

# The Modeling of Global Epidemics: Stochastic Dynamics and Predictability

V. Colizza<sup>a,\*</sup>, A. Barrat<sup>b</sup>, M. Barthélemy<sup>a,+</sup>, A. Vespignani<sup>a</sup>

<sup>a</sup>*School of Informatics and Center for Biocomplexity, Indiana University, Bloomington 47406, IN, USA*

<sup>b</sup>*Unité Mixte de Recherche du CNRS 8627, LPT, Bâtiment 210, Université de Paris-Sud, 91405 ORSAY Cedex, France*

Received: 11 July 2005 / Accepted: 17 November 2005 / Published online: 20 June 2006  
© Society for Mathematical Biology 2006

**Abstract** The global spread of emergent diseases is inevitably entangled with the structure of the population flows among different geographical regions. The airline transportation network in particular shrinks the geographical space by reducing travel time between the world's most populated areas and defines the main channels along which emergent diseases will spread. In this paper, we investigate the role of the large-scale properties of the airline transportation network in determining the global propagation pattern of emerging diseases. We put forward a stochastic computational framework for the modeling of the global spreading of infectious diseases that takes advantage of the complete International Air Transport Association 2002 database complemented with census population data. The model is analyzed by using for the first time an information theory approach that allows the quantitative characterization of the heterogeneity level and the predictability of the spreading pattern in presence of stochastic fluctuations. In particular we are able to assess the reliability of numerical forecast with respect to the intrinsic stochastic nature of the disease transmission and travel flows. The epidemic pattern predictability is quantitatively determined and traced back to the occurrence of epidemic pathways defining a backbone of dominant connections for the disease spreading. The presented results provide a general computational framework for the analysis of containment policies and risk forecast of global epidemic outbreaks.

**Keywords** Complex networks · Epidemiology

## 1. Introduction

Real-world populations often show a complicated age/social structure as well as heterogeneous patterns in the contact networks determining the transmission

---

\*Corresponding author.

E-mail address: vcolizza@indiana.edu.

<sup>+</sup>On leave from CEA-Centre d'Etudes de Bruyères-Le-Chatel, France.

dynamics (Anderson and May, 1992; Hethcote and Yorke, 1984; Kretzschmar and Morris, 1996; Keeling, 1999; Pastor-Satorras and Vespignani, 2001; Lloyd and May, 2001). All these factors have led to sophisticated modeling approaches including social structures, heterogeneous connectivity patterns, meta-population grouping, stochasticity and, more recently, to numerical approaches relying on agent based modeling that recreate entire populations and their dynamics at the scale of the single individual (Chowell et al., 2003; Eubank et al., 2004). In many cases, however, system complexity is not the same as the mere addition of complicated elements accounted for in sophisticated epidemic modeling (Ferguson et al., 2003). Complex features often imply a virtual infinite heterogeneity of the system and large-scale statistical fluctuations (Albert and Barabási, 2000; Amaral et al., 2000; Dorogovtsev and Mendes, 2003; Pastor-Satorras and Vespignani, 2003), which generally call for new theoretical frameworks and models (Pastor-Satorras and Vespignani, 2001; Lloyd and May, 2001; Meyers et al., 2005).

These considerations are particularly relevant in the study of the geographical spread of epidemics where the various long-range connections typical of modern transportation networks naturally give rise to a very complicated evolution of epidemics characterized by heterogeneous and seemingly erratic outbreaks (Institute of Medicine, 1992; Cohen, 2000; Cliff and Haggett, 2004) as recently documented for the SARS outbreak.<sup>1</sup> In this context, air-transportation represents a major channel of epidemic propagation, as put forward in the seminal work on global epidemic diffusion by Rvachev and Longini (1985) capitalizing on previous studies of the Russian airline network (Baroyan et al., 1969). Similar modeling approaches, even if limited by a very partial knowledge of the world-wide transportation network, have been used to reproduce specific outbreaks such as pandemic influenza (Longini, 1988; Grais et al., 2003, 2004), HIV (Flahault and Valleron, 1991), and very recently SARS (Hufnagel et al., 2004), with results in reasonable agreement with experimental data. The availability of the complete world-wide airport network dataset (WAN) and the recent extensive studies of its topology (Barrat et al., 2004; Guimerà et al., 2005) are finally allowing a full-scale computational study of global epidemics. In particular we are now in the position to rationalize the role of the large-scale properties of the airline transportation network in the heterogeneity and predictability of the epidemic pattern.

In the following we will consider a stochastic modeling and a computational framework for the study of global epidemics. A first input in the computational approach is the complete International Air Transport Association (IATA)<sup>2</sup> database, allowing an unprecedented level of detail in the study of the evolution of the spatiotemporal pattern of the epidemics. A second important input is given by the populations obtained from various sources and census of all cities associated to airports. The main result of the network analysis is that the WAN is highly heterogeneous both in the connectivity pattern and the traffic capacities. In particular the presence of broad statistical distributions and non-linear associations among the various quantities, contrary to linear relations used so far, might have

---

<sup>1</sup><http://www.who.int/csr/sars/en>.

<sup>2</sup><http://www.iata.org>.

a major impact in the ensuing disease spreading pattern. While previous studies have in general focused on the simulation and reproduction of case studies of real epidemics, the large-scale modeling allowed by the IATA database enables us to address more general theoretical issues such as (i) the spatiotemporal statistical properties of the epidemic pattern, (ii) their relation with the complex features of the underlying transportation network and (iii) the predictability of the epidemic evolution with respect to the inherent stochastic dynamics of the disease transmission.

The paper is organized as follows. In Section 2, we introduce the general modeling framework. We define the basic compartmental assumption used to describe the populations and the stochastic Langevin formulation of the epidemic dynamics within each city. We also define the stochastic transport operator defining the traveling of individuals among cities. In Section 3 we derive the set of formal equations describing the evolution of the global epidemic outbreaks by coupling the infection dynamics within cities with the action of the transport operator. The various sources of stochasticity, the general integration procedure and the input provided by the IATA network and population data are explained in detail, yielding the computational framework used to obtain the epidemic outbreak forecast. In Section 4 we implement the developed computational approach to the specific case of a susceptible-infected-removed (SIR) infection dynamics. We show how it is possible to obtain a detailed study of the quantities customarily tracked in epidemiological studies and provide a global analysis of the spatiotemporal evolution of the epidemic pattern. The rather heterogeneous spatiotemporal pattern emerging during the epidemic evolution might find its origin in one or more features of the underlying network as well as in the inherent stochastic dynamics of the disease transmission. In order to discriminate the various effects we introduce in Section 5 an entropic measure of the spatial disorder of the epidemic prevalence. In particular the key role of the transportation network is evidenced by comparing the results obtained in the WAN with those generated in network models representing different null hypotheses. The network structure appears to have a major effect in driving the epidemic evolution, therefore allowing the possibility of predicting an overall epidemic pattern on top of the stochastic fluctuations. Motivated by these findings, in Section 6 we investigate the predictability of the epidemic spread through the analysis of the statistical overlap of the epidemic pattern generated in different stochastic realizations of the spreading process. This overlap is quantitatively defined by using a statistical similarity analysis of the global distribution of infected individuals. Also in this case the comparison of different null hypothesis for the network topology and the real-world IATA network allows the determination of the main structural features affecting the predictability of the epidemic scenario. Finally, we conclude in Section 7 by providing an outlook on further open questions and the possible implementation of the presented framework for epidemic forecast and the assessment of containment policies.

## 2. Methods

In the following, we use a stochastic modeling of the epidemics based on the standard compartmentalization of the population into a few classes of individuals such

as susceptible (S), infected (I), latent (L), permanently recovered (R), etc. In each city  $j$  the population is  $N_j$  and  $X_j^{[m]}(t)$  is the number of individuals in the class  $[m]$  at time  $t$ . By definition it follows that  $N_j = \sum_m X_j^{[m]}(t)$ . The dynamics of individuals in each class is determined by two elements. Individuals may move from city  $j$  to city  $\ell$  by means of the airline transportation network and can change compartment because of the infection dynamics, similarly to the models in [Rvachev and Longini \(1985\)](#), [Flahault and Valleron \(1991\)](#), and [Grais et al. \(2003\)](#) and to the stochastic generalization of [Hufnagel et al. \(2004\)](#). More generally, this model can thus be used for the description of the effect of human flows between different locations on the spread of an epidemics.

2.1. Transport operator

The dynamics of individuals due to (air) travel between cities is described by the transport operator  $\Omega_j(\{X^{[m]}\})$  representing the net balance of individuals in a given class  $X^{[m]}$  that entered and left each city  $j$ . In each city  $j$  the transport operator reads as

$$\Omega_j(\{X^{[m]}\}) = \sum_{\ell} \left( \xi_{\ell j}(X_{\ell}^{[m]}) - \xi_{j\ell}(X_j^{[m]}) \right), \tag{1}$$

where  $\xi_{j\ell}(X_j^{[m]})$  are stochastic variables representing the numbers of individuals in the class  $X^{[m]}$  traveling on each connection  $j \rightarrow \ell$  at time  $t$ . In the case of air travel, this operator is a function of the passenger traffic flows and the city populations  $N_j$ , and might also include transit passengers on connecting flights. If we neglect multiple legs travels and if we assume the population homogeneity inside the cities, the probability that any individual travels from city  $j$  to city  $\ell$  during a time interval  $\Delta t$  is given by  $p_{j\ell} = \frac{w_{j\ell}}{N_j} \Delta t$ , where  $w_{j\ell}$  is the airplane passenger traffic per unit time. The stochastic variables  $\xi_{j\ell}(X_j^{[m]})$  will therefore follow the multinomial distribution

$$P(\{\xi_{j\ell}\}) = \frac{X_j^{[m]}!}{(X_j^{[m]} - \sum_{\ell} \xi_{j\ell})! \prod_{\ell} \xi_{j\ell}!} \left( \prod_{\ell} p_{j\ell}^{\xi_{j\ell}} \right) \left( 1 - \sum_{\ell} p_{j\ell} \right)^{(X_j^{[m]} - \sum_{\ell} \xi_{j\ell})}, \tag{2}$$

where  $(X_j^{[m]} - \sum_{\ell} \xi_{j\ell})$  identifies the number of non-traveling individuals of class  $[m]$  staying in city  $j$ . The mean and variance of the stochastic variables are  $\langle \xi_{j\ell}(X_j^{[m]}) \rangle = p_{j\ell} X_j^{[m]}$  and  $\text{Var}(\xi_{j\ell}(X_j^{[m]})) = p_{j\ell}(1 - p_{j\ell}) X_j^{[m]}$ , respectively. These expressions allow to recover the average transport operator

$$\langle \Omega_j(\{X^{[m]}\}) \rangle = \sum_{\ell} \left( p_{\ell j} X_{\ell}^{[m]} - p_{j\ell} X_j^{[m]} \right), \tag{3}$$

used in previous approaches ([Rvachev and Longini, 1985](#)). If the corresponding traveling fluxes are large, the fluctuations around the expected value are small and

one can indeed use this average expression. When dealing with a small number of individuals in a given class  $X^{[m]}$ , such as the number of infectious at the early stage of an epidemic, expression (3) is not valid and it is necessary to use the stochastic operator defined by Eqs. (1) and (2).

