the frequency of contacts in the general community, we assume homogeneous mixing among individuals, thus the columns of the matrix (superscript R) are proportional to the number of individuals by age.

In order to define the "adequate" contact matrix [41] we assume that the total rate of effective contacts in the households can be expressed as $N^H = N_{tot}\alpha^H$, where N_{tot} is the total rate of effective contacts and α^H is a constant providing the fraction of effective contacts within the household. Similar expressions can be written for all other settings. Since there is evidence that infection transmission is not uniform by setting [42,43] and the relevance of every setting (household, school, work, general community) in the transmission depends on the pathogen responsible for the disease, in principle it is possible to estimate empirically the fraction of transmission events ρ_K in each setting $K \in \{H,S,W,R\}$. In this study, in order to give a baseline, we assume values for influenza-like-illness (ILI) transmission following empirical estimates of the proportions of transmission in the different settings [5,42,44–46]: 0.3 in households, 0.18 in schools, 0.19 in workplaces and 0.33 in the general community. Although these weights are related to influenza, they have already been used in [11] to generate a synthetic contact matrix for the Italian population, which has been shown to capture Varicella and Parvovirus B19 data.

For each setting K we obtain the condition

$$N_{tot}\alpha^K \sum_{i=0}^{\omega}\sum_{j=0}^{\omega} M_{ij}^K = N_{tot}\varrho_K,$$

where ω is the maximum age of the population. This set of equations readily provides the values

$$\alpha^K = \frac{\varrho_K}{\sum_{i=0}^{\omega}\sum_{j=0}^{\omega} M_{ij}^K}.$$

The resulting adequate contact matrix M is therefore defined as the linear combination of the contact matrices in each setting:

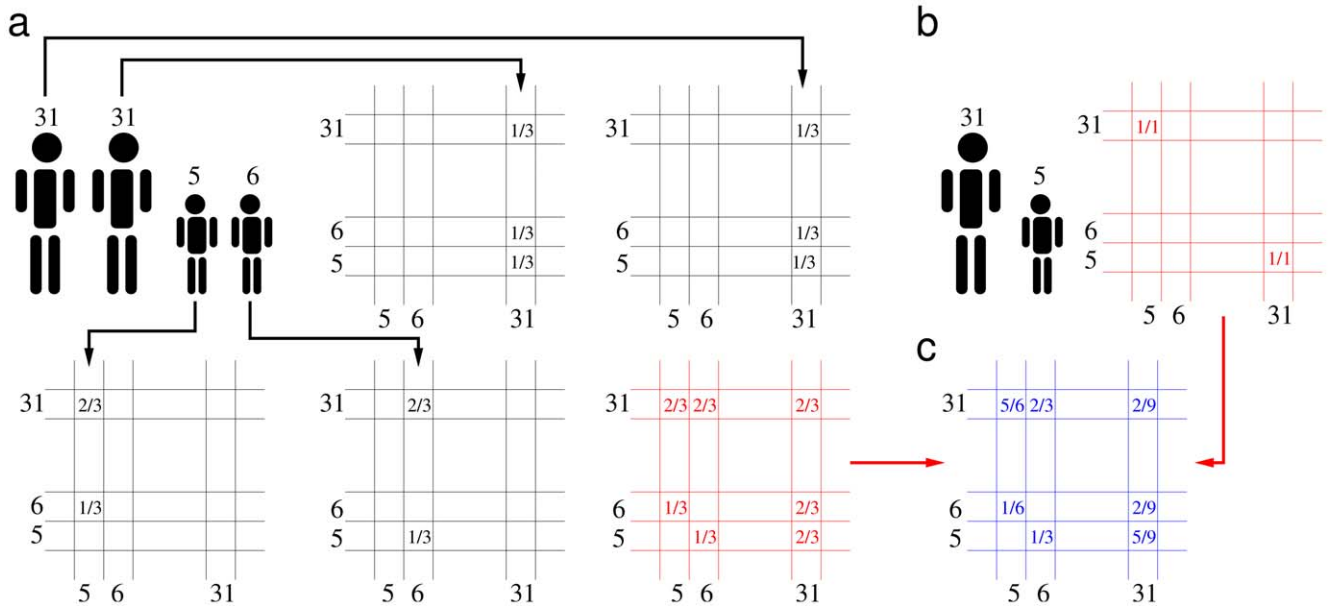

Figure 1. Example of the computation of household contact matrix. a Computation of contact frequencies for every member of a household composed by two adults aged 31 and two children of 5 and 6 years old. The sum of the four contributions gives contact frequencies within this household (in red). **b** Contact frequencies within a household composed of an adult aged 31 and a child aged 5. **c** Assuming that these two households constitute the whole population, the frequency of household contacts that individuals of age i have with individuals aged j is given by the sum of the contributions from each household, divided by the number of individuals aged i having at least one household contact.
doi:10.1371/journal.pcbi.1002673.g001

$$M_{ij} = \sum_K \alpha^K M_{ij}^K .$$

This notion of adequate matrix of total contacts appears to be appropriate, up to a scale factor, when the transmission coefficients of the infection depend only on settings (i.e. they are age-independent), and provided that the proportions of transmission in the different settings are roughly constant during the course of the epidemic.

Matrix M defines the contact pattern among ages up to a constant N_{tot} that has to be considered as an appropriate rescaling factor when comparing matrices defined according to different time scales or data aggregation processes. In the study of epidemic processes, by assuming that the probability of transmission q per effective contact is constant, the contact rate N_{tot} is usually absorbed in the definition of the transmissibility rate $\beta = qN_{tot}$ that is used as the scaling factor determining the reproduction number R_0 that characterizes the specific pathogen transmission. R_0 essentially represents the average number of secondary cases generated by a primary case in a completely susceptible population [47], and it is therefore the threshold parameter determining the dynamics of the epidemic.

Remarkably, our matrices are computed by considering one-year age brackets, from 0 to 100 and over years; this is the most refined version of our data on frequencies of contacts. They can however be aggregated in different ways, depending on the purpose for which they are used: for instance, for childhood diseases one may prefer to group contact data for children according to educational levels.

