The role of the airline transportation network in the prediction and predictability of global epidemics

V. Colizza, A. Barrat, M. Barthélemy, A. Vespignani

2006
Heterogeneity and predictability of global epidemics

Vittoria Colizza\textsuperscript{1}, Alain Barrat\textsuperscript{2}, Marc Barthélemy\textsuperscript{1}, and Alessandro Vespignani\textsuperscript{1}

April 26, 2005

\textsuperscript{1} School of Informatics and Biocomplexity Center, Indiana University, Bloomington, 47406, IN, USA
\textsuperscript{2} LPT and UMR du CNRS 8627, Bâtiment 210, Université de Paris-Sud, 91405 ORSAY Cedex - France

The global spread of emergent diseases is inevitably entangled with the structure of the population flows among different geographical regions\textsuperscript{1,2}. 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. Taking advantage of the entire world-wide airline database\textsuperscript{3} along with census data for urban populations, we propose a multi-scale epidemiological framework which allows us to relate the system’s complexity to the disease spreading pattern. We introduce an information theory approach that allows the quantitative characterization of the heterogeneity level of the spreading pattern and its predictability in presence of stochastic fluctuations. The level of spatio-temporal heterogeneity of the spreading pattern is globally characterized and found to be a direct consequence of the network statistical complexity. The epidemic pattern predictability is quantitatively determined and traced back to the occurrence of epidemic pathways defining a backbone of dominant connections.
in the disease spreading. The presented results provide a general framework for the analysis of containment policies and epidemic risk forecast.

The mathematical modeling of epidemics has often dealt with the problem of an appropriate description of real populations with complicated age, social and spatial structures and with heterogeneous patterns in the contact network\textsuperscript{4,5,6,7,8,9}. Recently, the availability of unprecedented computer power has led to numerical approaches relying on agent-based modeling that simulate entire populations and their dynamics at the scale of the single individual and on a minute-by-minute basis\textsuperscript{10,11}. On the other hand, the inherent complex features and emerging properties\textsuperscript{12,13,14} of the network in which epidemics occur are not the mere juxtaposition of complicated elements used for increased realism in sophisticated epidemic modeling\textsuperscript{15}. Indeed, networks’ complex properties often imply statistical fluctuations extending over several orders of magnitude and the breakdown of standard homogeneous approaches and results\textsuperscript{8,9}.

These considerations are particularly relevant in the study of the geographical spread of epidemics where the various long-range heterogeneous connections typical of modern transportation networks naturally give rise to a patchy evolution of epidemics characterized by very distant outbreaks. In this context, air-transportation represents a major channel of epidemic diffusion as recently documented for the SARS outbreak\textsuperscript{16}. The modeling of global epidemic diffusion via the air transportation network dates back to the seminal paper of Rvachev and Longini\textsuperscript{17} capitalizing on previous studies on the Russian network\textsuperscript{18}. Similar modeling approaches, even if limited by a very partial knowledge of the world-wide transportation network, have been used to study specific outbreaks such as pandemic influenza\textsuperscript{19,20,21}, HIV\textsuperscript{22}, and SARS\textsuperscript{23}. The availability of the complete world-wide airport network dataset (WAN) and the recent extensive studies of its topology\textsuperscript{24,25} are finally allowing a full scale computational study of global epidemics. Here, we use the International Air Transport Association (IATA) database\textsuperscript{3} 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. The resulting air-transportation network is therefore a weighted graph, comprising $V = 3,880$ vertices denoting airports and $E = 18,810$ edges.
whose weight \( w_{j\ell} \) represents the passenger flow between airports \( j \) and \( \ell \). This dataset has been complemented by the population \( N_j \) of the metropolitan area served by the airport \( j \) as obtained from different sources (see Supplementary Material). The resulting network 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 (see Fig. 1), contrary to linear relations used so far, indicate a possible major impact in the ensuing disease spreading pattern.