In many cases since the traffic flows are expressed as the number of available seats on a given connection, we have to consider that the transport operator is in general affected by fluctuations due to an occupancy rate of the airplanes not equal to 1. This introduces a further source of noise since we have to consider that on each connection  $(j, \ell)$  the flux of passengers at each time  $t$  is given by a stochastic variable

$$\tilde{w}_{j\ell} = w_{j\ell}[\alpha + \eta(1 - \alpha)], \tag{4}$$

where  $\alpha = 0.7$  corresponds to the average occupancy rate of 70% provided by official statistics and  $\eta$  is a random number drawn uniformly in the interval  $[-1, 1]$  at each time step.

A final note concerns the inclusion of traffic due to transit passengers in the transport operator. It is possible to obtain a full expression of the transport operator that considers up to two legs travels. This is obtained by including data on traffic passengers in the main airports. This result is a further technical complication whose detailed calculation is reported in the Appendix. It is worth remarking that the results obtained by including transit passenger flows are not qualitatively dissimilar to those presented here.

### 2.2. The epidemic Langevin equation

The dynamics of the individuals  $X^{[m]}$  between the different compartments depends on the specific disease considered. In general, the transition from a compartment to the other is specified by a specific rate that depends on the disease etiology, such as the infection transmission rate or the recovery or cure rate. In compartmental models there are two possible elementary processes ruling the disease dynamics. The first class of process refers to the spontaneous transition of one individual from one compartment  $[m]$  to another compartment  $[h]$

$$X^{[m]} \rightarrow X^{[m]} - 1, \tag{5}$$

$$X^{[h]} \rightarrow X^{[h]} + 1. \tag{6}$$

Processes of this kind are the spontaneous recovery of infected individuals ( $I \rightarrow R$ ) or the passage from a latent condition to an infectious one ( $L \rightarrow I$ ) after the incubation period. In this case the variation in the number of individuals  $X^{[m]}$  is simply given by  $\sum_h v_h^m a_h X^{[h]}$ , where  $a_h$  is the rate of transition from the class  $[h]$  and  $v_h^m \in \{-1, 0, 1\}$  is the change in the number of  $X^{[m]}$  due to the spontaneous process from or to the compartment  $[h]$ .

The second class of processes refers to binary interaction among individuals such as the contagion of one susceptible in interaction with an infectious ( $S + I \rightarrow 2I$ ). In the homogeneous assumption, the rate of variation of individuals  $X^{[m]}$  is given

by  $\sum_{h,g} v_{h,g}^m a_{h,g} N^{-1} X^{[h]} X^{[g]}$ , where  $a_{h,g}$  is the rate of transition rate of the process and  $v_{h,g}^m \in \{-1, 0, 1\}$  the change in the number of  $X^{[m]}$  due to the interaction. The factor  $N^{-1}$ , where  $N$  is the number of individuals, stems from the fact that the above expression considers the homogeneous approximation in which the probability for each individual of class  $[h]$  to interact with an individual of class  $[g]$  is simply proportional to the density  $X^{[g]}/N$  of such individuals (note that it is however possible to consider other cases (Anderson and May, 1992)). If we neglect fluctuations and use the above expression, it is possible to derive the usual deterministic reaction rate equations for the average number of individuals in the compartment  $[m]$  in each city  $j$  as

$$\partial_t \langle X_j^{[m]} \rangle = \sum_{h,g} v_{h,g}^m a_{h,g} N_j^{-1} \langle X_j^{[h]} \rangle \langle X_j^{[g]} \rangle + \sum_h v_h^m a_h \langle X_j^{[h]} \rangle. \tag{7}$$

If the total number of individuals is assumed to be constant (which is a realistic assumption if the disease spreads over a time-scale small compared to the typical time needed for a significant total population change), these equations must satisfy the conservation equation  $\sum_m \partial_t \langle X^{[m]} \rangle = 0$ . In order to go beyond the usual deterministic approximations, it is possible to work directly with the master equation for the processes described above and under the assumption of large populations, we obtain the Langevin equations in which we associate to each reaction process a noise term with amplitude proportional to the square root of the reaction term (Gardiner, 2004; Gillespie, 2000). The general Langevin equation for the evolution of  $X_j^{[m]}$  is

$$\begin{aligned} \partial_t X_j^{[m]} = & \sum_{h,g} N_j^{-1} v_{h,g}^m a_{h,g} X_j^{[h]} X_j^{[g]} + \sum_h v_h^m a_h X_j^{[h]} \\ & + \sum_{h,g} v_{h,g}^m \sqrt{a_{h,g} N_j^{-1} X_j^{[h]} X_j^{[g]}} \eta_{h,g} + \sum_h v_h^m \sqrt{a_h X_j^{[h]}} \eta_h, \end{aligned} \tag{8}$$

where  $\eta_{h,g}$  and  $\eta_h$  are statistically independent Gaussian noises. The Langevin formulation takes into account fluctuations and is widely used in chemical and reaction diffusion processes (Gillespie, 2000). It is worth remarking that the Langevin formulation is an approximation to the general master equations approach whose complete solution is computationally very expensive in the case of a large number of compartments. In the next section, we will see that the global epidemic spreading model which we are going to consider takes into account a total number of compartments equal to the number of urban areas (3100) times the number of states needed to describe the infection dynamics. At the cost of introducing an approximated treatment in the case of areas with a small population, the Langevin equation thus represents a viable approach to the evolution of the system.

### 3. Computational framework for the modeling of global epidemics

In this section we report the general computational approach that we use to model and simulate the spread of global epidemics. In particular we discuss the numerical

integration scheme of the epidemic Langevin equations in each city  $j$  in presence of the coupling induced by the transport operator. In addition, we review the properties of the worldwide airport network dataset that is used as the basic input in defining the structure and parameters of the computational approach.

### 3.1. Stochastic formulation of the global epidemic model

The dynamical evolution of the epidemic outbreak is specified by the changes in the population of the various classes in each city  $j$ . The epidemic Langevin equations are coupled among them by the transport operator that describes movements of individuals from one city to another and can be numerically solved (Hufnagel et al., 2004; Gardiner, 2004; Gillespie, 2000; Marro and Dickman, 1998) by considering the small time steps  $\Delta t$  discretization

$$\begin{aligned}
 X_j^{[m]}(t + \Delta t) - X_j^{[m]}(t) = & \sum_{h,g} N_j^{-1} v_{h,g}^m a_{h,g} X_j^{[h]}(t) X_j^{[g]}(t) \Delta t \\
 & + \sum_h v_h^m a_h X_j^{[h]}(t) \Delta t + \sum_{h,g} v_{h,g}^m \sqrt{a_{h,g} N_j^{-1} X_j^{[h]} X_j^{[g]}} \Delta t \eta_{h,g} \\
 & + \sum_h v_h^m \sqrt{a_h X_j^{[h]}} \Delta t \eta_h + \Omega_j(\{X^{[m]}\}), \tag{9}
 \end{aligned}$$

where  $\eta_{h,g}$  and  $\eta_h$  are now statistically independent Gaussian random variables with zero mean and unit variance. The stochastic transport operator  $\Omega$  is determined by the expressions (1) and (2) calculated with the time interval  $\Delta t$ . This simple discretization (9) leads to a well-known technical problem, in particular at the beginning of the spreading when some of the  $X_j^{[m]}$  are small so that a negative value of the noise term could produce an unphysical negative value of infected individuals. Various possibilities exist to avoid this problem and a first naive approach would consist in setting  $X_j^{[m]}$  to exactly zero whenever the numerical integration yields  $X_j^{[m]}(t + dt) \leq 0$ . This however corresponds to an asymmetric truncation of the noise and might introduce uncontrolled biases. An interesting alternative has been put forward by Dickman (1994) and consists in separating each variable  $X_j^{[m]}$  into its integer part  $[X_j^{[m]}]$  and its non-integer remaining  $\tilde{X}_j^{[m]}$  and to write the corresponding time evolution equations for these variables. This method used with a sufficiently small time-step for the numerical integration ensures that  $X_j^{[m]}$  always remain positive. Note also that this treatment is fully consistent with the stochastic procedure used for the transport term, in which the fields  $X_j^{[m]}$  are directly treated as integers.

### 3.2. The world-wide air transportation network and population data

The basic ingredient of the present computational approach is the dataset providing travel flows among urban areas. The flows  $w_{j\ell}$  indeed determine the structure of the dynamical equations by defining the couplings among the various cities and

determining the large-scale spread of the epidemics. In the following we use the IATA database containing the world list of airport pairs connected by direct flights and the number of available seats on any given connection for the year 2002. More precisely, each direct flight connection between airports  $j$  and  $\ell$  has a weight  $w_{j\ell}$  which is the number of available seats. The resulting world-wide air-transportation network (WAN) is therefore a weighted graph comprising  $V = 3880$  vertices denoting airports and  $E = 18810$  weighted edges accounting for the presence of a direct flight connection. The average degree (number of connections of each airport) of the network is  $\langle k \rangle = 2E/V = 9.70$ , while the maximal degree is 318. This dataset has been complemented by the population  $N_j$  of the large metropolitan area served by the airports as obtained by publicly available census databases on the WWW (such as the United Nations census<sup>3</sup> or the US census.<sup>4</sup>) The obtained network displays high levels of heterogeneity both in the connectivity pattern and in the traffic capacities, where the traffic  $T_j$  of a node  $j$  is defined as the sum of the weights of the links starting from  $j$

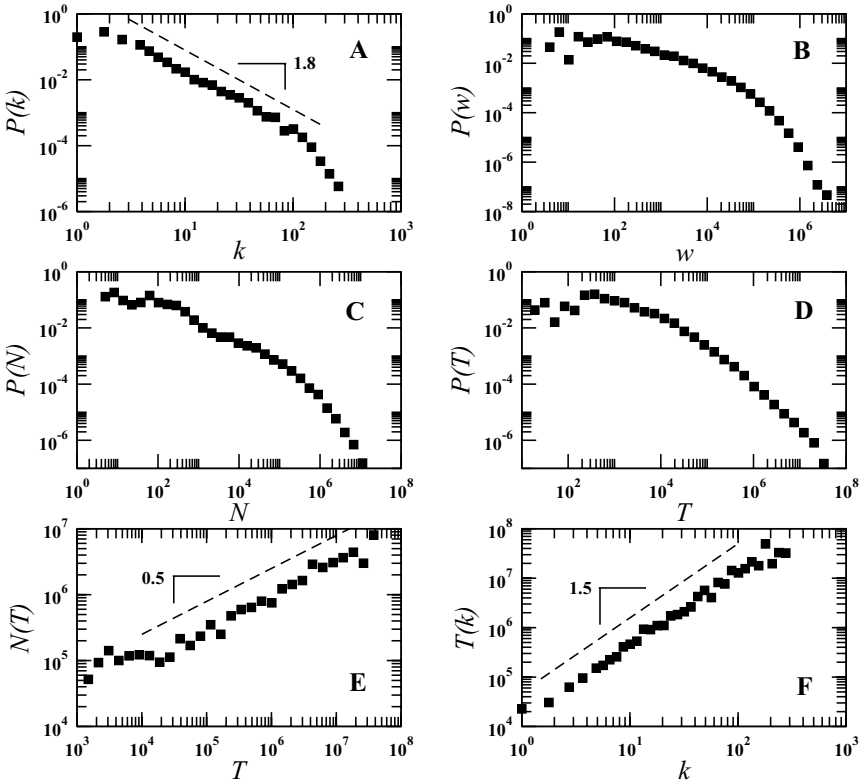
$$T_j = \sum_{\ell \in \mathcal{V}(j)} w_{j\ell} \quad (10)$$