Although we are dealing with very detailed data on the socio-demographic structures of European countries, there are a number of limitations and assumptions that it is worth stating. First of all, although the relevant statistics could be gathered from other sources, we consider Eurostat as the only source of data on occupation rates. This is the reason why we decided to exclude Belgium, Poland and Malta from our study in view of the incomplete information on employment and schooling rates. Furthermore, the different household structures considered in our virtual society cover about the 95% of the total number of households in Europe. We do not allow however families with an aggregate member or non-private households (such as rest homes, dorms, religious and military institutions). Finally, another limitation lies in the assumption of homogeneous mixing for the contacts occurring in the community at large (i.e., not occurring between household members, schoolmates and work colleagues). In fact, this implies to disregard any kind of preferential mixing, e.g. by age, and the level of activity of individuals, which may vary by age, as documented by Polymod data [9]. Clearly, the availability of more precise and complete data on any aspect of the socio-demographic structure of a population (e.g., number and composition of non-private households; size and attendance of nurseries) would allow a refinement of our virtual society.

SIR model with heterogeneous mixing patterns

In the classic SIR model the population is divided into three compartments: susceptible (individuals that can acquire infection), infectious (individuals that have been infected and are able to transmit the pathogen) and recovered (individuals that are immune to the disease–e.g. because they recovered from the infection). In order to include the mixing patterns encoded in the contact matrices, each group is characterized by an age structure. Every susceptible individual of age i (belonging to the S_i group) experiences an age-specific force of infection λ_i, which is determined by the average frequency M_{ij} of "adequate" contacts that an individual of age i has with individuals aged j, by the probability of contacting infectious individuals from every age class j, and by a transmissibility β_i that accounts for the probability of infection per contact. The force of infection yields the rate of transition of susceptible individuals into the infectious state I_i; individuals then leave this status according to the recovery rate γ_i (the inverse of the duration of the infectious period), entering the recovered compartment R_i. For the sake of simplicity we consider age independent transmissibility $\beta_i = \beta$ and recovery rate $\gamma_i = \gamma$. The set of equations governing the SIR model can be thus written as

$$\begin{aligned} \dot{S}_i &= -\sum_{j=1}^{n} \beta M_{ij} \frac{I_j}{N_j} S_i \\ \dot{I}_i &= \sum_{j=1}^{n} \beta M_{ij} \frac{I_j}{N_j} S_i - \gamma I_i \\ \dot{R}_i &= \gamma I_i \end{aligned} \qquad (1)$$

where N_j is the number of individuals in the population of age j and n is the number of age classes considered. R_0 is calculated as the spectral radius of the next generation matrix [48], that is $R_0 = \varrho(\beta\gamma^{-1}M)$.

To analyze post-pandemic H1N1 serological data collected in England and Wales in fall 2009 [49], we make use of a slightly modified version of model (1) accounting for age-specific susceptibility to infection, which is acknowledged as a further critical determinant of the force of infection for influenza. In particular, the equations for susceptibles and infectious become:

$$\begin{aligned} \dot{S}_i &= -\rho_i \sum_{j=1}^{n} \beta M_{ij} \frac{I_j}{N_j} S_i \\ \dot{I}_i &= \rho_i \sum_{j=1}^{n} \beta M_{ij} \frac{I_j}{N_j} S_i - \gamma I_i \end{aligned}$$

where $\rho_i = 1.0$ if $i \leq 15$, 0.5 otherwise as resulting from estimates reported in [50,51].

Results/Discussion

Synthetic contact matrices by age

Figure 2 shows the contact matrices obtained for the United Kingdom. They present a number of clear features that reflect the socio-demographic structure of the population: i) the matrix for households (Figure 2a) shows a dominant diagonal, representing contacts with siblings for young individuals, and with spouse for adults, whose ages are generally similar. There are also a lower and an upper diagonal, accounting for contacts that parents have with their children and vice versa; these contacts are generally absent for people aged over 60; ii) the structure of the educational system is clear from the matrix of contacts at school (Figure 2b) as young individuals mix mostly with persons of similar age, belonging to same school level; iii) in the workplaces (Figure 2c) most contacts occur between people from 20 to 60 years old, corresponding to the working age population. Finally, the matrix for contacts in the community, obtained by assuming homogeneous mixing, reflects the age structure of the overall population only (Figure 2d).

In Figure 2e the "adequate" contact matrix is reported; for young individuals, contacts within schoolmates of similar age are represented on the main diagonal; mixing in workplaces is prevalent for adults aged 20 to 60 years; people aged more than 65 years have most contacts with people of similar age. The

