www.sciencemag.org/cgi/content/full/342/6164/1337/DC1

Supplementary Materials for

The Hidden Geometry of Complex, Network-Driven Contagion Phenomena

Dirk Brockmann* and Dirk Helbing

*Corresponding author. E-mail: dirk.brockmann@hu-berlin.de

Published 13 December 2013, *Science* **342**, 1337 (2013)
DOI: 10.1126/science.1245200

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/342/6164/1337/DC1)

Movies S1 to S3

Supplementary Materials for

The hidden geometry of complex, network-driven contagion phenomena

Dirk Brockmann & Dirk Helbing

correspondence to: dirk.brockmann@hu-berlin.de

This PDF file includes:

- Supplementary Text
- Figs. S1 to S20
- Tables S1 to S4
- Captions for Movies S1 to S3

Other Suplementary Materials for this manuscript includes the following:

- Movies S1 to S3

Supplementary Text

The global mobility network

The global mobility network (GMN) used in the main text (see Fig.1) is constructed from data provided by OAG (Official Airline Guide) Ltd. (http://www.oag.com) [45]. The dataset includes a total of 4,069 airports and the number of seats on scheduled commercial flights between pairs of airports over a period of three years, see also Ref. [39]. Assuming that the number of seats on scheduled commercial flights is proportional to the number of passengers traveling, the data can be represented as a weighted graph in which nodes represent airports and weighted, directed links represent the traffic between them. That is, F_{ij} is the number of passengers that travel from airport i to airport j per day. The global mobility network represented by matrix $\mathbf{F}$ exhibits a high degree of statistical symmetry, $F_{ij} \approx F_{ji}$ ($R^2 = 0.9979$). For

the sake of simplicity, we consider a strictly symmetric, undirected network using the symmetrization operation

$$F_{ij} \to (F_{ij} + F_{ji}). \tag{S1}$$

This way, F_{ij} measures the combined number of individuals traveling between airports i and j. The resulting network consists of $4,069$ nodes and $25,453$ undirected weighted links. The total amount of passengers is

$$\Phi = \sum_{i,j} F_{ij} = 8.91 \times 10^6 \,\text{passengers/day} \tag{S2}$$

with a mean link weight of only $\langle F \rangle \approx 175$ passenger / day, but a large coefficient of variation ($c_V(F) = 2.6$). It is important to note that for the model used in the main text, the absolute passenger flux along direct connection is not important. Our model only depends on relative flux fractions $f_{ij} = F_{ij}/\Phi$. A number of topological characteristics of the network are provided in Table S1, see also Refs. [5,39,40,41,44]. A well-known property of the network is its structural heterogeneity reflected in broad distributions of network characteristic quantities such as the degree k (number of connection of a node), node strength S (total traffic through a node) and link weights F (total traffic across a link), see Fig. S1.

In the analysis and application to the spread of the 2009 H1N1 pandemic (below) we also use a projected network GMNc in which nodes represent entire countries and the links that connect them represent the aggregated traffic between countries. This network has $N = 189$ nodes and $L = 5004$ links.

Derivation of the model

The foundation of the dynamical system in the main text (Eq. 3) is a stochastic SIR model governed by the mass action reaction kinetics

$$I_n + S_n \xrightarrow{\alpha} 2I_n \tag{S3}$$

$$I_n \xrightarrow{\beta} R_n \tag{S4}$$

$$X_m \xrightarrow{w_{nm}} X_n, \tag{S5}$$

where S_n, I_n, R_n denote the states (susceptible, infected, recovered) of an individual in population $n = 1, ..., M$. The first reaction represents disease transmission, the second recovery and the third equation movement of a host from population m to n (X is a placeholder for all three classes of individuals, i.e. S, I and R). The parameters of this system are the population averaged, effective per capita transmission rate α [in units d^{-1}] the population averaged recovery rate β [in units d^{-1}] and the mobility rate w_{nm} [in units d^{-1}][1]. All these rates are probability rates. The conditional probability $p(n, t+\Delta t|m,t)$ of a randomly chosen individual in population m at time t being located at population $n \neq m$ a time Δt later is given by