where  $\mathcal{V}(j)$  denotes the set of neighbors of node  $j$ . In Fig. 1 we show the behavior of the basic quantities characterizing the world-wide network. The degree probability distribution  $P(k) = n(k)/V$ , where  $n(k)$  is the number of airports with  $k$  connections (i.e. of degree  $k$ ), is broadly distributed with a degree range covering almost two decades, and can be approximated (Guimerà et al., 2005) by a power-law with exponent close to 2. Moreover, weights and traffic display a strong heterogeneity revealed by very broad distributions  $P(w)$  and  $P(T)$  spanning more than 5 orders of magnitude. It is clear that such large heterogeneities will have a strong impact on any dynamical process—in particular spread of epidemics—taking place on the considered network. Finally, to each airport corresponds a city whose population  $N$  is heavy-tailed distributed in agreement with the general result of Zipf's (1949) law for city size. Strikingly, these quantities appear to have non-linear relations among them. The traffic  $T$  handled by each airport varies with the corresponding number of connections  $k$  as a power law  $T \sim k^\beta$  with exponent  $\beta \simeq 1.5$  (Barrat et al., 2004). The city population and the traffic handled by the corresponding airport obey the non-linear relation  $N \sim T^\alpha$  with  $\alpha \simeq 0.5$  in contrast with the linear behavior assumed in previous studies (Hufnagel et al., 2004). Numerical simulations will consider the 3100 airports with highest traffic  $T$ , which are complemented by population data of the corresponding urban area. This fraction corresponds to 80% of the total number of airports and carries more than 99% of the total traffic. It is worth noticing that in such a heterogeneous frame, it is very important to take into account as much as possible the fluctuations and large-scale variability of the world-wide airport network. Considering small subsets of the network leads to serious approximations on the structure of the network.<sup>5</sup>

<sup>3</sup><http://unstats.un.org/>.

<sup>4</sup><http://www.census.gov>.





**Fig. 1** Properties of the world-wide airport network. Statistical fluctuations are observed over a broad range of length scales. (A) The degree distribution  $P(k)$  follows a power-law behavior on almost two decades with exponent  $1.8 \pm 0.2$ . (B) The distribution of the weights (fluxes) is skewed and heavy-tailed. (C) The distribution of populations is power-law distributed, following Zipf's law (Zipf, 1949). (D) The traffic  $T$  displays large fluctuations with a distribution  $P(T)$  spanning more than 6 orders of magnitude. (E) The city population varies with the traffic of the corresponding airport as  $N \sim T^\alpha$  with  $\alpha \simeq 0.5$ , in contrast with the linear behavior postulated in previous works (Hufnagel et al., 2004). (F) The traffic  $T$  varies with connectivity as a power law with exponent  $1.5 \pm 0.1$ .

**4. Spatiotemporal pattern of global epidemic outbreaks**

Global epidemic forecast would be extremely relevant in the case of the emergence of a new pandemic influenza. Indeed, pandemic influenza killed millions of people world-wide throughout this century (1918, 1957 and 1968) and the occurrence of a new outbreak has to be expected. Pandemic influenza spreads rapidly and substantial transmission occurs before the onset of case-defining symptoms. In the following we adopt a minimal model for a pandemic spread with the aim

<sup>5</sup>Smaller fractions of the network used in earlier studies such as the 50 (resp. 500) largest airports correspond to only 2% (resp. 12%) of the total network in terms of number of nodes and carry 20% (resp. 80%) of the total traffic, according to the 2002 IATA database.

of understanding the effect of the underlying air-transportation network structure on the general properties of the spatiotemporal outbreak pattern. We use the very simplistic approximation of the SIR dynamics in which a fully mixed population is assumed within each city. A refined modeling would include additional aspects such as detailed information about contact networks, age structure, seasonal effects, stages of infection of the specific disease. However, our study aims at the discussion of the general statistical properties of the global spreading and the role of the transportation network heterogeneity in the disease spreading. From this point of view, while we provide the most possible details on the transportation network, we make use of the minimal SIR model in order to provide a general discussion that is not hindered by the use of very complicate disease transmission mechanisms. Specific results of the present computational approach in the case of realistic disease transmission and the comparison with real case studies will be reported elsewhere.

#### 4.1. The SIR case

In the basic standard compartmentalization of the SIR, each individual can only exist in one of the discrete states such as susceptible (S), infected (I) or permanently recovered (R). In each city  $j$  the population  $N_j$  is given by  $N_j = S_j(t) + I_j(t) + R_j(t)$ , where  $S_j(t)$ ,  $I_j(t)$  and  $R_j(t)$  represent the number of susceptible, infected and recovered individuals at time  $t$ , respectively (in the following we replace for simplicity the general notation  $X_j^{[m]}$  by  $S_j, I_j, R_j$ ). The epidemic evolution is governed by the basic dynamical evolution of the SIR model where the probability of a susceptible individual acquiring the infection from any given infected individual in the time interval  $dt$  is proportional to  $\beta dt$ . Here  $\beta$  is the transmission parameter that captures the aetiology of the infection process. At the same time, infected individuals recover with a probability  $\mu dt$ , where  $\mu^{-1}$  is the average duration of the infection. The relevant parameter describing the epidemic is the basic reproduction number  $R_0 = \beta/\mu$  (Anderson and May, 1992) given by the average number of secondary cases that each infected individual generates in a susceptible population. If  $R_0 > 1$  and the initial density of susceptibles is larger than  $R_0^{-1}$ , then an epidemic will develop in the city. Using the scheme presented in Section 3 and writing the discretized epidemic Langevin equation for each class of the SIR model we obtain

$$S_j(t + \Delta t) - S_j(t) = -\beta \frac{I_j(t)S_j(t)}{N_j} \Delta t + \sqrt{\beta \frac{I_j S_j}{N_j}} \Delta t \eta_{j,1}(t) + \Omega_j(\{S\}) \quad (11)$$

$$I_j(t + \Delta t) - I_j(t) = +\beta \frac{I_j(t)S_j(t)}{N_j} \Delta t - \mu I_j(t) \Delta t - \sqrt{\beta \frac{I_j S_j}{N_j}} \Delta t \eta_{j,1}(t) + \sqrt{\mu I_j \Delta t} \eta_{j,2}(t) + \Omega_j(\{I\}) \quad (12)$$

$$R_j(t + \Delta t) - R_j(t) = +\mu I_j(t) \Delta t - \sqrt{\mu I_j \Delta t} \eta_{j,2}(t) + \Omega_j(\{R\}), \quad (13)$$

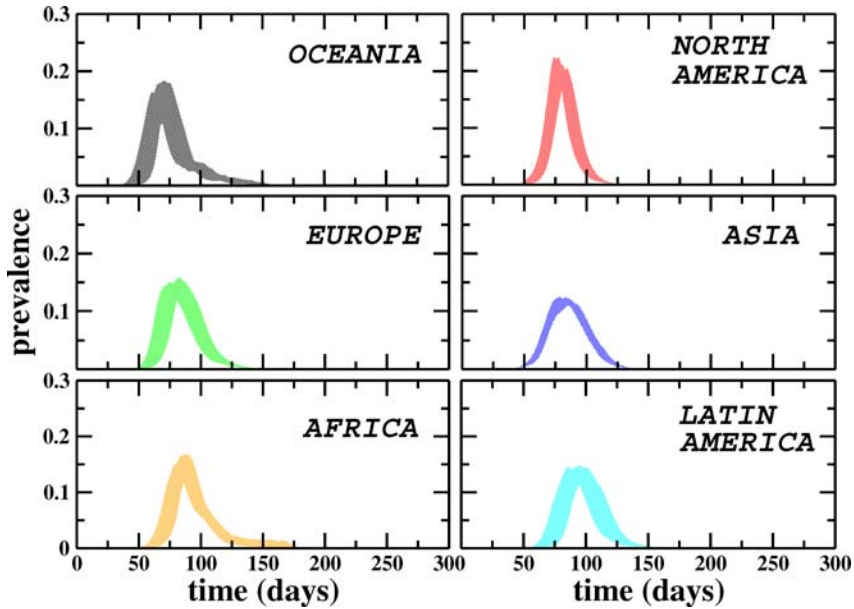
where  $\eta_{j,1}$  and  $\eta_{j,2}$  are statistically independent Gaussian random variables with zero mean and unit variance, and  $\Omega_j(\{X\})$  is the stochastic travel operator (defined in Section 2.1) depending on the traveling probabilities (obtained from the IATA dataset)  $p_{j\ell} = w_{j\ell}\Delta t/N_j$  of any individual to travel from city  $j$  to city  $\ell$ .

The model is thus a compartmental system of  $3100 \times 3$  differential equations whose integration provides the disease evolution in every urban area corresponding to an airport. The system can be solved numerically by using standard Cauchy–Euler discretization methods for small time steps  $\Delta t$ . In the limit of making the step size small, this procedure represents a good approximation to the underlying differential equations. In order to avoid un-physical negative values of infected individuals, specific numerical techniques must be used, as those illustrated in Section 3.1.

The presented computational framework may be used to obtain detailed forecast of the geographic spread of an emerging disease starting in differential initial seeds. At first instance, it is possible to monitor standard epidemiological quantities such as the level of infected individuals, the morbidity and the prevalence at different granularity levels; i.e. country, state or administrative regions. The values of the disease parameters  $\beta$  and  $\mu$  have been chosen according to (Grais et al., 2004) to ensure biologically sound values and kept constant during the evolution. This amounts to assume that no restrictions on traveling and targeted prophylaxis measure are implemented during the outbreak. In this perspective, it is worth remarking that the presented results have to be considered just in a theoretical perspective. We performed several different realizations of the noise for an epidemic outbreak starting from a given initially infected city. A sensitivity analysis of the model respect to the initial conditions has been performed by specifying different origins of the pandemic outbreak, namely Hong Kong, New York, Los Angeles, San Francisco and London. Moreover, global properties have been studied by averaging over different initially infected cities.