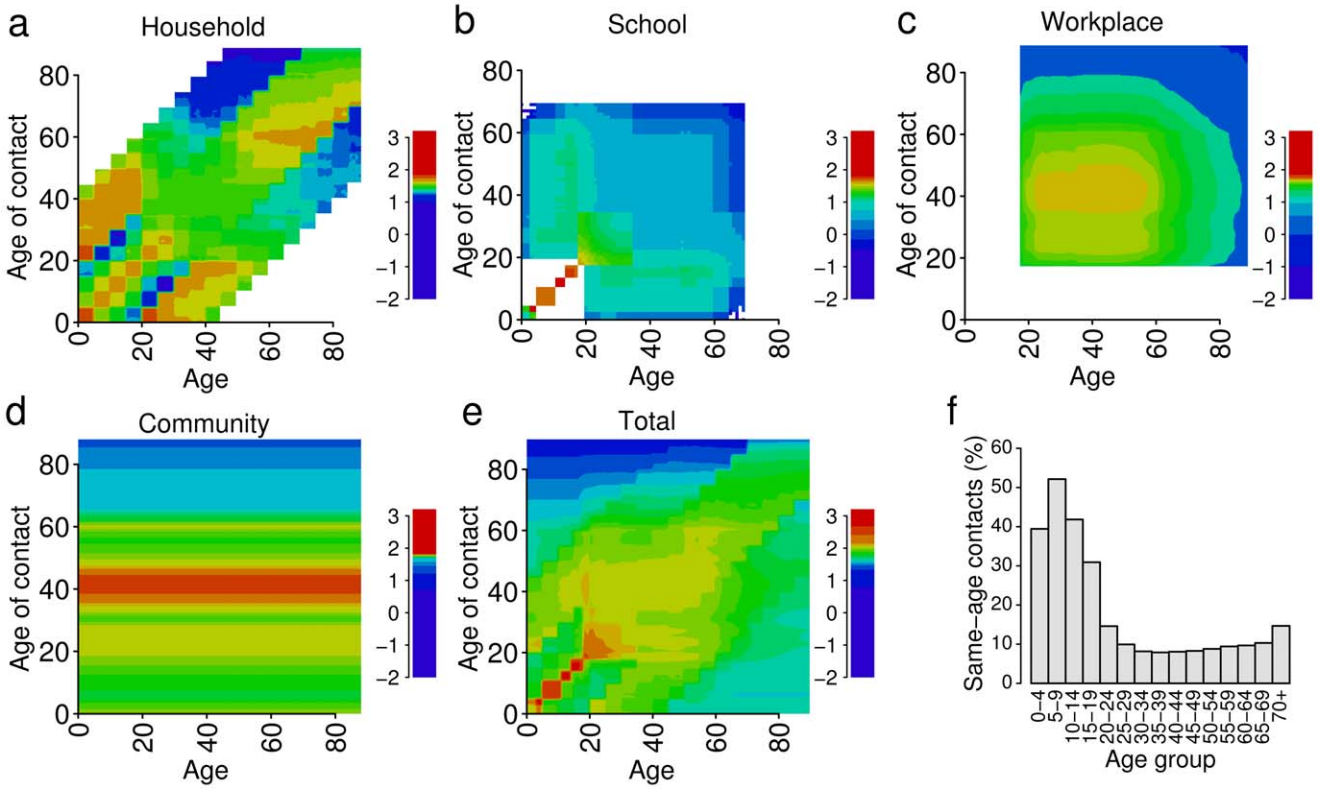

Figure 2. Mixing patterns by age in the UK. Representations in logarithmic scale of contact matrices by one-year age brackets for the United Kingdom in the different social settings. Frequency of contacts (in arbitrary units) increases from blue to red. **a** Household. **b** School. **c** Workplace. **d** General community. **e** The total matrix obtained as a linear combination of the matrices represented in **a–d**; the coefficients used are the proportions of transmission in the four settings: 0.3 in households, 0.18 in schools, 0.19 in workplaces and 0.33 in the general community [3,11,42,44–46]. **f** Proportions of contacts with individuals of the same age group, from the total matrix.
doi:10.1371/journal.pcbi.1002673.g002

difference between the upper left and the lower right entries of the matrix is a consequence of the age structure of the population, which is characterized by a small fraction of individuals aged over 80. Figure 2f reports the proportions of same-age contacts by 5-year age groups of this matrix: from 40 to more than 50% of the contacts of individuals under 15 years are within the same age group, basically with schoolmates and siblings at home. Adults contacts instead are much less assortative because the working period spans a wide range of ages.

Country-specific features of synthetic contact matrices

Although similar attributes can be observed in the synthetic contact matrices for all 26 countries under consideration (the representation of all matrices other than UK are reported in Text S1), some features are distinctive of specific regions. For instance, household mixing is always characterized by the three diagonals representing contacts between siblings/spouse or parents and children, but in Northern Europe and France the upper diagonal is shorter, reflecting the fact that people tend to leave home earlier than in the other countries. The contribution of educational systems is always represented on the lower part of the main diagonal: the observed heterogeneity among countries is driven by the organization of school cycles (e.g. the arrangement of primary and lower secondary schools into either single or separate structures is clearly visible). The central part of the matrices is associated to contacts at work, which depend on the age structure of the working population. Moreover, we can observe that mixing with individuals of about 60 years of age tends to be higher in Northern countries and lower in the others, particularly in Southern Europe: this is probably due to differences in retirement age [52].

In order to infer more rigorously whether similarities between contact matrices can be identified to characterize specific groups of countries, we use a hierarchical cluster algorithm. The algorithm uses the average dissimilarity between two matrices x and y (treated as vectors) as measured by the Canberra distance $d(x,y)=\sum_{k=1}^{n}\frac{|x_k-y_k|}{|x_k|+|y_k|}$; this choice was made because the entries of contact matrices range over several orders of magnitude, and this distance, differently from L_1 and L_2 distances, is appropriate to measure average relative dissimilarities rather than absolute dissimilarities [53]. We found that contact matrices can be clustered in a way mainly reflecting the geographical location of the country. This may be motivated with the observation that neighboring states, for historical and cultural reasons, show marked similarities in school organization and demographic structures. The latter are well explained by the common demographic trajectories followed after World War II with respect to major structural changes, from the baby boom in the Sixties, to the fall towards low fertility [54].

In particular, we isolated four main clusters (see Figure 3): Ireland and Cyprus, Eastern, Southern and Northern Europe. This grouping can be partly explained in terms of some macroscopic indicators such as average age (which is a proxy for the number of students in the population), household size (see inset of Figure 3) and school organization. For instance, Ireland and

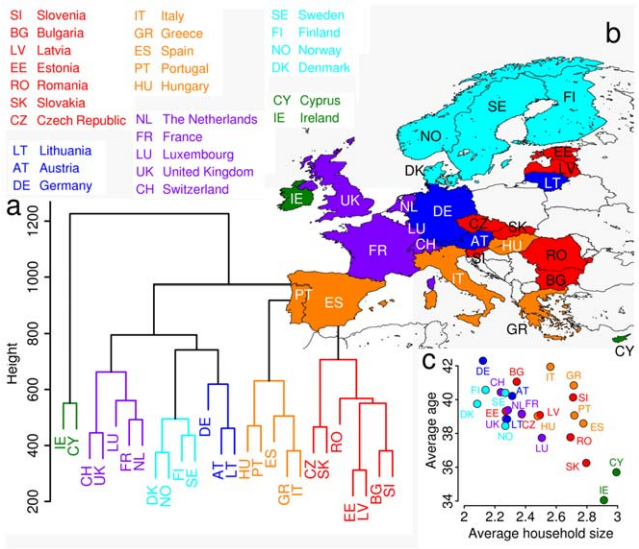

Figure 3. Characterization of synthetic contact matrices. Clustering of countries on the basis of total matrices. **a** Dendrogram of cluster analysis based on the Canberra distance. **b** Map of Europe and grouping of countries made by the algorithm; countries having the same color belong to the same cluster. **c** Average age and household size for the 26 countries considered. Colors as in the map.
doi:10.1371/journal.pcbi.1002673.g003

Cyprus are the youngest European countries: the average ages are 34.1 and 35.7 years respectively, while the overall average age in Europe is 39.2. Moreover, they have the largest household size (2.88 individuals for Ireland, 2.98 for Cyprus, 2.49 for Europe). All Scandinavian countries (Norway, Sweden, Finland, Denmark) are grouped into a unique sub-cluster which is characterized by large average age and small household size. Eastern and Southern Europe are similar in terms of average age and household size, but in Southern countries elementary and lower secondary schools are organized as separate structures, whereas a unique cycle is predominant in Eastern countries. Interestingly, in Eastern Europe two sub-clusters can be identified: Czech Republic-Slovakia, which became two independent states few years ago, and Estonia-Latvia. Lithuania instead is grouped with Germany and Austria: this distinction may be due, at least partially, to Lithuanian school organization, which is similar to Central and Western Europe.