In each city we consider the basic standard compartmentalization\(^4\) of the susceptible-infected-removed (SIR) in which each individual can only exist in discrete states such as susceptible (S), infected (I) or recovered (R), whose corresponding numbers in city \( j \) at time \( t \) are \( S_j(t), I_j(t) \) and \( R_j(t) \) (with \( S_j(t) + I_j(t) + R_j(t) = N_j(t) \) and \( \sum_j N_j(t) = N \)). The dynamics of individuals due to travels between cities is described by the stochastic transport operator \( \Omega_j \{ \{ X \} \} \) representing the net balance of individuals in a given class \( X \) (S, I, or R) that entered and left each city \( j \) (see Methods). This operator is a function of the traffic flows \( w_{j\ell} \) and the city populations \( N_j \) and might also include transit passengers on connecting flights (see Supplementary Material). While we provide as many details as possible on the transportation network, we use the SIR model at the city level in order to provide a general discussion that is not hindered by the use of a very complicated disease transmission mechanism. In this framework, the probability that a susceptible individual acquires the infection from any given infected individual in the time interval \( dt \) is proportional to \( \beta dt \), where \( \beta \) is the transmission parameter that captures the etiology of the infection process. Infected individuals recover with a probability \( \mu dt \), where \( \mu^{-1} \) is the average duration of the infection. The basic reproduction number\(^4\) defined as the average number of secondary cases generated by an infected individual in a susceptible population is \( R_0 = \beta / \mu \). If \( R_0 > 1 \) and if the initial density of susceptibles is larger than \( R_0^{-1} \), then an epidemic will develop in the city.

To describe the disease evolution we consider the standard homogeneous approximations for the SIR model in its Langevin formulation\(^26\) taking into account stochastic fluctuations which
are important in the outbreak evolution. The operator $\Omega_j$ acts as a coupling among the different cities, yielding for each city $j$ the set of dynamical equations

\[
\frac{dS_j}{dt} = -\beta \frac{I_j S_j}{N_j} + \sqrt{\beta \frac{I_j S_j}{N_j}} \eta_{j,1}(t) + \Omega_j(S)
\]

(1)

\[
\frac{dI_j}{dt} = +\beta \frac{I_j S_j}{N_j} - \mu I_j - \sqrt{\beta \frac{I_j S_j}{N_j}} \eta_{j,1}(t) + \sqrt{\mu I_j \eta_{j,2}(t) + \Omega_j(I)}
\]

(2)

\[
\frac{dR_j}{dt} = +\mu I_j - \sqrt{\mu I_j \eta_{j,2}(t) + \Omega_j(R)}
\]

(3)

where $\eta_{j,1}$ and $\eta_{j,2}$ are independent Gaussian white noises. This model is thus a compartmental system of $3880 \times 3$ stochastic differential equations whose integration provides the disease evolution in the corresponding urban areas. In the integration of the equations we also consider the fluctuations of the number of individuals of a given class traveling on each route and of the flow values $w_{j\ell}$. A complete description of the stochastic formulation and the integration scheme is provided in the Supplementary Material.