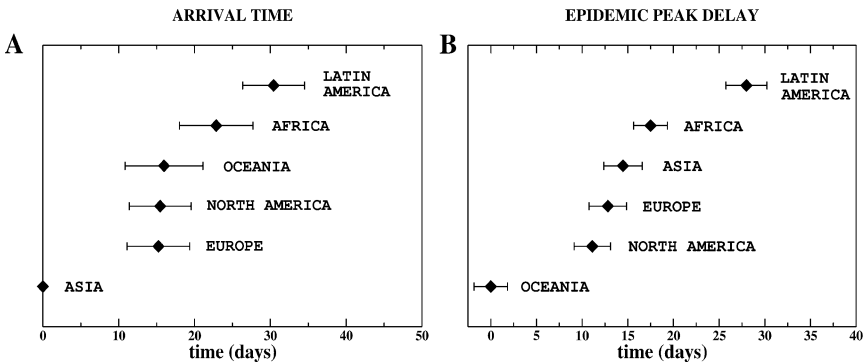
Figure 2 shows the profiles of the prevalence, defined as the fraction of infected individuals, restricted to the World continents (with America split in North America and Latin America) for an epidemics starting in Hong Kong. The maximal dispersion as obtained with different realizations of the noise is shown for each continent. Differences can be noted in the time evolution of the epidemics in different regions of the World. Figure 3 A shows the arrival time of the infection for each continent. Starting from Asia, the infection is very soon detected in Europe, North America and Oceania, which are connected to the initially infected city through direct flight connections carrying a large traffic, and much later in Africa and Latin America. The epidemic peak is first reached in Oceania (see Fig. 3B) followed by North America, Europe, Asia and Africa. The maximum of the prevalence is much more delayed in Latin America with respect to other continents since it is the last one to be infected.

The prevalence profiles shown in Fig. 2 give only a global measure of the time evolution of the epidemics in a certain region. Further insight in the geographic spread of an emerging disease can be obtained by representing the level of infected individuals at the different granularity levels through the use of maps. In Figs. 4 and 5 we present the dynamical evolution in the United States and in Europe,

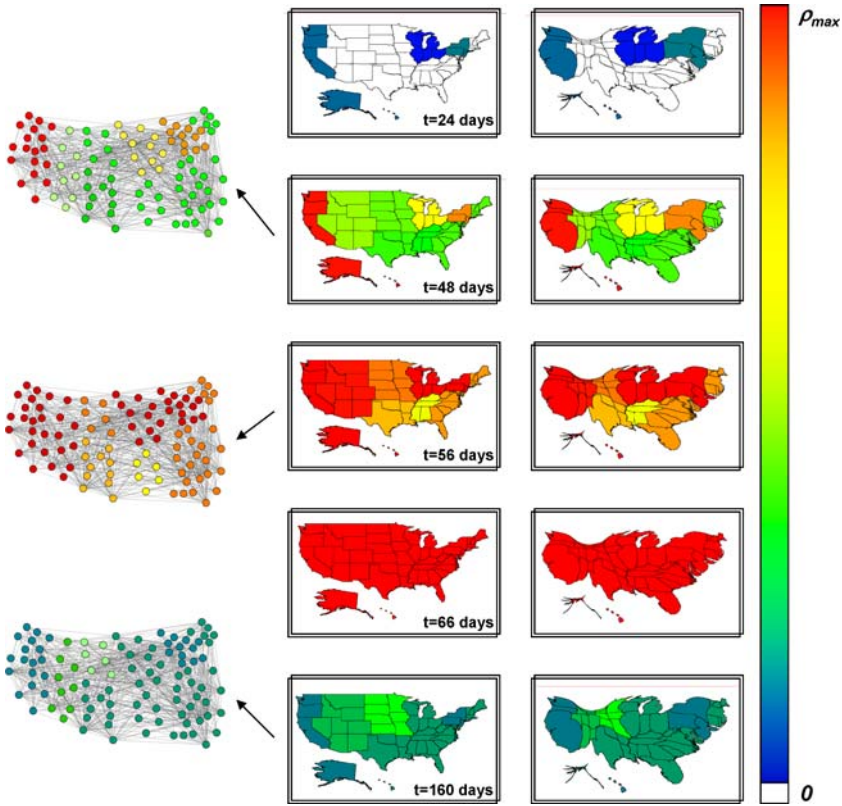


**Fig. 2** Prevalence profiles for each continent for an epidemics starting in Hong Kong. The shaded area represents the maximal dispersion of the prevalence profile obtained with different realizations of the noise.

respectively, of a pandemic starting in Hong Kong. The evolution of epidemic outbreaks is monitored by recording at each time step (1 day) the density of individuals in each class ( $S, I, R$ ) present in each city. We group the US states in the nine



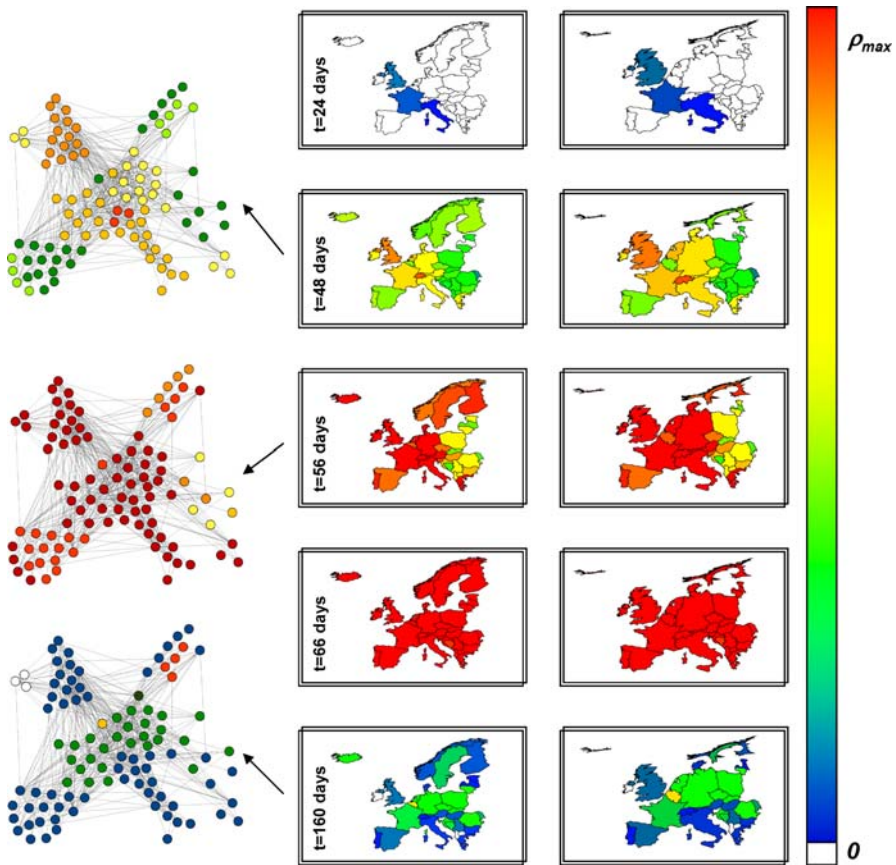
**Fig. 3** Arrival time (A) and epidemic peak delay (B) restricted to the World continents for an epidemics originated in Hong Kong. Starting from Asia, the epidemics quickly infects cities in Europe, North America and Oceania, then reaches Africa and much later Latin America. The time evolution of the epidemics is faster in Oceania where the maximum of the prevalence is first reached, followed by North America, Europe, Asia and Africa. The last continent which experiences the outbreak of the epidemics is Latin America, which is also the last one to be infected.



**Fig. 4** Geographical representation of the evolution in the US of the SIR epidemic specified in the text with Hong Kong as initial seed. States are grouped according to the nine influenza surveillance regions. The color code corresponds to the prevalence in each region, from 0 to the maximum value reached ( $\rho_{max}$ ). The first set of maps provides the original US maps, while the second shows the corresponding cartograms obtained by rescaling each region according to its population (Gastner and Newman, 2004). Three representations of the airport network restricted to the United States are also shown, corresponding to three different snapshots of the epidemic diffusion. For the sake of visualization, only the 100 airports with largest traffic in the US are shown, however the data have been obtained by using the full data set including 3100 airports. The color code is the same adopted for the maps.

influenza surveillance regions<sup>6</sup> which are identical to the nine divisions of the US census. Two different visualization strategies are used for both the US and Europe. In the first set of maps, regions are drawn with their original size and a color code gives the prevalence of the infection in each region. This representation readily shows the high heterogeneity of the pandemic evolution. While useful, such a visualization might be misleading, since the same prevalence obtained in different regions might correspond to very different values in the number of infected individuals if the two regions are very differently populated. Moreover, it is common to

<sup>6</sup><http://www.cdc.gov/flu/>.



**Fig. 5** Geographical representation of the evolution in Europe of the SIR epidemics starting in Hong Kong. The color code corresponds to the prevalence in each European country, from 0 to the maximum value reached ( $\rho_{\max}$ ). The first set of maps provides the original maps of Europe, while the second shows the corresponding cartograms obtained by rescaling each country according to its population (Gastner and Newman, 2004). Three representations of the airport network restricted to Europe are also shown, corresponding to three different snapshots of the epidemic diffusion. For the sake of visualization, only the 100 airports with largest traffic in Europe are shown, however the data have been obtained by using the full data set including 3100 airports. The color code is the same adopted for the maps.

find strong heterogeneities in the population density, and it is not easy to detect visually a large level of contamination in a small but densely populated geographical area. In order to obtain a geographical representation which is able to carry at the same time information both on the level of infection and on the number of infection cases in each region, we have constructed the corresponding cartograms of the original maps in which the size of each geographic region—influenza surveillance regions in the US and countries in Europe—is rescaled according to its population. Several methods for constructing cartograms have been developed (see Gastner and Newman, 2004, and references therein); here we adopt the diffusion-based

method (Gastner and Newman, 2004), which produces cartograms by equalizing the population density through a linear diffusion process. This method depicts simultaneously both the prevalence and the total number of cases and conveys properly the impact in terms of infection cases despite the strong heterogeneities in population density. The obtained maps strengthen the evidence of a very heterogeneous pattern of the epidemic evolution. Countries and states show different levels of prevalence following a diverse temporal pattern. The differences in the epidemic peaks that we observe at the continental granularity are present at the country and state level. While the visual inspection provides a very intuitive evidence for the complex evolution of the epidemics, we still lack a quantitative analysis of the epidemic pattern. This is particularly relevant if we aim at connecting the observed epidemic evolution with the underlying transport network structure and the stochasticity of the infection transmission.

4.2. Deterministic approximation

In order to gain some analytical understanding on the time evolution of the epidemics it is convenient to consider the deterministic versions of the stochastic equations (11), (12), (13) that read as

$$\frac{dS_j}{dt} = -\beta \frac{I_j S_j}{N_j} + \langle \Omega_j(\{S\}) \rangle \tag{14}$$

$$\frac{dI_j}{dt} = \beta \frac{I_j S_j}{N_j} - \mu I_j + \langle \Omega_j(\{I\}) \rangle \tag{15}$$

$$\frac{dR_j}{dt} = \mu I_j + \langle \Omega_j(\{R\}) \rangle. \tag{16}$$

These equations correspond to the average behavior and contain only the averaged expression of the transport operator. At the early stage of the epidemics, the number of infected individuals is relatively small in all cities and it is possible to linearize the evolution equations for the number  $I_j(t)$  of infected ( $S_j \simeq N_j$ ) as

$$\partial_t I_j = \Lambda_j I_j + \sum_{\ell} \frac{w_{\ell j}}{N_{\ell}} I_{\ell} \tag{17}$$

where  $\Lambda_j = \beta - \mu - T_j/N_j$ . The solution of this partial differential equation can be written as the solution of the following integral equation

$$I_j(t) = I_j(0)e^{\Lambda_j t} + \sum_{\ell} \frac{w_{\ell j}}{N_{\ell}} \int_0^t d\tau I_{\ell}(\tau)e^{\Lambda_j(t-\tau)}. \tag{18}$$

This shows in particular that the ratio  $T_j/N_j$  is a relevant variable in the determination of the time behavior of  $I_j$  and in differentiating among the epidemics evolution in different cities. The underlying network structure affects the epidemics evolution also by the heterogeneity of the connectivity pattern and the weight distribution through the second term of the above equation. This term contains the

ratios  $w_{\ell j}/N_{\ell}$ , thus determining the number of the couplings of  $I_j$  with the infection of other cities and the strength of these couplings. A heterogeneous behavior for the infection behavior might therefore find its origin both in a heterogeneous connectivity pattern as well as in a heterogeneous traffic flows distribution. This is a striking evidence of the intricate nature of the interplay between the various levels of heterogeneity in the system, that contribute simultaneously to the dynamical behavior of the epidemic.