Validation with Polymod contact matrices

In order to validate the data driven modeling approach at the origin of the synthetic contact matrices we compared our matrices with those obtained by the Polymod project [9]. Polymod currently represents the most accurate and extensive study on mixing patterns in Europe. We considered only the subset of countries common to both our approach and the survey study, namely Germany, Finland, United Kingdom, Italy, Luxembourg and the Netherlands. By jointly regressing the matrices for all countries, each of them normalized so that the sum of all its elements is one, the value of the coefficient of determination R^2 for the linear regression model is 0.72. However, we found that the estimated value of the intercept is very close to 0 ($-2.9\cdot10^{-4}$, p-value = 0.01). Thus we fit a linear model with zero intercept and a single slope coefficient (Figure 4a), concluding that most statistical variation between Polymod matrices and ours can be captured by a single scale factor. Taking every country singularly and applying the linear model with zero intercept to the original matrices (without normalization), we got a coefficient of determination ranging from $R^2 = 0.59$ in Germany to $R^2 = 0.81$ in the United Kingdom (Figure 4b). The estimated scale factor depends on the considered country, because of the large variability in the average number of contacts of an individual in the Polymod matrices. A further comparison highlighting the similarities between Polymod and our contact matrices is shown in Text S1.

Contact matrices and seroprevalence profiles

Another way to assess and validate our approach consists in the analysis of the prevalence by age profile generated by using Polymod matrices and our synthetic matrices in ILI epidemic models. As an example, we considered the epidemic prevalence generated by an age structured SIR model in a fully susceptible population (as detailed in the Materials and Methods section). The model includes the heterogeneity of contacts by age by introducing a force of infection across age groups modulated by the matrix M_{ij}. We considered a basic reproduction number $R_0 = 1.4$ and compared the results obtained by using the contact matrices derived from both the Polymod survey and our model. The results reported in Figure 4d show that Polymod matrices and synthetic matrices yield qualitatively comparable profiles of seroprevalence by age for all countries, with a few noticeable deviations in Germany and the Netherlands (i.e. the countries having the lowest values of R^2 for the linear model, see Figure 4b). As a double check with regards to the Netherlands, we performed the simulations also by considering the contact matrix reported in [7] and obtained by a survey prior to the Polymod one. In this case the results are in good agreement for age groups under 60 years old, while for elderly individuals our matrix gives results closer to those obtained by using the Polymod matrix.

Furthermore, in order to validate numerical simulations against empirical data, we compared predictions of our and Polymod contact matrices to seroprevalence data collected in England and Wales at the end of the second wave of the 2009 H1N1 pandemic influenza [49]. We simulated this epidemic by using both the Polymod and our contact matrix for the United Kingdom. An age-specific susceptibility to infection was assumed; however, this parameter has not been fitted to epidemic data, but its value has been set to 2.0 for children under 16 years of age, following the estimate in [50] later confirmed in [5,51]. The only degree of freedom in the fit is therefore represented by the scale factor which can be tuned to obtain different values of R_0. The results are shown in Figure 4c: our model is able to reproduce well the seroprevalence of individuals in the classes 0–4, 5–14 and 65+ years old, while it underestimates intermediate age groups, where the Polymod matrix instead performs slightly better. Overall, model simulations belong to the 95% confidence interval of serological data in five of the six age groups considered in [49].

It is worth remarking that profiles predicted by employing our matrices are smooth because the proportions of contacts are derived from the entire simulated population. Polymod matrices instead are based on the observation of a sample of the population, and this leads to a less regular seroprevalence profile (as can be seen for instance for the Netherlands, where prevalence for individuals aged 19–29 appears to be much lower than for the adjoining age groups). More in general, the prevalence predicted from the synthetic mixing patterns is higher among school-age children, intermediate for working ages and progressively declining in the elderly; prevalence among little children is at an intermediate level. This pattern is mainly driven by country-specific employment and schooling rates, along with the scholastic organization. Simulated seroprevalences, using our contact matrices in the same epidemic setting, for the countries not