The geographic spread of an emerging disease starting at different initial conditions is monitored by recording the infection prevalence at different granularity levels (country, state or administrative regions). Indeed, the detailed inclusion of the airport network allows finer scale resolutions and a very detailed analysis of pandemic outbreaks. In Fig. 2 we present the dynamical evolution in the US of a pandemic starting in Hong-Kong. The parameters $\beta$ and $\mu$ are chosen according to\textsuperscript{21} and kept constant during the evolution. Different values do not lead to different overall conclusions. We divide the US according to the nine influenza surveillance regions defined by the Center for Disease Control and use two different visualization strategies. In the first set of maps we simply use a color code corresponding to the prevalence of the infection in each region and in the second set we use cartograms in which the size of each region is proportional to its actual population\textsuperscript{27}. 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 quantitative analysis of the heterogeneity of the epidemic evolution relies on monitoring the relative prevalence in each city $j$ and analyzing the obtained sequence by using the entropy
function $H$ customarily used in information theory. The entropy measures at each time step $t$ the disorder of the sequence thus characterizing quantitatively the level of geographical heterogeneity of the disease prevalence (see Methods). Starting from $H = 0$ which corresponds to one initial infected city - the most localized and heterogeneous situation - $H(t)$ increases as more cities become infected thus reducing the level of heterogeneity (see Fig. 3a). In order to uncover the effect of the network structure on the epidemic pattern heterogeneity, we compare the results obtained on the actual network with those obtained on different network models providing null hypotheses (see Fig. 3). In particular, we consider a fully homogeneous network ($HOMN$) with homogeneous connectivity, traffic and population values, and a network ($HETN$) with the connectivity pattern of the real network but with homogeneous fluxes and populations chosen to be equal to the corresponding averages in the actual air-transportation network. The differences in the behavior observed in the $HOMN$, the $HETN$ and in the real case provide striking evidence for a direct relation between the network structure and the epidemic pattern. The homogeneous network displays a homogeneous evolution (with $H \approx 1$) of the epidemics during a long time window, with sharp changes at the beginning and at the end of the spread. We observe a different scenario for heterogeneous networks where $H$ is significantly smaller than one most of the time, with long tails signalling a long lasting heterogeneity of the epidemic behavior. Indeed, the analytical inspection of the epidemic equations points out 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 (see supplementary material). Strikingly, the curves obtained for both the real network and the $HETN$ are similar, pointing out that in the case of the airport network the broad nature of the degree distribution determines to a large extent the overall properties of the epidemic pattern. In Fig. 3b,c,d we report results obtained for the spreading starting from a given city. In particular, we report in Fig. 3b 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.
A further major question in the modeling of global epidemics consists in providing adequate information on the predictability of the resulting pattern. A measure of the predictability of the epidemic pattern is yielded by the statistical similarity of the history of epidemics outbreaks starting with the same initial conditions and subject to different noise realizations. The similarity is quantitatively measured by the overlap $\Theta(t)$ between the sequences defined by the partition of infected individuals among cities in two different outbreaks (see Methods). The overlap is maximal ($\Theta(t) = 1$) when the very same cities are infected with the very same prevalence in both realizations, and $\Theta(t) = 0$ if the two realizations do not have any common infected cities at time $t$. In the HOMN we find an significant overlap ($\Theta > 80\%$, see Fig. 4) even at the early stage of the epidemics - the one more prone to statistical fluctuations. The picture is different if we consider the HETN and the real airport network where especially at the initial stage of the epidemics the predictability is much smaller. These results may be rationalized by relating the level of predictability to the presence of a backbone of dominant spreading channels defining specific epidemic pathways which are weakly affected by the stochastic noise. Epidemic pathways are the outcome of the conflict between two different properties of the network. On one hand, the heterogeneity of the connectivity pattern provides a multiplicity of equivalent channels for the travel of infected individuals depressing the predictability of the evolution. On the other hand, the heterogeneity of traffic flows introduces dominant connections which select preferential pathways increasing the epidemic predictability. The heterogeneous connectivity pattern of the HETN and the WAN thus generates a multiplicity of channels that decreases the predictability. In the real case the lowering of the epidemic predictability also indicates the dominant effect of the topological heterogeneity that wins over the opposite tendency of the traffic heterogeneity. The above framework is confirmed by the two distinct behaviors depending on the degree of the initial infected city. Epidemics starting in initial cities with a hub airport generate realizations whose overlap initially decreases to 50-60% because of the many possible equivalent paths resulting in a larger differentiation of the epidemic history in each stochastic realization. On the contrary, outbreaks from poorly connected initial cities display a
large overlap due to the few available connections that favor the selection of specific epidemic pathways.