## 5. Epidemic heterogeneity and network structure

While the analysis of the deterministic equations provides a qualitative insight on the role of the network structure on the heterogeneity of the epidemic pattern, we aim at a more quantitative analysis able to discriminate the effect of the specific structural properties of the WAN. Indeed, this heterogeneity might find its origin just in the stochastic nature of the infectious process or be determined by the structural properties of the transportation network. In the latter case, it is possible to envision the possibility of a larger predictability of the epidemic behavior that would reflect the underlying network structure. In order to quantify the heterogeneity of the epidemic spread at a global level, we monitor for the  $V$  cities the behavior of the prevalence in time,  $i_j(t) = I_j(t)/N_j$ , and define the normalized vector  $\vec{\rho}(t)$  with components  $\rho_j(t) = i_j(t)/\sum_{\ell} i_{\ell}(t)$  which contains the relevant information on the epidemic pattern at time  $t$ . The level of heterogeneity of the disease prevalence is then measured by quantifying the disorder encoded in the vector  $\vec{\rho}(t)$ . In order to do that, we use for the first time the normalized entropy function

$$H(t) = -\frac{1}{\log V} \sum_j \rho_j(t) \log \rho_j(t) \quad (19)$$

that is customarily used in information theory to quantify the level of disorder of a signal or system. If the epidemics is homogeneously distributed among all nodes (i.e. all cities have the same value of prevalence at the same time), the entropy is  $H = 1$ . On the other hand, the most heterogeneous situation corresponds to only one city being infected while all the others are not infected, thus leading to an entropy equal to zero,  $H = 0$ . It is important to stress that in the present context the entropy does not have any thermodynamical implications. It must be just considered as the appropriate mathematical tool able to quantify the statistical disorder of a complicate spatiotemporal signal. Without being related to any specific biological meaning, this quantity is able to provide a measure of the geographic heterogeneity of the disease pattern at any given time  $t$ .

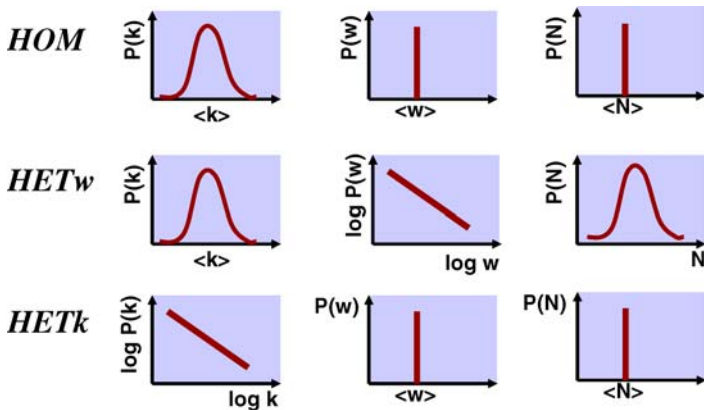
To uncover the effect of the network structure, we will compare the results obtained on the actual network with those obtained on three different network models, which we will denote by HOM, HET $w$  and HET $k$ . The first one (HOM) is a homogeneous Erdős-Rényi random graph with the same number of vertices  $V$  as the WAN, and is obtained as follows: for each pair of vertices  $(j, \ell)$ , an edge is drawn independently, with uniform probability  $p = \langle k \rangle / V$  where  $\langle k \rangle$  is the average



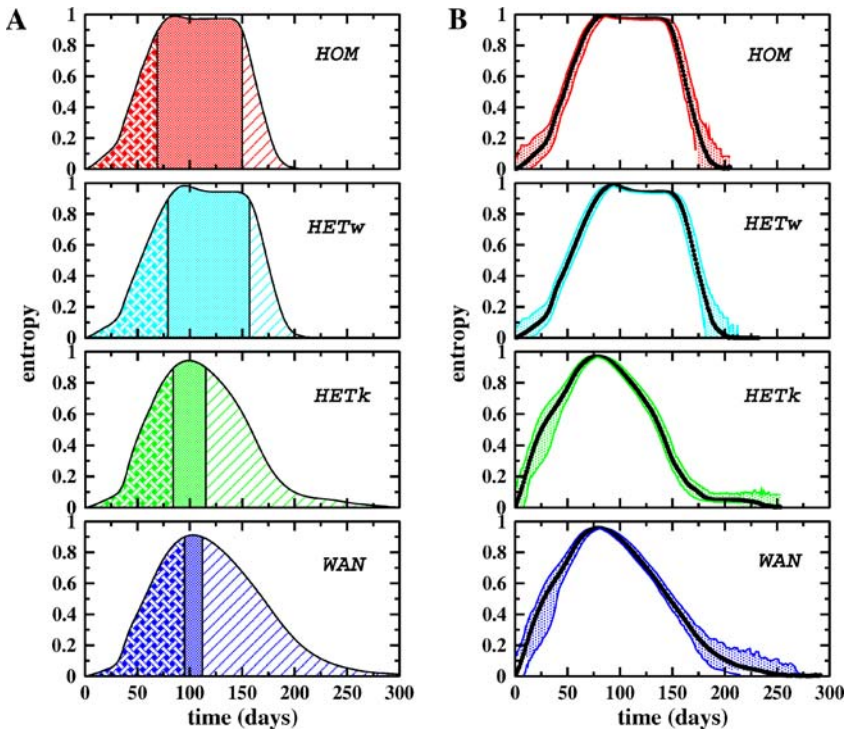
degree of the WAN. A typical instance of a random graph is thus obtained, characterized by a Poissonian degree distribution, peaked around the average value  $\langle k \rangle$  and decreasing faster than exponentially at large degree values, in strong contrast with the true degree distribution of the WAN. Moreover, the weights of the links and the city populations are taken as uniform and equal to their average values in the WAN. The second network model (HETw) is again characterized by a homogeneous connectivity pattern, but has heterogeneous traffic values between airports. In particular, the weights are distributed according to the same skewed distribution  $P(w)$  as in the real case. To the urban area surrounding each airport we assign a population value as obtained from the non-linear relation found in the WAN between the traffic  $T$  handled by each airport and its population  $N$ , i.e.  $T \propto N^\alpha$  with  $\alpha \simeq 0.5$ . Finally, for the third model (HETk), we consider the real topology of the WAN, but weights and populations are taken uniform and equal to the average values  $\langle w \rangle$  and  $\langle N \rangle$  obtained from the distributions of the WAN,  $P(w)$  and  $P(N)$ , respectively. A schematic representation of the probability distributions of degrees, weights and population values adopted for each network model is shown in Fig. 6.

Comparisons between the actual air-transportation network and the three null models yield insight on the role distinctly played by each level of heterogeneity. Starting from the complete absence of heterogeneity both in the topology and in the flux/population data in HOM, the various levels of heterogeneity are investigated separately: HETw incorporates only the large fluctuations of weights and populations, while HETk allows to study the role of a broad degree distribution in the epidemic spread pattern. The real network WAN contains finally all heterogeneities.

We show in Fig. 7 the results obtained on the actual network and those obtained on the three different network models. Differences in the behavior observed in the null cases and in the real case provide striking evidence for a direct relation between the network structure and the epidemic pattern. The completely homogeneous network HOM displays a homogeneous evolution (with  $H \approx 1$ ) of the

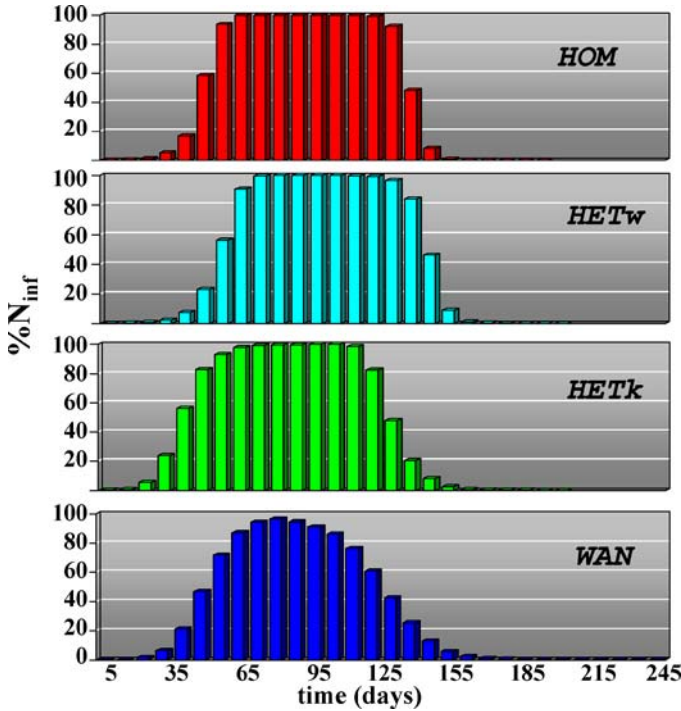


**Fig. 6** Schematic representation of the probability distributions characterizing the network models considered. For each network model  $P(k)$ ,  $P(w)$  and  $P(N)$  are shown.