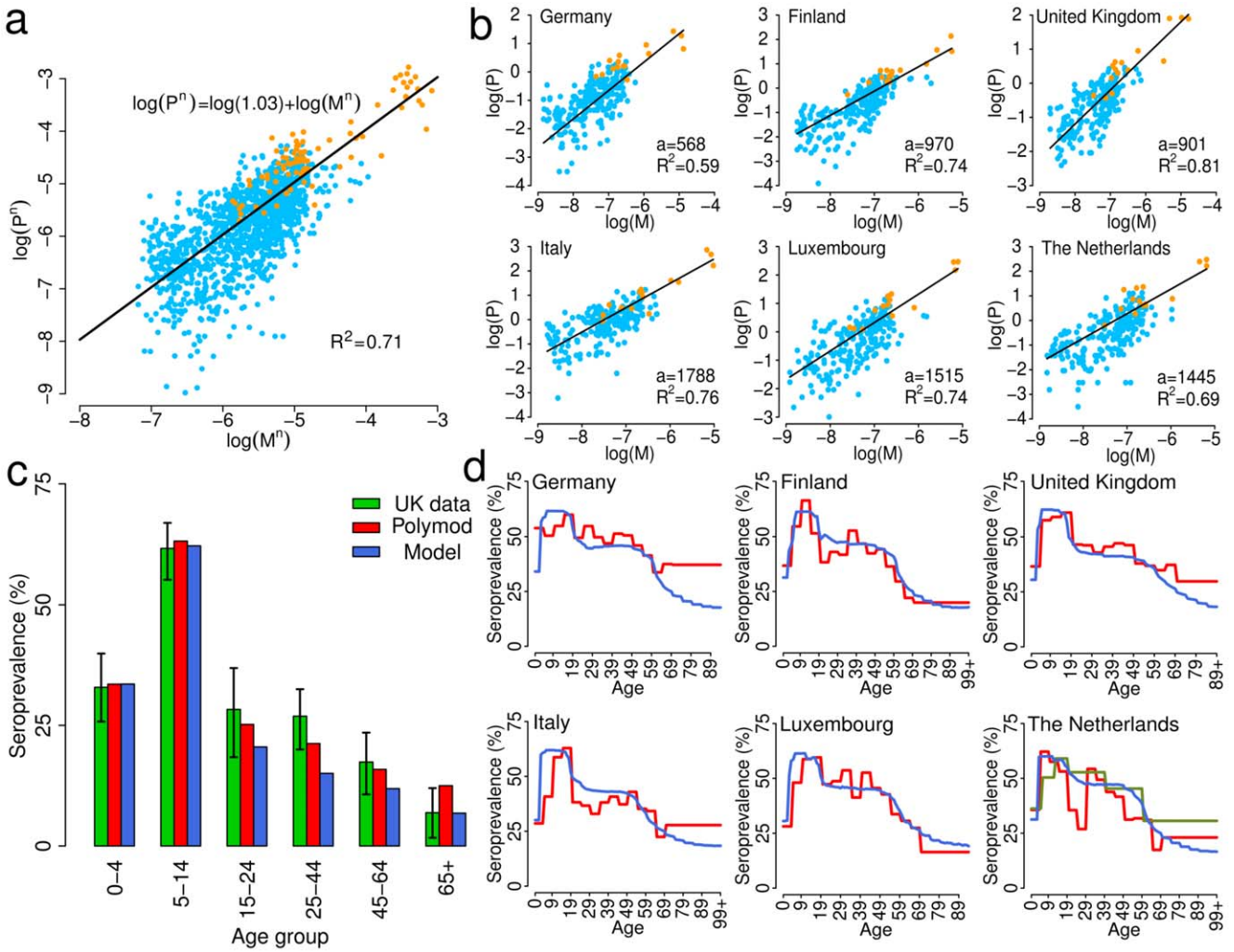

Figure 4. Comparison with Polymod contact matrices. a Linear regression model with zero intercept for Polymod matrices [9] P^n against those from our model, M^n (results shown in logarithmic scale). All countries are considered together and every matrix is normalized so that the sum of its elements is one. Yellow dots refer to the terms on the diagonal, light blue dots correspond to the other entries of the matrices. The value for the regression coefficient is 1.03 and the coefficient of determination R^2 results to be 0.71. **b** As in **a** but for each country singularly, without matrix normalization. In every plot the values for the regression coefficient a and the coefficient of determination R^2 are reported. **c** Green bars represent the average seroprevalence of H1N1 influenza infections in England and Wales during the 2009 pandemic as estimated in a serosurvey [49] (in that study a titre ≥ 32 for haemagglutination inhibition has been considered for defining seroconversion in the population) and the black lines represent the 95%CI. Blue bars represent the seroprevalence as obtained by simulating a SIR model with $R_0 = 1.46$ using our contact matrix. Red bars represent the seroprevalence as obtained by simulating a SIR model with $R_0 = 1.42$ using the Polymod contact matrix. **d** Simulated seroprevalence profiles by age. using Polymod (red) and our matrices (blue), for an epidemic emerging in a completely susceptible population, assuming $R_0 = 1.4$. In the plot for the Netherlands the profile obtained using the matrix from [7] is also shown (dark green).
doi:10.1371/journal.pcbi.1002673.g004

covered by Polymod are provided in Text S1. The shapes are all similar; however, some differences are visible: for instance, following the decline of prevalence after school age, a second steep decay occurs in all countries at a variable age, generally higher (around 60) in Northern Europe and lower (around 50) elsewhere and more markedly in Southern countries. This is probably an effect of the differences in the retirement age across Europe, as we previously pointed out.

Comparison with average European matrix

An intermediate choice between homogeneous mixing and country-specific contact patterns would be to consider mixing patterns as derived by appropriately averaging over the 26 country-specific contact matrices; therefore in this section we compare results obtained by assuming homogeneous mixing, the average European matrix and country-specific matrices. We considered an SIR model where all the basic parameters and scaling factors are set on the baseline yielding a basic reproduction number $R_0 = 1.4$ for the European average. For each country we used the synthetic contact matrices aggregated by one-year age brackets up to 84 years of age (so that all matrices have the same dimension). All other parameters being equal, the different contact matrices in each country define different values of R_0 and different epidemic behaviors in each country. In particular, the reproduction number R_0 can be calculated (see Materials and Methods section) for each contact matrix.

The improvement obtained by using either the European average matrix or country-specific matrices compared to the homogeneous model is evident e.g. in terms of final attack rate which, as already noticed in previous computational studies [26],

is always overestimated by the homogeneous assumption (Figure 5a). Although we found a strong positive correlation between R_0 and attack rate following the country-specific approach (Pearson correlation test 0.71, p-value <0.001), the result is clearly far from the homogeneous mixing model that does not consider any contact structure.

By applying to every country the average European contact matrix, large differences in terms of attack rate and peak day can be observed compared to the results obtained with the country-specific mixing patterns, especially for values of the basic reproduction number consistent with influenza epidemics (Figures 5b–c). These variations are driven only by the structure of contacts used, as the population considered is the same. In particular, depending on R_0, peak days may differ by several weeks (Figure 5c). This is clear also from Figures 5d–g, where epidemic profiles corresponding to $R_0=1.4$ for four countries are shown. Moreover, the use of the average contact matrix yields a general alignment of epidemics in the different countries, while differences in timing are clearly visible when country-specific matrices are used.

Socio-demographic structure and disease epidemiology

The synthetic contact matrices allow us to analyze the effect of the different social and demographic structure of countries on the evolution of infectious diseases characterized by the same natural history. For the sake of simplicity we considered an SIR model with basic parameters and scaling factors corresponding to $R_0=1.4$ in the UK. In each country we used the synthetic contact matrices aggregated by 5-year age brackets up to the class 70+.