The analysis provided here shows that the present computational approach, augmented with additional features (such as specific disease transmission mechanisms, seasonal effects, geographical heterogeneity of the transmission), may be used to obtain risk analysis of emergent disease outbreaks, especially for the evaluation of control strategies and vaccination deployment.
Methods

The transportation operator

If we neglect multiple legs travels and if we assume 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 rate of passengers in the unit time (in our case 1 day). We can therefore define a transportation operator

$$\Omega_j(\{X\}) = \sum_\ell (\xi_{j\ell}(X_j) - \xi_{\ell j}(X_\ell)),$$

where $\xi_{j\ell}(X_j)$ are stochastic variables extracted from the corresponding multinomial distribution defined in each city $j$ by the traveling probabilities on each connection $p_{j\ell}$ and the number of individuals $X_j$ in the corresponding compartmental class. On the average we have that $\langle \xi_{j\ell} \rangle = p_{j\ell} X_j$ that recovers the deterministic approximation used in Ref. 17. A detailed analysis of the stochastic transport operator and a more refined approach including passengers on connecting flights are described in the supplementary material.

Entropic measure of the geographical heterogeneity

In order to quantify the heterogeneity of the spread at a global level, we introduce for the set of $V$ cities the normalized entropy measure commonly adopted in information theory. We associate to each city $j$ a weight that is given by the normalized prevalence impact $\rho_j / \sum_\ell \rho_\ell$ (with $\rho_j = I_j / N_j$) at time $t$ and the entropy reads as :

$$H(t) = -\frac{1}{\log V} \sum_j \left( \frac{\rho_j}{\sum_\ell \rho_\ell} \right) \log \left( \frac{\rho_j}{\sum_\ell \rho_\ell} \right).$$

If the epidemic is homogeneously distributed among all nodes (i.e. all prevalences are equal), the entropy is $H = 1$. On the other hand, the extreme opposite situation corresponds to only one city being infected which gives an entropy equal to zero $H = 0$.

It is also possible to investigate the heterogeneity of the spread on a smaller scale by restricting the sums in equation (5) to the cities belonging to a given region. For example, Fig. 3d
shows the evolution of the entropy restricted to the US cities, for an epidemic spread starting in Hong Kong.

**The overlap function**

In order to quantify the predictability of the spread, we introduce the overlap function $\Theta(t)$ which is a convenient measure of the similarity of two different stochastic realizations of the pandemic evolution (with the same initial conditions). At a given time $t$, the infection in the system can be characterized by the normalized partition of infected individuals in the set of $V$ cities, identified by the vector $\vec{V}(t)$ with components $V_j(t) = \sqrt{I_j(t)/\sum I_\ell(t)}$ where $I_j(t)$ is the number of infected individual in the city $j$. The scalar product $\vec{V}_I(t) \cdot \vec{V}_{II}(t)$ of the two infection partitions for two different noise realizations is a measure of their similarity where the index $I^{(I)}$ refers to the two different stochastic realizations of the pandemic evolution. The overlap of normalized partitions is however the same in the case of partitions differing only by a multiplicative factor. We take into account the heterogeneity of the total number of infected individuals by the specific function $\Theta_N(t) = \sqrt{p^IP^{II} + \sqrt{(1-p^I)(1-p^{II})}}$, where $p^{I(II)} = \sum_j I_j^{(I)(II)}$ and $N$ is the total population. The global overlap function is therefore defined as

$$\Theta(t) = \Theta_N(t)\vec{V}_I(t) \cdot \vec{V}_{II}(t) = \Theta_N(t) \sum_j \left[ \frac{I_j^I(t)}{\sum I_\ell^I(t)} \frac{I_j^{II}(t)}{\sum I_\ell^{II}(t)} \right].$$

(6)

If the global density of infected individuals is exactly the same in the two realizations $\Theta_N = 1$ and it decreases when differences appear. Similarly, the other term appearing in the definition of $\Theta$ is equal to 1 if the propagation patterns are identical, i.e. if for each city $I_j^I(t) = I_j^{II}(t)$. If the patterns are strongly different in the two realizations, $I_j^I I_j^{II}$ will systematically be small and so will be the overlap.