$$p(n, t+\Delta t|m,t) \approx \Delta t \, w_{nm} \quad n \neq m \tag{S6}$$

for small Δt. The key parameter that controls the magnitude of a local outbreak is the basic reproduction ratio given by

$$R_0 = \alpha/\beta, \tag{S7}$$

i.e. the expected number of secondary infections caused by one infected individual in an entirely susceptible population during the episode of being infectious. We consider a system with constant total population

$$\Omega = \sum_n N_n \tag{S8}$$

where N_n is the size of population n. We consider a system which is equilibrated with respect to the movement kinetics (Eq. (S5)), which imposes a restriction on the rates w_{nm}. A reasonable way to establish an equilibrium with respect to the movement between subpopulations is to assume detailed flux balance

$$F_{nm} = w_{nm} N_m = w_{mn} N_n = F_{mn}, \tag{S9}$$

[1] Here and in the following the unit "d" denotes days.

which means that flux [individuals / time] F_{nm} is identical to the flux in the opposite direction. In principle, the matrix F_{nm} is provided by traffic data and N_m by census data such that the model rates w_{nm} can be computed using the above balance equation. However, although it is straightforward to measure F_{nm}, assessing the effective population is more subtle. The number of individuals that effectively participate in the dispersal N_m is not necessarily the same as population data provided by census. We will discuss this point in more depth below and show that a plausible assumption can eliminate the need to know N_m.

Based on the above, the mean-field dynamics for the expected number of infecteds I_n, susceptibles S_n and recovered R_n (recycling the symbols S, I, R for the number of individuals in a given state) are given by

$$\begin{aligned} \partial_t I_n &= \alpha S_n I_n/N_n - \beta I_n + \sum_{m\neq n} (w_{nm} I_m - w_{mn} I_n) \\ \partial_t S_n &= -\alpha S_n I_n/N_n + \sum_{m\neq n} (w_{nm} S_m - w_{mn} S_n) \\ \partial_t R_n &= \beta I_n + \sum_{m\neq n} (w_{nm} R_m - w_{mn} R_n) \end{aligned} \tag{S10}$$

In order to simplify the system it is helpful to define a population's relative size

$$c_n = N_n/\Omega \tag{S11}$$

and the flux fraction

$$f_{nm} = F_{nm}/\Phi \tag{S12}$$

where

$$\Phi = \sum_{n,m} F_{mn} \tag{S13}$$

is the entire mobility flux (total passengers per day). Eq. (S9) implies $f_{nm} = f_{mn}$ and thus

$$\gamma f_{nm} = w_{nm} c_m, \tag{S14}$$

where γ is the global mobility rate, i.e. the passenger flux per individual in the entire population:

$$\gamma = \Phi/\Omega. \tag{S15}$$

Assuming that approx. 2-6 billion people are *effective* members of all populations together one obtains:

$$\gamma = (1.48 - 4.45) \times 10^{-3} \mathrm{d}^{-1}. \tag{S16}$$

The particular numerical value is not essential for the results discussed in the manuscript. It is important to note, however, that the global mobility rate is at least two orders of magnitude smaller than disease transmission and recovery rates α and β, respectively. The parameter γ is also equal to the expected exit rate with respect to the entire population. Denoting the exit rate at location n by $\omega_n = \sum_m w_{mn}$ we have

$$\begin{aligned} \langle \omega \rangle &= \sum_m \omega_m c_m = \sum_m c_m \left[\sum_k w_{km} \right] = \sum_{k,m} w_{km} c_m \\ &= \frac{1}{\Omega} \sum_{k,m} w_{km} N_m = \frac{1}{\Omega} \sum_{k,m} F_{km} \\ &= \Phi/\Omega = \gamma. \end{aligned} \tag{S17}$$