The average age of the population is the single factor best explaining the basic reproduction number (correlation -0.61, p-value $=0.001$; and the linear regression model with the average age as the only independent variable gives a coefficient of determination $R^2=0.37$, p-value $=0.001$). We found that a linear model having both average age and matrix assortativeness (measured by the Q-index [11], see Text S1) as independent variables represents the best option to explain the variability of R_0 between countries: $R^2=0.69$, p-value <0.0001. However, matrix assortativeness cannot be derived directly from routinely collected data, but is characteristic of the specific contact matrices, therefore it is unknown *a priori*. Nonetheless, matrix assortativeness is strongly related to the duration of the primary school cycle (correlation -0.72, p-value <0.001). Therefore, we decided to add the duration of the primary school cycle as a proxy for matrix assortativeness in the linear model for explaining R_0 having average age as independent variable; this model gives $R^2=0.46$ and the analysis of variance reveals a statistically significant improvement (p-value $=0.05$) with respect to the model considering the average age as the only independent variable (see Figure 6a). As regards the final attack rate, we found that the best single socio-demographic factor explaining the variability

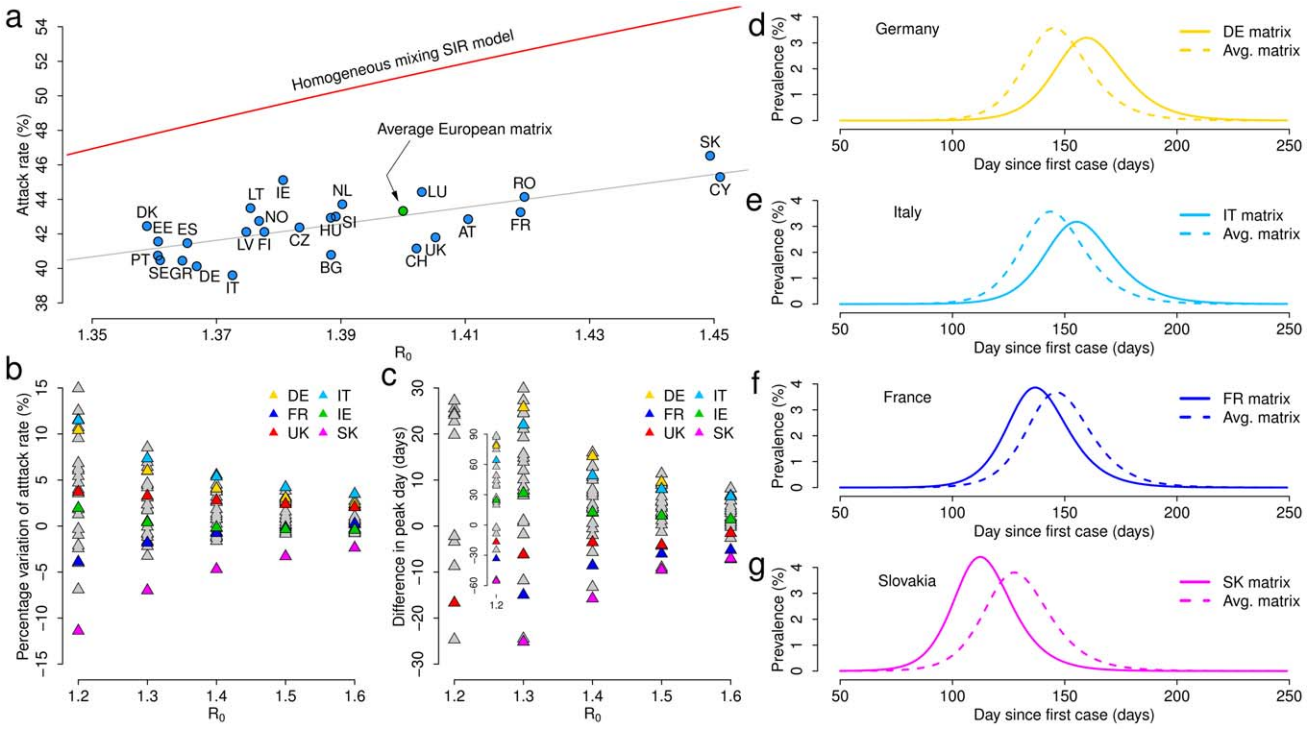

Figure 5. Country-specific matrices and European average. a Final infection attack rate as a function of the basic reproduction number R_0 in the different countries (blue dots) by adopting country-specific matrices and by assuming the same probability of transmission q in all countries – specifically, the value resulting in $R_0=1.4$ by adopting the average European matrix (green dot). The attack rate corresponding to the average European matrix is computed by assuming the average European age structure in the model. Red line represents the attack rate of the homogeneous mixing SIR model for values of R_0 in the range of variability of the basic reproduction number of country-specific matrices. Grey line represents the best fit of the linear model to data points related to the use of country-specific matrices. **b** Percentage variation of infection attack rate for increasing values of R_0 of models based on country-specific matrices with respect to models based on the average European matrix (with country-specific age structure). **c** As **b** but for the variation of the peak day. **d–g** Daily prevalence over time of models with $R_0=1.4$ based on either the country-specific matrix (solid lines) or the average European matrix (dashed lines, with country-specific age structure) in Germany, Italy, France and Slovakia respectively. In this figure we assume the generation time to be 3.1 days.
doi:10.1371/journal.pcbi.1002673.g005

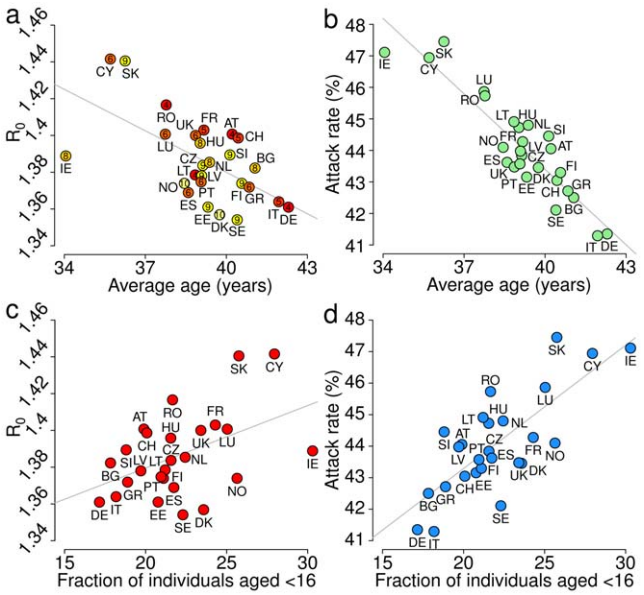

Figure 6. Socio-demography and disease epidemiology. a Basic reproduction number R_0 as a function of the average age of the population in the different countries. Numbers inside the circles represent the duration (in years) of the primary school cycle; colors from red to yellow are proportional to those numbers. **b** Final attack rate as a function of the average age of the population in the different countries. **c** Basic reproduction number as a function of the fraction of individuals younger than 16 years of age in the different countries. **d** Final attack rate as a function of the fraction of individuals younger than 16 years of age in the different countries.
doi:10.1371/journal.pcbi.1002673.g006

between countries is the average age of the population (correlation -0.91, p-value < 0.001; see Figure 6b). Finally, less strong correlations between socio-demographic features of the populations and epidemiologically relevant quantities are shown in Text S1. We also found a relationship between the proportion of individuals less than 16 years old in the population and the basic reproduction number (correlation 0.46, p-value $= 0.016$; Figure 6c), in line with a recent study on the 2009 H1N1 influenza pandemic [55]; the correlation of this fraction of the population with the attack rate is even stronger (correlation 0.76, p-value < 0.00001; Figure 6d). These results highlight that the epidemic spread is affected by the presence of younger individuals in the population, who are generally exposed to a larger force of infection than the elderly.