**Fig. 7** Analysis of the heterogeneity of the epidemic pattern in the actual network (WAN) compared with the three network models HOM, HET $w$  and HET $k$ . (A) Entropy  $H(t)$  averaged over distinct initial infected cities and over noise realizations. Each profile is divided into three different phases each represented by different patterns, the central one corresponding to  $H > 0.9$ , i.e. to a homogeneous geographical spread of the disease. This phase is much longer for the network models characterized by a homogeneous degree distribution—HOM and HET $w$ —than for the real airport network. The behavior observed in HET $k$  is strikingly close to the real case, thus meaning that the connectivity pattern plays a crucial role in the epidemic behavior. (B) Average value of the entropy, with the maximal dispersion (shaded area) obtained from  $2 \times 10^2$  noise realizations of an epidemics starting in a given city (Hong Kong for the WAN). Fluctuations have a mild effect in all cases.

epidemics during a long time window, with sharp changes at the beginning and at the end of the spread. A similar picture is obtained when considering HET $w$ , i.e. a network characterized by a homogeneous connectivity pattern, but with traffic capacities and populations which are broadly distributed. We observe a different scenario for topologically heterogeneous networks where  $H$  is significantly smaller than one for most of the time, with long tails signaling a long lasting heterogeneity of the epidemic behavior. The previous analytical inspection of the deterministic epidemic equations has shown that the broad variability of the contact pattern (degree distribution) and  $w_{j\ell}/N_j$  play an important role in the heterogeneity of the spreading pattern. Strikingly, networks sharing the same connectivity pattern, but with remarkably different traffic and population distributions (such as HOM/HET $w$  or HET $k$ /WAN), display very similar entropic profiles. Moreover, differences in the nature of the degree distribution of the underlying network lead



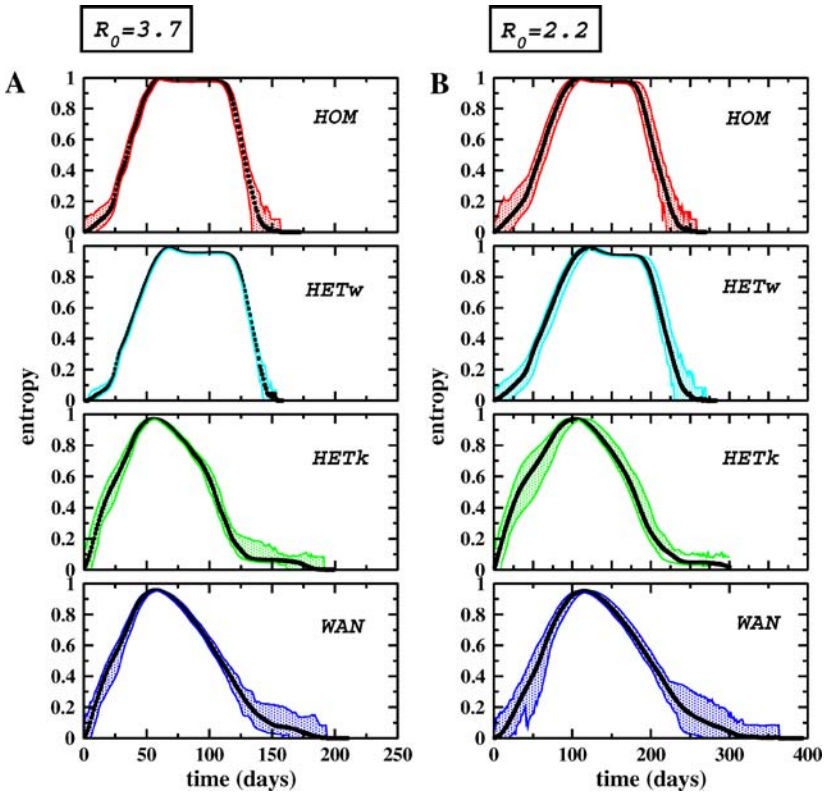
**Fig. 8** Percentage  $N_{inf}$  of infected cities as a function of time. The HOM and the  $HET_w$  cases display a large interval in which all cities are infected. The  $HET_k$  and the real case show a smoother profile with long tails, signature of a long lasting geographical heterogeneity of the epidemic diffusion.

to clearly different results in the spatiotemporal heterogeneity of the spreading pattern, thus pointing out that the broad distribution of degrees of the airport network represents a crucial feature in determining the overall properties of the epidemic pattern.

In Fig. 7B we also show the results obtained for the spreading starting from a given city, by reporting the average entropy profile together with the maximal dispersion obtained from different realizations of the noise. It is clear that the noise has a mild effect and that the average behavior of the entropy is very representative of the behavior obtained in each realization.

Figure 8 displays the evolution of the percentage  $N_{inf}$  of infected cities as a function of time. For the HOM and the  $HET_w$  case, a large interval in which all cities are infected is observed, confirming the occurrence of a long lasting homogeneous infection pattern. In contrast, the  $HET_k$  and the real case show a smoother profile with long tails, corresponding to a geographical heterogeneity of the epidemic diffusion.

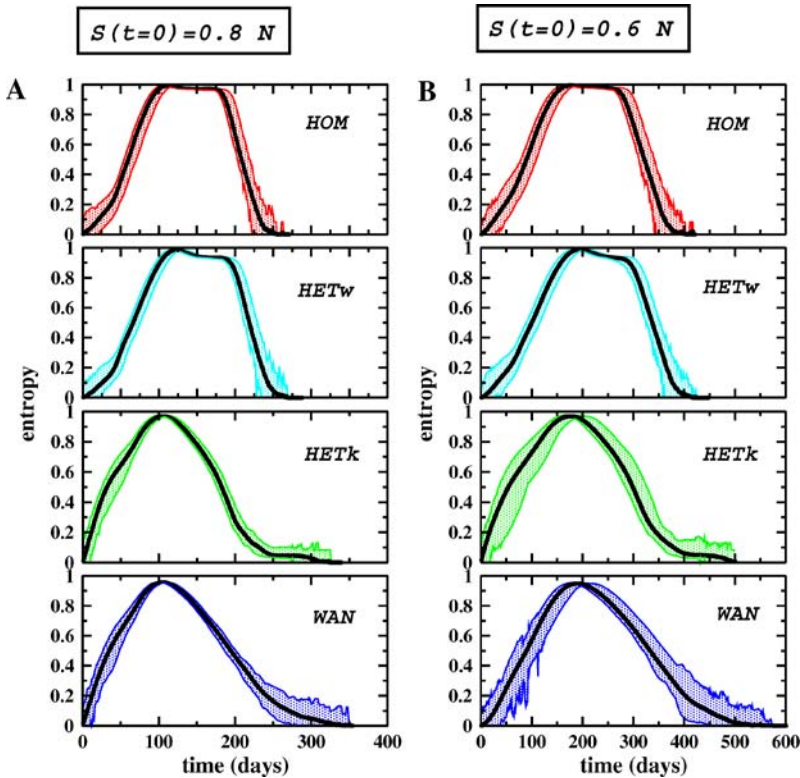
We have investigated different values of the parameters  $\beta$  and  $\mu$  and different initial conditions in order to test the reliability of the results obtained concerning the heterogeneity level of the spreading pattern. In Fig. 9 we show the entropy profile for two epidemic diseases starting in a given city, namely Hong Kong for the



**Fig. 9** Reliability of entropic profiles with respect to changes in  $R_0$ . Entropy profiles for two different disease strains starting in a given city (Hong Kong for the WAN) characterized by two different values of the reproductive number  $R_0$ . Values are those estimated for the SARS outbreak in Hong Kong: A,  $R_0 = 3.7$ ; B,  $R_0 = 2.2$ . From top to bottom: HOM, HET $w$ , HET $k$ , WAN.

WAN, and characterized by two different values of the reproductive number  $R_0$ . We consider low to moderate transmissibility, as estimated for the early epidemic stage of the SARS outbreak in Hong Kong (Lipsitch et al., 2003), with values equal to 2.2 and 3.7. For each value, the three null models—HOM, HET $w$ , and HET $k$ —and the real case (WAN) are shown. Changes in the parameter values clearly leads to different time scales for the global spread, but do not affect the overall conclusions regarding the geographical heterogeneity of the epidemic diffusion pattern.

We have also studied the effect of different initial conditions, such as different initial infected cities and several different initial fractions of susceptible population. In the following, we compare results obtained with two different values of the initial percentage of susceptibles in each city, namely 80% and 60%. The propagation time scales are affected, but the entropy profiles display the same features already discussed for the absence of initial immunity ( $S(t = 0) = N$ ), as shown in Fig. 10. In all cases, HOM and HET $w$  display a strong homogeneity and sharp transitions at the early and final stages of the epidemics, while HET $k$  and WAN



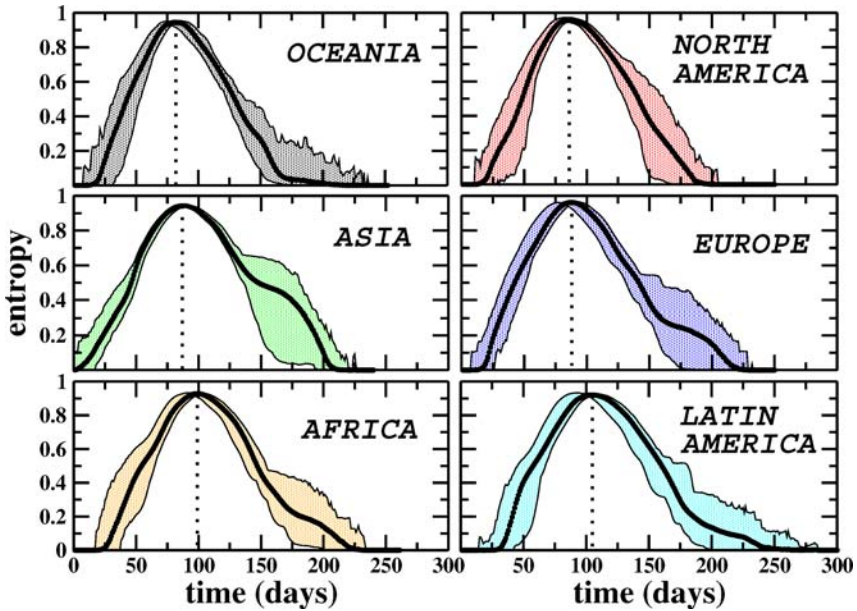
**Fig. 10** Reliability of entropic profiles with respect to the initial immunity. Entropy profiles of an epidemics starting in a given city (Hong Kong in the WAN case), where the initial percentage of susceptible population in each city is assumed to be equal to 80% or 60%. Susceptible fraction values are chosen in agreement with the estimates provided by research studies on different strains of influenza disease: A,  $S_j(t = 0) = 0.8 \cdot N_j$  for all  $j$ ; B,  $S_j(t = 0) = 0.6 \cdot N_j$  for all  $j$ . From top to bottom: HOM, HETw, HETk, WAN.

profiles are characterized by long tails and shorter homogeneous phases ( $H \approx 1$ ), thus confirming the overall results discussed previously.

The entropy profiles shown in Figs. 7, 9 and 10 give a quantitative characterization of the level of heterogeneity of the epidemic pattern observed by considering the complete world-wide airport network. It is also possible to investigate the heterogeneity of the spread on a smaller scale by restricting the sums in Eq. (19) to the cities  $j$  belonging to a given region  $G$

$$H_G(t) = -\frac{1}{\log V_G} \sum_{j \in G} \rho_j(t) \log \rho_j(t), \tag{20}$$

where  $V_G$  is the number of cities in the region  $G$ . As an example, in Fig. 11 we plot the entropy profile restricted to the World continents for a pandemic starting in Hong Kong. The entropic profiles relative to North America and Europe



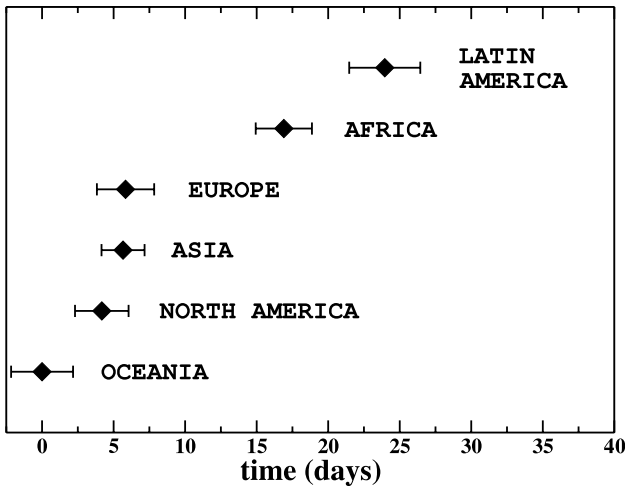
**Fig. 11** Evolution of the entropy restricted to the World continents—North America, Latin America, Europe, Asia, Africa and Oceania—for an epidemics starting in Hong Kong. Shaded areas show the maximal dispersion obtained for various realizations of the noise. The dotted line corresponds to the maximum value of the entropy reached in each region, thus representing the most homogeneous distribution of prevalence among cities belonging to that region.

correspond to the spatiotemporal spread of the disease illustrated in the maps of Figs. 4 and 5, respectively. It is interesting to note that all continents show essentially the same behavior, displaying very similar entropic profiles. There are however small differences and the evolution towards the most homogeneous phase (indicated by the dotted line in Fig. 11) is faster in Oceania, North America, Asia and Europe with respect to Africa and Latin America. Figure 12 shows the entropy peak delay of each continent, i.e. the time period after which the maximum value of their entropic profile is reached since the first entropy peak observed in Oceania. When the epidemic pattern reaches its maximum level of homogeneity in Africa and Latin America, the other continents are already experiencing an increase in the heterogeneity of the epidemic prevalence distribution.