With this information, Eqs. (S10) are equivalent to

$$\begin{aligned}\partial_t j_n &= \alpha s_n j_n - \beta j_n + \omega_n \sum_{m\neq n} P_{mn}\,(j_m - j_n)\,,\\ \partial_t s_n &= -\alpha s_n j_n + \omega_n \sum_{m\neq n} P_{mn}\,(s_m - s_n)\,,\end{aligned} \tag{S18}$$

where $j_n = I_n/N_n$ and $s_n = S_n/N_n$ are the local fractions of infected and susceptibles. The recovered population can be computed by $r_n = 1 - s_n - j_n$, because the local population sizes are constant. In dynamical system (S18) the matrix

$$P_{mn} = f_{mn}/f_n \quad \text{with} \quad \sum_m P_{mn} = 1 \tag{S19}$$

governs the coupling between populations and quantifies the fraction of individuals that leave n and go to m. P_{mn} can be interpreted as the conditional probability that an individual that left n moved to m. In these equations parameters α, β can be estimated for specific diseases and matrix P_{mn} is provided by the network. Note that absolute traffic flux numbers are not required to quantify P_{mn}. In fact, multiplying the original flux matrix F_{mn} by a constant leaves P_{mn} invariant.

The only obstacle that remains is the population specific exit rate ω_n or, equivalently, the expected dwell time $\tau_n = 1/\omega_n$. In most reaction-diffusion meta-population models these rates are assumed to be uniform, $\omega_n = \omega$, and assigned a plausible numerical value. Interestingly, uniform mobility rates ω_n directly follow from a deeper and plausible assumption that relates airport capacity to the population size of its catchment area. If one requires that the total traffic fraction into (or out of) a node is proportional to the relative population size of the node,

$$F_n = \sum_m F_{mn} \propto N_n, \tag{S20}$$

it follows that

$$f_n = \sum_m f_{mn} = c_n. \tag{S21}$$

In combination with Eq. (S14) we find $\omega_n = \gamma$ for all populations and the dynamical system simplifies to

$$\begin{aligned}\partial_t j_n &= \alpha s_n j_n - \beta j_n + \gamma \sum_{m\neq n} P_{mn}\,(j_m - j_n)\\ \partial_t s_n &= -\alpha s_n j_n + \gamma \sum_{m\neq n} P_{mn}\,(s_m - s_n)\end{aligned} \tag{S22}$$

A few things are important to note:

1. Absolute traffic data is not required to fix the parameters in this dynamical system.
2. Neither relative nor absolute sizes of the populations are required.
3. The impact of flux data and population sizes accumulates in the global mobility rate

$$\gamma = \Phi/\Omega \approx 10^{-3}\mathrm{d}^{-1} - 10^{-2}\mathrm{d}^{-1}. \tag{S23}$$

The dynamical system above describes the dynamics of expectation values in the absence of fluctuations. Fluctuations, however, can play a dominant role, particularly when the number of infecteds are small $I_n \ll N_n$. In this regime fluctuations dominate the dynamics and when e.g. $I_n = \mathcal{O}(1)$ random recovery events can lead to disease extinction which is not captured by the deterministic system. In this case a systematic approach is based on solving the associate master-equation for the process which

yields Eqs. (S22) in the limit of infinite population sizes. However, because the deterministic equations describe the dynamics reasonably well when infecteds have passed a certain threshold and when $R_0 > 1$ a phenomenological modification of the above dynamics is based on the idea of incorporating an effective invasion threshold, assuming that a local epidemic can only take off when I_n exceeds a fixed small fraction of the population [18,30,32], i.e. when

$$I_n/N_n > \varepsilon. \tag{S24}$$

We account for this by changing the dynamical system according to:

$$\begin{aligned} \partial_t j_n &= \alpha(j_n/\varepsilon)s_n j_n - \beta j_n + \gamma \sum_{m\neq n} P_{mn}(j_m - j_n), \\ \partial_t s_n &= -\alpha(j_n/\varepsilon)s_n j_n + \gamma \sum_{m\neq n} P_{mn}(s_m - s_n), \end{aligned} \tag{S25}$$

where we have introduced the modifier function $\alpha(j_n/\varepsilon)$ as a factor in the nonlinear growth terms. We chose a sigmoid Hill-type function

$$\alpha(x) = \alpha \frac{x^\eta}{x^\eta + 1} \tag{S26}$$

as a threshold function with gain parameter $\eta \gg 0$. Other sigmoid functional choices work as well. When $\eta \to \infty$ this function becomes a Heaviside step function. One can get a rough idea of the typical value for ε in a realistic scenario: For example, if $\Omega \approx 10^9$ individuals are distributed across $M \approx 10^3$ populations then $\varepsilon \approx 10^{-6}$. Results discussed in the manuscript are qualitatively robust against variations in this threshold, we have performed analyses for thresholds ranging from 10^{-6} to 10^{-3}.

In summary, the above model has four important, global parameters

$$\alpha, \beta \ll \gamma \quad \text{and} \quad \varepsilon \ll 1. \tag{S27}$$

In addition to these parameters, the transport matrix P_{nm} plays a key role and is is provided by the GMN.

Characteristics of simulated epidemics

The global time course of an epidemic is typically quantified by the global prevalence curve, i.e. the fraction of infecteds as a function of time

$$\mathscr{I}(t) = \frac{1}{\Omega}\sum_n I_n(t) = \sum_n j_n(t) f_n = \sum_n j_n(t) c_n = \mathbf{j}(t)\cdot\mathbf{f}. \tag{S28}$$

For a chosen set of outbreak locations this quantity is depicted in Fig. S3. Two outbreak locations from each regions, one OL with highest capacity in the region, the second with size comparable to the mean of the airports in the region. A characteristic of the time course of $\mathscr{I}(t)$ is a strong dependence on size and regional coordinates of the corresponding OL. Each prevalence curve can be characterized by the epidemic duration T_d defined as the point in time when $\mathscr{I}(t)$ crosses a minimum threshold from above (i.e. after an epidemic has waned), the peak infecteds $\mathscr{I}_\mathrm{max} = \max(\mathscr{I}(t))$ and the peak time T_p defined as the point in time where $\mathscr{I}(T_\mathrm{p}) = \mathscr{I}_\mathrm{max}$. Fig. S3 shows that these quantities depend on the specific choice of OL. To assess this variability quantitatively, we ran simulations for all possible 4069 OLs and measured global characteristics. Distributions of epidemic duration T_d, epidemic peak time T_P, and maximum prevalence $\mathscr{I}_\mathrm{max}$ are shown in Fig. S4. All quantities exhibit significant variability with respect to the ensemble of OLs. For instance, T_d can vary between 70 and 125 days for the choice of rate parameters. The nature of distribution of the quantities is robust with respect to variations of the these parameters.

Fig. S5 shows how typical peak time, duration, and peak infection. As expected peak time, duration decrease with R_0 whereas peak infecteds increase. The variability of global quantities typically increases as R_0 approaches the critical value $R_0 = 1$ from above.

Fig. S6 shows how global properties (epidemic duration, peak time, maximum prevalence) of an epidemic correlate with properties of the OL as for example node degree k_i and node capacity F_i.

Supplementing Fig. 1, Fig. S7 depicts scatter plots of arrival time vs. geographic distance for a set of 16 OLs (a subset of locations depicted in Fig. S3). Additional information on the set if OLs is provided in Tab. S2, for example the total capacity of the OL, the estimated speed of spread based on a linear regression on the correlation of arrival T_a and geographic distance, and the squared correlation coefficient of both quantities. Epidemic parameters in all simulations were: $R_0 = 1.5$, $\beta = 0.285\,\mathrm{d}^{-1}$, $\gamma = 2.8 \times 10^{-3}\,\mathrm{d}^{-1}$ and $\varepsilon = 10^{-6}$.