Conclusion

In this work we propose a method, based on the analysis of the contact network in a highly detailed virtual society, and compute the related matrices of adequate contacts for 26 European countries.

Our analysis highlights well defined correlations between epidemiological parameters and socio-demographic features of the populations. Specifically, we found that the basic reproduction number is well explained by a linear model having average age of the population and duration of primary school cycle as independent variables, whose values are easily derivable from routinely collected social and demographic data. In addition, the average age appears as the main determinant in explaining differences in final attack rates between countries. In this perspective, the use of synthetic contact matrices helps in improving the accuracy of mathematical models predictions, which are increasingly used for supporting public health decisions.

It is worth remarking that the presented approach is based on routinely collected data, and it can be easily extended to every country for which socio-demographic data are available. Notably, by providing information by one-year age brackets, our contact matrices are particularly suitable when dealing with childhood diseases which require detailed information on contact patterns in the youngest age classes. Finally, our method may be used also retrospectively, in order to reconstruct contact patterns in the past by using data from previous census rounds; this would be useful to review classic results based on indirect estimates of contacts, such as WAIFW matrices [47].

Supporting Information

Table S1 Total matrices of adequate contacts. Frequencies of total contacts by age for 26 European countries. Please note that disease transmission models usually require the average frequencies of contacts by age; these can be obtained by dividing the symmetric matrices given in this table by the age structure of the population.
(XLS)

Table S2 Setting-specific contact matrices. Matrices of contacts in households, schools, workplaces and in the general community for 26 European countries. Please note that disease transmission models usually require the average frequencies of contacts by age; these can be obtained by dividing the symmetric matrices given in this table by the age structure of the population.
(XLS)

Text S1 Supporting text. Supporting text containing methodological details and additional results.
(PDF)

Acknowledgments

The authors would like to thank three anonymous reviewers for their helpful comments.

Author Contributions

Conceived and designed the experiments: MA PM AV SM. Performed the experiments: LF MA. Analyzed the data: LF MA PM AV SM. Wrote the paper: LF MA PM AV SM. Developed the model: LF MA SM.

References

1. Rohani P, Zhong X, King AA (2010) Contact Network Structure Explains the Changing Epidemiology of Pertussis. Science 330: 982–985.
2. Kretzschmar M, Teunis PFM, Pebody RG (2010) Incidence and Reproduction Numbers of Pertussis: Estimates from Serological and Social Contact Data in Five European Countries. PLoS Med 7: e1000291.
3. Merler S, Ajelli M (2010) The role of population heterogeneity and human mobility in the spread of pandemic influenza. Proc R Soc B 277: 557–565.
4. Nishiura H, Chowell G, Safan M, Castillo-Chavez C (2010) Pros and cons of estimating the reproduction number from early epidemic growth rate of influenza A (H1N1) 2009. Theor Biol Med Mod 7: 1.
5. Merler S, Ajelli M, Pugliese A, Ferguson NM (2011) Determinants of the spatiotemporal dynamics of the 2009 H1N1 pandemic in Europe: Implications for real-time modelling. PLoS Comput Biol 7: e1002205.
6. Edmunds WJ, O'Callaghan CJ, Nokes DJ (1997) Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections. Proc R Soc B 264: 949–957.