## 6. Predictability and epidemic forecast

A major issue in the modeling of global epidemics is represented by the reliability of the obtained forecasts. In other words it is crucial to answer the question “are epidemics predictable?”. A positive answer to this question amounts to state that stochastic fluctuations are less important than the constraints imposed by the transport network structure that impose an overall pattern to the epidemic evolution. In this respect, global properties provide insights on the general features of

ENTROPY PEAK DELAY



**Fig. 12** Time delay at which each continent reaches the most homogenous phase corresponding to the maximum entropy (dotted line in Fig. 11). The first entropy peak is here observed in Oceania and serves as the initial time  $t = 0$ . Continents are ranked according to their entropy peak delay.

the epidemic spreading in relation with the underlying network but do not provide adequate information on the predictability of the resulting pattern. Indeed, even when global quantities have small fluctuations from one stochastic realization to another, the infected cities might be very different, thus resulting in a very different epidemic scenario. This is a basic issue of global epidemic modeling since the reliability of any epidemic forecast depends on the fact that any given stochastic realization is reasonably similar to the average prediction or, in other words, that any two given stochastic realizations follow similar evolutions.

A quantitative measure of the predictability of the epidemic pattern is therefore given by the statistical similarity of the history of epidemics outbreaks starting with the same initial conditions and subject to different noise realizations. A convenient quantity to monitor in this case is the vector  $\vec{\pi}(t)$  whose components are  $\pi_j(t) = I_j(t) / \sum_l I_l$ ; i.e the normalized probability that an infected individual is in city  $j$ . The similarity between two outbreaks realizations is quantitatively measured by the statistical similarity of two realizations (denoted by  $I$  and  $II$ ) of the global epidemic characterized by the vectors  $\vec{\pi}^I$  and  $\vec{\pi}^{II}$  respectively. There are many possible measures of distributional similarity and we consider here the Hellinger affinity which leads to the similarity

$$\text{sim}_H(\vec{\pi}^I, \vec{\pi}^{II}) = \sum_j \sqrt{\pi_j^I \pi_j^{II}}. \tag{21}$$

This measure has the appealing property of belonging to the interval  $[0, 1]$  and to take the value  $\text{sim}(\vec{\pi}^I, \vec{\pi}^{II}) = 1$  when the two distributions are identical and

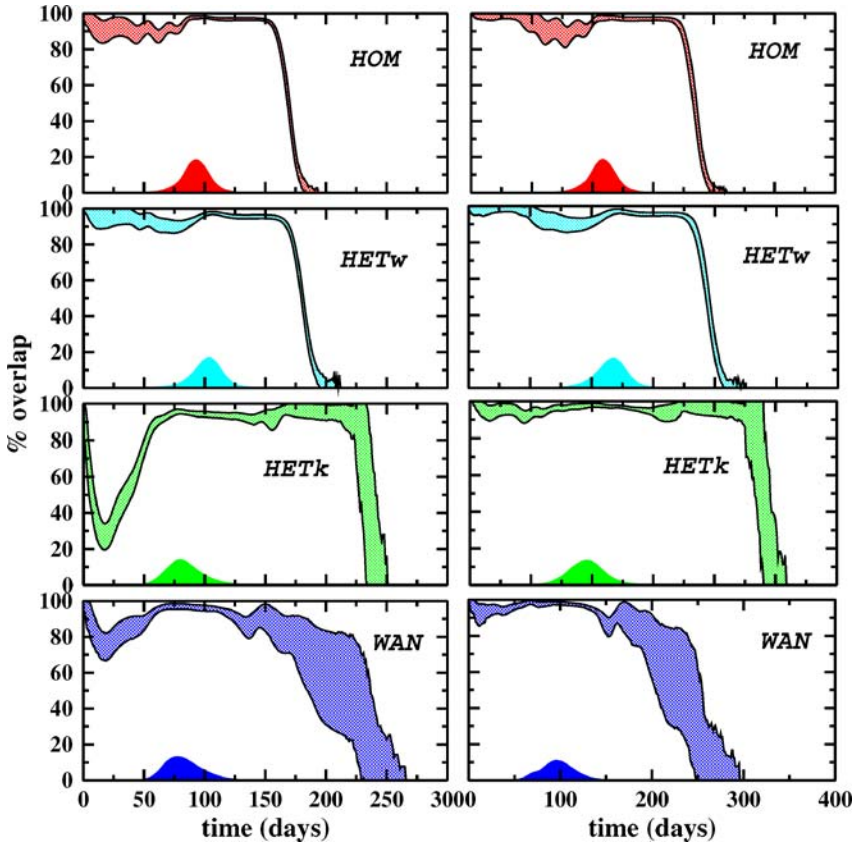
the lowest value  $\text{sim}(\vec{\pi}^I, \vec{\pi}^{II}) = 0$  when there is no overlap at all. While in the following we will report results obtained with the Hellinger distance, we have measured for completeness also another similarity based on the Jensen–Shannon entropy (Wong and Yao, 1987; Lee, 1999) which leads to nearly identical results confirming the independence of our results with respect to the choice of the similarity measure. These normalized similarities measures do not however account for the difference in the total epidemic prevalence. Indeed, normalized partition of infected individuals differing only of a multiplicative factor will generate the same vector  $\vec{\pi}$ . It is thus necessary to consider as well the quantity  $\text{sim}(\vec{i}^I, \vec{i}^{II})$  where  $\vec{i} = (i, 1 - i)$  and  $i(t)$  is the worldwide epidemic prevalence  $\sum_j I_j(t)/\mathcal{N}$ , with  $\mathcal{N} = \sum_j N_j$ . This vector takes into account the similarity of the global prevalence in two stochastic realizations. The overlap function between two different stochastic realizations of an epidemic with the same starting initial conditions and disease parameters is therefore defined as

$$\Theta(t) = \text{sim}_H(\vec{i}^I(t), \vec{i}^{II}(t)) \times \text{sim}_H(\vec{\pi}^I(t), \vec{\pi}^{II}(t)). \quad (22)$$

This overlap is maximal ( $\Theta(t) = 1$ ) when the very same cities have the very same number of infectious individuals in both realizations, and  $\Theta(t) = 0$  if the two realizations do not have any common infected cities at time  $t$ .

The evolution of the overlap during the epidemic spread is shown in Fig. 13 for the various null models and for the WAN. In the HOM and HET<sub>w</sub> case, we find an impressively large overlap ( $\Theta > 80\%$ ) even at the early stage of the epidemics—the most relevant phase for early detection. The picture is different if we consider the HET<sub>k</sub> and the real airport network where two distinct behaviors emerge depending on the degree of the initial infected city. In the case of poorly connected initial cities the overlap is above 80% and decreases only at the very end of the epidemics reflecting the fact that the epidemic behavior is initially constrained to move along the few connections of the starting city, providing an imprint of the outbreak evolution. On the contrary, initial cities with a hub airport generate realizations whose overlap initially decreases to 50–60% (up to 20% in the HET<sub>k</sub>): the possibility of a wider range of possible channels for the epidemic diffusion results in a larger differentiation of the epidemic history in each stochastic realization. By comparing HET<sub>k</sub> with the real case—i.e. networks which share the same connectivity pattern but have different traffic distributions—we observe a higher predictability at the early stage of the epidemics in the WAN than in the null model with homogeneous traffic. The comparison of the overlap results obtained for networks with connectivity pattern of different nature but with same traffic capacities—HOM/HET<sub>k</sub> or HET<sub>w</sub>/WAN—clearly shows the effect of large degree fluctuations in decreasing the predictability of the epidemic pattern, while the emergence of a backbone of dominant connections given by highly heterogeneous weights is evident in the increase of the overlap values obtained for WAN and HET<sub>w</sub> compared to HET<sub>k</sub> and HOM, respectively. We can thus conclude at this stage that the observed behaviors are the result of the balance between two competitive trends. On the one hand, the topological heterogeneity tends to lower the predictability of the global spread of a disease by providing several possible





**Fig. 13** Predictability of the epidemics spread as measured by the average overlap and corresponding standard deviations (obtained with  $5 \times 10^3$  couples of different realizations) versus time. We observe two different behaviors depending on the degree of the initially infected city. The initial decrease of the overlap is larger in the case of airport hubs (left column) compared to poorly connected cities (right column) where few spreading channels are available. We also report the prevalence profile as a function of time showing that the maximum predictability corresponds to the prevalence peak.

channels for the travels of infected individuals. On the other hand, the heterogeneity in the traffic handled by each airport contributes to increase the predictability of the epidemic diffusion by identifying emerging epidemic pathways, i.e. main spreading channels for the spreading disease.

It is important to stress that other factors contribute to the effective predictability of the epidemic evolution. For instance, epidemics that reach a large value of infected individuals are less prone to fluctuations and have a large predictability. For the same reason, it is possible to see that the overlap peak is usually associated with the prevalence peak. On the other hand, the initial and late stage of the epidemics are naturally less predictable and prone to stochastic effects. Finally, the time variability of epidemic parameters due to vaccination and control measure are obviously important factor that might change the overall predictability. The

mathematical tools developed here are therefore providing a probe to assess the epidemic forecast reliability that may be used on each case and specific infection scenario.

## 7. Conclusions and perspectives

From our study, it emerges that the air transportation network properties are responsible for the global pattern of emerging disease. In this perspective, the complex features characterizing this network are the sources of the heterogeneous and seemingly erratic spreading on a global scale of diseases such as SARS. We have also shown that it is possible to obtain quantitative measurements of the predictability of epidemic patterns by providing a tool that might be used to obtain confidence intervals in epidemic forecast and in the risk analysis of containment scenarios. In addition, we were able to relate different levels of predictability with two conflicting effects: on the one hand, the dominant traffic flows which define the main channels of epidemic spreading—the “epidemic pathways”; on the other hand, the underlying heterogeneous connectivity pattern leading to a multiplicity of possible channels. Obviously, in order to make the forecast more realistic, it is necessary to introduce more details in the disease dynamics. Moreover, seasonal effects and geographical heterogeneity in the basic transmission rate (due to different hygienic conditions and health care systems in different countries) should be addressed. Finally, the interrelation of the air transportation network with other transportation systems such as railways and highways could be very useful for forecast on longer time scales. We believe that large-scale mathematical models that take fully into account the complexity of the transportation matrix can be used to obtain detailed forecast of emergent disease outbreaks. This first general and theoretical study thus opens the path to further works including historical case studies of specific diseases such as the SARS, the influenza, etc, and the detailed forecast of future epidemic scenarios. Both the historical and future forecasts may represent a valuable tool to test and optimize traveling restrictions and vaccination policies in the case of new pandemic events.