Most probable path and effective distance

The best way to understand the definition of effective distance (Eq. 4 in the main text) is the following. We can interpret the mobility matrix P_{nm} as a transition matrix of a homogenous stochastic jump process the state of which is the current location $n(t)$ of a randomly moving particle, with an equilibrium probability density $c_n = N_n/\Omega$. The dynamics is governed by the master equation

$$\partial_t p_n = \sum_{m \neq n} \left(w_{nm} p_m - w_{mn} p_n \right). \tag{S29}$$

for the time evolution of the conditional probability $p_n = p(n,t|n_0,t_0)$ of a random walker (that started at n_0 at time t_0) of being located at n at time t. w_{nm} is the transition rate. We can write this as

$$\partial_t p_n = -\gamma p_n + \gamma \sum_m P_{nm} p_m, \tag{S30}$$

because $\sum_m w_{mn} = \gamma_m = \gamma$. We can eliminate the parameter γ by rescaling time and obtain

$$\partial_t p_n = -p_n + \sum_m P_{nm} p_m. \tag{S31}$$

This system can be approximated by the time discrete process

$$p_n(t+1) = \sum_m P_{nm} p_m(t), \tag{S32}$$

where a walker waits for a constant time at a node, and then hops to another node.

Consider now a path $\Gamma = \{n_1, n_2, ..., n_L\}$ of L steps that starts at n_1 and ends at n_L. Fixing the endpoints, we can associate with any path that connects these end points the probability

$$W(\Gamma) = P_{n_L n_{L-1}} \times ..... \times P_{n_2 n_1} = \prod_{i=1}^{L-1} P_{n_{i+1} n_i}. \tag{S33}$$

Note that this probability fulfills

$$W(\Gamma) c_{n_1} = W(-\Gamma) c_{n_L},$$

where $-\Gamma$ denotes the identical path in reverse order. We establish a connection between a high probability of a path and short effective distance from origin to destination by assigning an effective directed length to a link from m to n in the network by

$$l_{nm} = 1 - \log P_{nm}. \tag{S34}$$

Note that because of $P_{mn} \neq P_{nm}$ we also have:

$$l_{nm} \neq l_{mn}. \tag{S35}$$

In fact

$$\begin{aligned} l_{nm} &= 1 - \log P_{nm} \\ &= 1 - \log P_{mn} - \log c_n + \log c_m, \end{aligned} \tag{S36}$$

so

$$l_{nm} - \log c_m = l_{mn} - \log c_n \tag{S37}$$

This implies that if the effective distance from a large population to a small one is larger than in the opposite direction, i.e. if $c_n > c_m$, then if follows that $l_{mn} > l_{nm}$. This captures the intuitive notion that a randomly chosen individual in a large population is less likely to relocate to a small, connected population, than an individual from the small population relocating to the large population (despite the fact that in equilibrium a flux balance exists between both populations). The total effective (directed) length of path Γ is then given by

$$\lambda(\Gamma) = L - \log W(\Gamma). \tag{S38}$$

If W is large then λ is small and also $\lambda \geq L$. This definition takes into account that a walker spends a unit of time at every node. Two paths with different number of legs but certain transitions along each leg have different effective lengths, equal to the number of legs. Note again, that directionality matters for effective lengths

$$\begin{aligned} \lambda(-\Gamma) &= L - \log W(-L) \\ &= L - \log W(L) - \log c_{n_1} + \log c_{n_2} \\ \lambda(-\Gamma) - \log c_{n_L} &= \lambda(\Gamma) - \log c_{n_1} \end{aligned} \tag{S39}$$