**References**


**Acknowledgments** M.B. is on leave of absence from CEA-Centre d’Etudes de Bruyères-le-Châtel, Département de Physique Théorique et Appliquée BP12, 91680 Bruyères-Le-Châtel, France. We thank the International Air Transport Association (IATA) for providing us with the world-wide airline database. A.B. and A.V are partially funded by the European Commission - Fet Open project COSIN IST-2001-33555 and contract 001907 (DELIS).
Figure 1: **Properties of the North-American airport network.** Statistical fluctuations are observed over a broad range of length scales. **a, b,** The probability distributions that any airport has $k$ connections to other airports (degree) or handles a number $T$ of passengers (traffic) are heavy tailed. **c,** Analogously, the probability that a connection has a traffic $w$ is skewed and heavy-tailed. **d,** To each airport we can associate a city whose population $N$ is power-law distributed in agreement with the general result of Zipf’s law for city size. **e,** The traffic varies with connectivity as a power law with exponent $1.7$. **f,** The city population varies with the traffic of the corresponding airport as $N \sim T^{\alpha}$ with $\alpha \simeq 0.65$, in contrast with the linear behavior postulated in previous works. Similar results are obtained for the world-wide network. Taking into account the whole network instead of just a fraction is quite relevant: e.g. the 50 (500) largest airports, corresponding to 2 (resp. 12%) of the total network, carry 20% (resp. 80%) of the total traffic, according to the 2002 IATA database.
Figure 2: **Geographical spreading Maps.** Geographical representation of the disease evolution in the US for an epidemics starting in Hong Kong. States are collected according to the nine influenza surveillance regions. The color code corresponds to the prevalence in each region, from 0 to the maximum value reached ($\rho_{\text{max}}$). On the left the original US maps are shown, while on the right we provide the corresponding cartograms obtained by rescaling each region according to its population. Three representations of the airport network restricted to the US are also shown, in correspondence to three different snapshots. The nodes represent the 100 airports in the US with highest traffic $T$; the color is assigned in accordance to the color code adopted for the maps.
Figure 3: **Analysis of the heterogeneity of the epidemic pattern.** The spread of epidemic diseases is investigated in the actual network (WAN) and compared to the results obtained in two different null models: the completely homogeneous network, (HOMN), and a network in which the actual airport network is coupled to homogeneous travel flows (HETN).  

**a,** Entropy $H(t)$ as a function of time, averaged over distinct initial infected cities and over noise realizations. Each profile is divided into three different phases, the central one corresponding to $H > 0.9$, i.e. to a homogeneous geographical spread of the disease. This central phase is much longer for the HOMN than for the real airport network. The behavior observed in HETN is close to the real case meaning that the connectivity pattern plays a leading role in the epidemic behavior.  

**b,** The average value of the entropy is shown together with the maximal dispersion obtained from $2 \cdot 10^2$ noise realizations of an epidemic starting in Hong Kong. Fluctuations have a mild effect in all cases.  

**c,** Percentage of infected cities as a function of time. The HOMN case displays a large interval in which all cities are infected. The HETN and the real case show a smoother profile with long tails, signature of a long lasting geographical heterogeneity of the epidemic diffusion.  

**d,** Evolution of the entropy and of the percentage of infected cities restricted to the US, corresponding to the heterogeneity pattern of figure 2.
Figure 4: **Epidemics predictability.** Average overlap and corresponding standard deviations (obtained with $5 \cdot 10^3$ couples of different realizations) versus time. Topological heterogeneity plays a dominant role in reducing the overlap in the early stage of the epidemics. We observe two different behaviors depending on the degree of the initially infected city: a larger decrease in the case of airport hubs (left) compared to poorly connected cities (right). Large fluctuations at the end of the epidemics are observed in the $HETN$ and in the real case, due to the different lifetime of the epidemic in distinct realizations induced by the heterogeneity of the network. We also report the prevalence profile as a function of time showing that the maximum predictability corresponds to a prevalence peak.