## Acknowledgements

We thank the International Air Transport Association (IATA) for providing us with the world-wide airline database. We also thank P. De Los Rios, A. Flammini, F. Menczer and S. Schnell for useful discussions and suggestions. A.B. and A.V are partially funded by the European Commission—Fet Open project COSIN IST-2001-33555 and contract 001907 (DELIS).

## Appendix: Transport operator including transit passengers

Here we want to provide an expression for the transport operator that contains also transit passengers. Considering only two-legs travel, and neglecting the

fraction of passengers with more than one stop, we can rewrite the total flux  $w_{j\ell}$  of passengers traveling from city  $j$  to city  $\ell$  as a sum of different contributions (see Fig. A.1): the first one, denoted  $w_{j\ell}^{(1)}$ , corresponds to passengers taking a one leg flight from  $j$  to  $\ell$ ; another contribution,  $w_{j\ell}^{(2)}$ , comes from passengers departing from  $j$  and transiting in  $\ell$  for a connecting flight towards another final destination  $m'$ ; finally, a last term has to take into account passengers departing from an airport  $m$  neighbor of  $j$ , transiting in  $j$  and finally reaching their destination  $\ell$ . For each  $m$  neighbor of  $j$ , this last term is thus proportional to  $w_{mj}^{(2)}$  and to the probability  $w_{j\ell}/T_j$  that passengers from  $m$  take in  $j$  a flight towards  $\ell$ . Finally we obtain

$$w_{j\ell} = w_{j\ell}^{(1)} + w_{j\ell}^{(2)} + \sum_{m \in \mathcal{V}(j), m \neq \ell} w_{mj}^{(2)} \frac{w_{j\ell}}{T_j}. \tag{A.1}$$

We can assume that the number of passengers in transit in  $\ell$  is a small fraction  $\alpha$  of the total traffic between  $j$  and  $\ell$ :  $w_{j\ell}^{(2)} = \alpha w_{j\ell}$ . We then obtain the expression for one leg travel passengers from  $j$  to  $\ell$ :

$$w_{j\ell}^{(1)} \simeq (1 - \alpha)w_{j\ell} - \sum_{m \in \mathcal{V}(j), m \neq \ell} \alpha w_{mj} \frac{w_{j\ell}}{T_j}. \tag{A.2}$$

We now derive the expression for the transport operator with the inclusion of two legs traveling. For the sake of simplicity we will present the derivation for the average transport operator. for a given city  $j$ , passengers of a category  $X$  can leave  $j$  either through one-leg or two-leg traveling, with respective probabilities  $p_{j\ell}^{(1)} = w_{j\ell}^{(1)}/N_j$  and  $p_{j\ell}^{(2)} = w_{j\ell}^{(2)}/N_j$ . This gives an average outflow

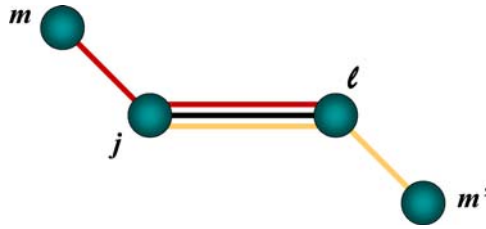
$$\sum_{\ell \in \mathcal{V}(j)} p_{j\ell}^{(1)} X_\ell + \sum_{\ell \in \mathcal{V}(j)} p_{j\ell}^{(2)} X_\ell. \tag{A.3}$$

On the other hand, passengers of category  $X$  arriving to  $j$  can come either directly from neighbors  $\ell$  of  $j$ , with probability  $p_{\ell j}^{(1)}$ , or arrive from farther nodes  $m$  after a transit through these neighbors. The average influx can thus be written as

$$\sum_{\ell \in \mathcal{V}(j)} p_{\ell j}^{(1)} X_\ell + \sum_{\ell \in \mathcal{V}(j)} \sum_{m \in \mathcal{V}(\ell), m \neq j} p_{m\ell}^{(2)} \frac{w_{\ell j}}{T_\ell} X_m. \tag{A.4}$$

Finally the average full transport operator for city  $j$ , including two-leg travels, reads

$$\begin{aligned} \langle \Omega_j(\{X\}) \rangle &= \sum_{\ell \in \mathcal{V}(j)} \left( p_{\ell j}^{(1)} X_\ell - p_{j\ell}^{(1)} X_j \right) \\ &+ \sum_{\ell \in \mathcal{V}(j)} \sum_{m \in \mathcal{V}(\ell), m \neq j} p_{m\ell}^{(2)} \frac{w_{\ell j}}{T_\ell} X_m - \sum_{\ell \in \mathcal{V}(j)} p_{j\ell}^{(2)} X_j. \end{aligned} \tag{A.5}$$



**Fig. A.1 Transit term.** There are three different contributions to the travel from  $j$  to  $l$ : Some individuals travel only from  $j$  to  $l$  (black line), while others make a two-leg travel from either  $m$  to  $l$  (red line) or from  $j$  to  $m'$  (yellow line).

A stochastic version of the two leg transport operator consists in a straightforward generalization of the multinomial stochastic process in which the probabilities for single leg and two legs traveling are separated.

## References

- Albert, R., Barabási, A.-L., 2000. Statistical mechanics of complex networks. *Rev. Mod. Phys.* 74, 47–97.
- Amaral, L.A.N., Scala, A., Barthélemy, M., Stanley, H.E., 2000. Classes of small-world networks. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11149–11152.
- Anderson, R.M., May, R.M., 1992. *Infectious Diseases in Humans*. Oxford University Press, Oxford, p. 4
- Baroyan, O.V., Genchikov, L.A., Rvachev, L.A., Shashkov, V.A., 1969. An attempt at large-scale influenza epidemic modelling by means of a computer. *Bull. Int. Epidemiol. Assoc.* 18, 22–31.
- Barrat, A., Barthélemy, M., Pastor-Satorras, R., Vespignani, A., 2004. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U.S.A.* 101, 3747–3752.
- Chowell, G., Hyman, J.M., Eubank, S., Castillo-Chavez, C., 2003. Scaling laws for the movement of people between locations in a large city. *Phys. Rev. E* 68, 066102.
- Cliff, A., Haggett, P., 2004. Time, travel and infection. *Br. Med. Bull.* 69, 87–99.
- Cohen, M.L., 2000. Changing patterns of infectious disease. *Nature* 406, 762–767.
- Dickman, R., 1994. Numerical study of a field theory for directed percolation. *Phys. Rev. E* 50, 4404–4409.
- Dorogovtsev, S.N., Mendes, J.F.F., 2003. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford.
- Eubank, S., Guclu, H., Anil Kumar, V.S., Marathe, M.V., Srinivasan, A., Toroczkai, Z., Wang, N., 2004. Modelling disease outbreaks in realistic urban social networks. *Nature* 429, 180–184.
- Ferguson, N.M., Keeling, M.J., Edmunds, W.J., Gani, R., Grenfell, B.T., Anderson, R.M., Leach, S., 2003. Planning for smallpox outbreaks. *Nature* 425, 681–685.
- Flahault, A., Valleron, A.-J., 1991. A method for assessing the global spread of HIV-1 infection based on air-travel. *Math. Pop. Studies* 3, 1–11.
- Gardiner W.C., 2004. *Handbook of Stochastic Methods for Physics, Chemistry and Natural Sciences*, 3rd ed. Springer, New York.
- Gastner, M.T., Newman, M.E.J., 2004. Diffusion-based method for producing density-equalizing maps. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7499–7504.
- Gillespie, D.T., 2000. The chemical Langevin equation. *J. Chem. Phys.* 113, 297–306.
- Grais, R.F., Hugh Ellis, J., Glass, G.E., 2003. Assessing the impact of airline travel on the geographic spread of pandemic influenza. *Eur. J. Epidemiol.* 18, 1065–1072.
- Grais, R.F., Hugh Ellis, J., Kress, A., Glass, G.E., 2004. Modeling the spread of annual influenza epidemics in the U.S.: The potential role of air travel. *Health Care Manage. Sci.* 7, 127–134.

- Guimerà, R., Mossa, S., Turtschi, A., Amaral, L.A.N., 2005. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7794–7799.
- Hethcote, H.W., Yorke, J.A., 1984. *Gonorrhea: Transmission Dynamics and Control*. Lecture Notes in Biomathematics 56. Springer-Verlag, Berlin.
- Hufnagel, L., Brockmann, D., Geisel, T., 2004. Forecast and control of epidemics in a globalized world. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15124–15129.
- Institute of Medicine 1992. *Emerging Infections: Microbial Threats to Health in the United States*. National Academy Press, Washington, DC.
- Keeling, M.J., 1999. The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. Lond. B* 266, 859–867.
- Keeling, M.J., et al., 2001. Dynamics of the 2001 UK foot and mouth epidemic: Stochastic dispersal in a heterogeneous landscape. *Science* 294, 813–817.
- Kretzschmar, M., Morris, M., 1996. Measures of concurrency in networks and the spread of infectious disease. *Math. Biosci.* 133, 165–195.
- Lipsitch, M. et al., 2003. Transmission dynamics and control of Severe Acute Respiratory Syndrome. *Science* 300, 1966.
- Lee, L., 1999. Paper presented in the 37th Annual Meeting of the Association for Computational Linguistics, pp. 25–32.
- Lloyd, A.L., May, R.M., 2001. How viruses spread among computers and people. *Science* 292, 1316–1317.
- Longini, I.M., 1988. A mathematical model for predicting the geographic spread of new infectious agents. *Math. Biosci.* 90, 367–383.
- Marro, J., Dickman, R., 1998. *Nonequilibrium Phase Transitions and Critical Phenomena*. Cambridge University Press, Cambridge, UK.
- Meyers, L.A., Pourbohloul, B., Newman, M.E.J., Skowronski, D.M., and Brunham, R.C., 2005. Network theory and SARS: Predicting outbreak diversity. *J. Theor. Biol.* 232, 71–81.
- Murray, J.D., 1993. *Mathematical Biology*, 2nd ed. Springer, New York.
- Pastor-Satorras, R., Vespignani, A., 2001. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.* 86, 3200–3203.
- Pastor-Satorras, R., Vespignani, A., 2003. *Evolution and Structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, Cambridge, UK.
- Riley, S., et al., 2003. Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science* 300, 1961–1966.
- Rvachev, L.A., Longini, I.M., 1985. A mathematical model for the global spread of influenza. *Math. Biosci.* 75, 3–22.
- Wong, S.K.M., Yao Y.Y., 1987. *SIGIR '87: Proceedings of the 10th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, New York, pp. 3–12.
- Zipf, G.K., 1949. *Human Behavior and the Principle of Least Efforts*. Addison-Wesley, Reading, MA.