Eq. (S38) defines the length of a path with fixed endpoints. Finally, we define the effective distance $d(m|n)$ from a reference node n to any other node m as the minimum of $\lambda(\Gamma)$ with respect to all paths Γ that go from n to m:

$$d(m|n) = \min_{\Gamma} \lambda(\Gamma) \tag{S40}$$

and the shortest path Γ_s as the (generically unique) path of that length. Note that again,

$$d(n|m) \neq d(m|n). \tag{S41}$$

In fact, the shortest paths that connect two endpoints in either direction not necessarily contain the same set of intermediate nodes. One can use the set of shortest paths using the above distance measure to obtain the dynamic diameter of the network as the expectation value of all shortest paths

$$D_{\text{dyn}} = \langle d(m|n) \rangle = \frac{1}{M^2} \sum_{n,m} d(m|n).$$

The dynamic diameter D_{dyn} can be roughly estimated and is related to the topological diameter D_{top} defined by the expected number or legs in a shortest path. If we assume that the average degree in the network is k_0 and that transition probabilities W at each node are roughly distributed equally across the links so that $W \approx 1/k_0$ then

$$D_{\text{dyn}} = D_{\text{top}} + \log k. \tag{S42}$$

For the GMN, the distribution of shortest path distances $p(d)$ is shown in Fig. S9, as well as the distribution conditioned on selected reference nodes, $p(d|n)$. Fig. S8 illustrates graphically the link between the most probable and shortest path.

An important feature of effective distances in strongly heterogeneous networks and a consequence of their asymmetry is shown Fig. S9b. For a fixed reference node, we can compute the average shortest path *to* all other notes as well as the average shortest path *from* all other nodes,

$$d_{\rightarrow}(n) = \frac{1}{M} \sum_m d(m|n) \quad \text{and} \quad d_{\leftarrow}(n) = \frac{1}{M} \sum_m d(n|m),$$

respectively. $d_{\rightarrow}(n)$ quantifies how far the rest of the network is from the perspective of node n, whereas $d_{\leftarrow}(n)$ quantifies how far node n is from the rest of the network. Fig. S9b shows these quantities as a function of node size. We see that from the perspective of a hub the entire network is closer than from the perspective of smaller nodes. However, from the perspective of the network (or a location in it) hubs are equally far away as smaller airports. This has the following two implications

1. If a disease breaks out in a hub, it will spread throughout the network faster than if seeded in a smaller, remote node.

2. However, if we average over all initial outbreak locations, the arrival time at a node is roughly independent of its size. On average, remote places are reached as quickly as hubs.

These properties are very typical for multi-hierarchical, star-like networks such as the GMN.

In the GMN, we expect effective distance to be correlated with geographic distance but due to strong long range connections this correlation is expected to be small. This is confirmed in Fig. S10. Although, on average effective distance increases with geographic distance, this systematic increase is weak, and for a given geographic distance, effective distance exhibits substantial variability.

Spreading patterns using effective distance

The material in this section supplements the results shown in Fig. 2. Fig. S11 depicts temporal snapshots of simulated epidemics (parameters $R_0 = 1.5$, $\beta = 0.2857\,\mathrm{d}^{-1}$, $\gamma = 1.42 \times 10^{-3}\,\mathrm{d}^{-1}$ and $\varepsilon = 10^{-6}$) at times $T = 56$, 70, 84, 98, 112, and 126 days after initial outbreak and for 5 sample epicenters in the effective distance/shortest path tree representation. The concentric spreading pattern is a generic feature in all scenarios. For the same sample OLs as in Fig. S7 it depicts the correlation between effective distance and arrival time. Table S3 compares the correlation coefficients of arrival time and effective distance and of arrival time and geographic distance. The relative residuals of the linear regression and the fidelity for the same OLs is also listed in the table.

Fig. S12 illustrates the strong correlation between effective distance and arrival time and their linear relationship for a sample of 16 different OLs that were chosen from different geographic regions and according to different size categories. Despite their differences, the linear relationship and the predictive fidelity using effective distance to determine arrival times is robust with respect to changes in OL. Tab. S3 provides quantitative data on the correlations.