7. Wallinga J, Teunis P, Kretzschmar M (2006) Using Data on Social Contacts to Estimate Agespecific Transmission Parameters for Respiratory-spread Infectious Agents. Am J Epidemiol 164: 936–944.
8. Beutels P, Shkedy Z, Aerts M, Van Damme P (2006) Social mixing patterns for transmission models of close contact infections: exploring self-evaluation and diary-based data collection through aWebbased interface. Epidemiol Infect 134: 1158–1166.
9. Mossong J, Hens N, Jit M, Beutels P, Auranen K, et al. (2008) Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 5: e74.
10. Zagheni E, Billari FC, Manfredi P, Melegaro A, Mossong J, et al. (2008) Using Time-Use Data to Parameterize Models for the Spread of Close-Contact Infectious Diseases. Am J Epidemiol 168: 1082–1090.
11. Iozzi F, Trusiano F, Chinazzi M, Billari FC, Zagheni E, et al. (2010) Little-Italy: An Agent-Based Approach to the Estimation of Contact Patterns- Fitting Predicted Matrices to Serological Data. PLoS Comput Biol 6: e1001021.
12. Horby P, Thai PQ, Hens N, Yen NTT, Thoang DD, et al. (2011) Social contact patterns in Vietnam and implications for the control of infectious diseases. PLoS ONE 6: e16965.
13. Hens N, Ayele GM, Goeyvaerts N, Aerts M, Mossong J, et al. (2009) Estimating the impact of school closure on social mixing behaviour and the transmission of close contact infections in eight European countries. BMC Infect Dis 9: 187.
14. Eames KTD, Tilston NL, Brooks-Pollock E, Edmunds WJ (2012) Measured Dynamic Social Contact Patterns Explain the Spread of H1N1v Influenza. PLoS Comput Biol 8: e1002425.
15. Eubank S, Guclu H, Kumar VSA, Marathe MV, Srinivasan A, et al. (2004) Modelling disease outbreaks in realistic urban social networks. Nature 429: 180–184.
16. Ferguson NM, Cummings DAT, Cauchemez S, Fraser C, Riley S, et al. (2005) Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 437: 209–214.
17. Germann TC, Kadau K, Longini IM, Macken CA (2006) Mitigation strategies for pandemic influenza in the United States. Proc Natl Acad Sci U S A 103: 5935–5940.
18. Colizza V, Barrat A, Barthélemy M, Vespignani A (2007) Predictability and epidemic pathways in global outbreaks of infectious diseases: the SARS case study. BMC Med 5: 34.
19. Ciofi Degli Atti ML, Merler S, Rizzo C, Ajelli M, Massari M, et al. (2008) Mitigation Measures for Pandemic Influenza in Italy: An Individual Based Model Considering Different Scenarios. PLoS ONE 3: e1790.
20. Pitzer VE, Viboud C, Simonsen L, Steiner C, Panozzo CA, et al. (2009) Demographic Variability, Vaccination, and the Spatiotemporal Dynamics of Rotavirus Epidemics. Science 325: 290–294.
21. Balcan B, Hao H, Gonçalves B, Bajardi P, Poletto C, et al. (2009) Seasonal transmission potential and activity peaks of the new influenza A(H1N1): a Monte Carlo likelihood analysis based on human mobility. BMC Med 7: 45.
22. Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco JJ, et al. (2009) Multiscale mobility networks and the spatial spreading of infectious diseases. Proc Natl Acad Sci U S A 106: 21484–21489.
23. Ajelli M, Merler S (2009) An individual-based model of hepatitis A transmission. J Theor Biol 259: 478–488.
24. Baguelin M, Hoek AJV, Jit M, Flasche S, White PJ, et al. (2010) Vaccination against pandemic influenza A/H1N1v in England: A real-time economic evaluation. Vaccine 28: 2370–2384.
25. Chao D, Halloran M, Obenchain V, Longini I (2010) FluTE, a publicly available stochastic influenza epidemic simulation model. PLoS Comput Biol 6: e1000656.
26. Ajelli M, Gonçalves B, Balcan D, Colizza V, Hu H, et al. (2010) Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. BMC Infect Dis 10: 190.
27. Ajelli M, Merler S, Pugliese A, Rizzo C (2011) Model predictions and evaluation of possible control strategies for the 2009 A/H1N1v influenza pandemic in Italy. Epidemiol Infect 139: 68–79.
28. Cauchemez S, Bhattarai A, Marchbanks TL, Fagan RP, Ostroff S, et al. (2011) Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. Proc Natl Acad Sci U S A 108: 2825–2830.
29. Silhol R, Boëlle PY (2011) Modelling the Effects of Population Structure on Childhood Disease: The Case of Varicella. PLoS Comput Biol 7: e1002105.
30. Van den Broeck W, Gioannini C, Gonçalves B, Quaggiotto M, Colizza V, et al. (2011) The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. BMC Infect Dis 11: 37.
31. Guzzetta G, Ajelli M, Yang Z, Merler S, Furlanello C, et al. (2011) Modeling socio-demography to capture tuberculosis transmission dynamics in a low burden setting. J Theor Biol 289: 197–205.
32. Longini IM, Halloran ME, Nizam A, Yang Y (2004) Containing Pandemic Influenza with Antiviral Agents. Am J Epidemiol 159: 623–633.
33. Meyers LA, Pourbohloul B, Newman MEJ, Skowronski DM, Brunham RC (2005) Network theory and SARS: predicting outbreak diversity. J Theor Biol 232: 71–81.
34. Meyers LA, Newman MEJ, Pourbohloul B (2006) Predicting epidemics on directed contact networks. J Theor Biol 240: 400–418.
35. Rvachev L, Longini I (1985) A mathematical model for the global spread of influenza. Math Biosci 75: 3–22.
36. Epstein J, Goedecke D, Yu F, Morris R, Wagener D, et al. (2007) Controlling pandemic flu: The value of international air travel restrictions. PLoS ONE 2: e401.
37. Colizza V, Barrat A, Barthélemy M, Valleron AJ, Vespignani A (2007) Modeling the Worldwide spread of pandemic influenza: baseline case and containment interventions. PLoS Med 4: e13.
38. Balcan D, Gonçalves B, Hu H, Ramasco J, Colizza V, et al. (2010) Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. J Comput Sci 1: 132–145.
39. Kenah E, Chao D, Matrajt L, Halloran M, Longini I (2011) The global transmission and control of influenza. PLoS ONE 6: e19515.
40. Statistical Office of the European Commission (Eurostat) (2011) Database by themes. Available: http://epp.eurostat.ec.europa.eu.
41. Hethcote H (2000) The Mathematics of Infectious Diseases. SIAM Review 42: 599–653.
42. Cauchemez S, Carrat F, Viboud C, Valleron AJ, Boëlle PY (2004) A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. Stat Med 23: 3469–3487.
43. Melegaro A, Jit M, Gay M, Zagheni E, Edmunds WJ (2011) What types of contacts are important for the spread of infections? Using contact survey data to explore European mixing patterns. Epidemics 3: 143–151.
44. Ferguson NM, Cummings DAT, Fraser C, Cajka JC, Cooley PC, et al. (2006) Strategies for mitigating an influenza pandemic. Nature 442: 448–452.
45. Cauchemez S, Valleron AJ, Boëlle PY, Flahault A, Ferguson NM (2008) Estimating the impact of school closure on influenza transmission from Sentinel data. Nature 452: 750–754.
46. Kwok KO, Leung GM, Riley S (2011) Modelling the Proportion of Influenza Infections within Households during Pandemic and Non-Pandemic Years. PLoS ONE 6: e22089.
47. Anderson RM, May RM (1991) Infectious diseases of humans: dynamics and control. Wiley.
48. Diekmann O, Heesterbeek JAP, Metz JAJ (1990) On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. J Math Biol 28: 365–382.
49. Hardelid P, Andrews NJ, Hoschler K, Stanford E, Baguelin M, et al. (2010) Assessment of baseline age-specific antibody prevalence and incidence of infection to novel influenza A/H1N1 2009. Health Technol Assess 14: 115–192.
50. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, et al. (2009) Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings. Science 324: 1557–1561.
51. Cauchemez S, Donnelly CA, Reed C, Ghani AC, Fraser C, et al. (2009) Household Transmission of 2009 Pandemic Influenza A (H1N1) Virus in the United States. N Engl J Med 361: 2619–2627.
52. Statistical Office of the European Commission (Eurostat) (2011) Average exit age from the labour force by gender. Available: http://epp.eurostat.ec.europa.eu, table code tsiem030.
53. Critchlow DE (1985) Metric Methods for Analyzing Partially Ranked Data. Lecture Notes in Statistics. Springer.
54. Kohler HP, Billari FC, Ortega JA (2002) The Emergence of Lowest-Low Fertility in Europe During the 1990s. Popul Dev Rev 28: 641–680.
55. Opatowski L, Fraser C, Griffin J, de Silva E, Van Kerkhove M, et al. (2011) Transmission Characteristics of the 2009 H1N1 Influenza Pandemic: Comparison of 8 Southern Hemisphere Countries. PLoS Pathog 7: e1002225.