Fig. S13 compiles evidence that, using the effective distance representation, contagion dynamics on complex networks is equivalent to simple Fisher-type wave propagation. It also provides further evidence for the validity of factor separation, i.e. Eq. (6) in the main text.

The observed patterns of effective wave front dynamics is best understood by first considering a regular 2D lattice and two arbitrary sites A and B separated by some distance. Now consider an infinite ensemble of random walkers that start at site A a subset of which will eventually reach site B. In a spatially continuous system the concentration profile $u(x,t)$ is governed by the diffusion equation

$$\partial_t u = D\partial_x^2 u. \tag{S43}$$

One can now ask, which ones of the individual stochastic paths that start at A and terminate at B have a high probability of occurring. One can show that trajectories that most closely resemble a straight line between points A and B have the highest probability.

A large class of epidemic models naturally yields local exponential growth from an unstable disease free state (e.g. SIR, SIS and related models). This leads to pulled front dynamics, for example in the a spatially extended system

$$\partial_t u = \lambda u(1-u) + D\partial_x^2 u, \tag{S44}$$

where u is the concentration of the quantity of interest is such a system. Local exponential growth in combination with diffusion yields constant speed wave fronts. In these systems, the most probable path that a single random walker can take from one point to another coincides with the path a propagating wave takes if the time scale of local proliferation is much smaller than the time scale of diffusion. This does not mean, however, that the stochastic motion that individual particles take are shortest paths. This condition is typically met in disease dynamics, where the global mobility rate γ is a few orders of magnitude smaller than recovery rates and infection rates.

The key idea behind the approach proposed in the original manuscript is that the shortest probabilistic path from a node A in a network to another node B when considering a simple random walk process still determines the path an epidemic wave front takes as in the spatially homogeneous case. The only difference is that the structure of most probable paths is different and more complex.

Earlier attempts[**?**] to impose a relation between network topology with effective distance measures using standard heuristics[**?**], e.g reciprocal weights,

$$d_{\text{eff}} \sim \frac{1}{w} \tag{S45}$$

neither lead to a strong correlation of effective distance with arrival time, nor do they allow the definition of a constant effective speed. The reason for this is the strong heterogeneity in real transportation networks and their hierarchical structure. Furthermore, the essential breakthrough implied by Eq.(S40) is the derivation from the underlying probabilistic interpretation of matrix P_{ij} in the dynamical system and the implication that non-linear spreading processes are shaped by most probable paths.

What happens if paths are degenerate and a large class of paths exist that acquire probabilities of the same order of magnitude? This could be equivalent to an epidemic taking one of many alternative routes through the networks. While this is true in arbitrary networks, human transportation networks are strongly heterogeneous. Under these conditions the actual spread is dominated by the unique path that has the highest probability. Support for this argument comes from the link salience analysis provided in Sec. , which shows that a small subset of links attract all the shortest paths in all shortest path trees of all nodes. The physical analogy of this is a random process that evolves in a strongly structured external potential, predominantly in the valleys of this potential. In a way, therefore, epidemics on real transportation networks evolve along "classical" paths that can be accounted for by mean-field descriptions. This is, we believe, the underlying reason why deterministic epidemic meta population models have been surprisingly successful in the past. Intuitively, although particles that start at some node A may move along very different paths to reach node B, those that get their first (the ones that moved along the shortest path), will initiate a new outbreak at B before others reach that destination.

Epicenter reconstruction of a real or simulated epidemic, based on network knowledge and a temporal snapshots of the spreading pattern, can be accomplished in various ways. Each method has to quantify (1) how close the effective epidemic wave front is and (2) how much variability exists in distances to the wave front. If both, the average as well as the variance (or standard deviation) of distances to the wave front are small, the chosen location is a viable candidate. For the reconstructions depicted in Fig. 4D and E, we used the following algorithm. The state of the system is given by prevalence $j_n(t)$ at time t. Based on this full information we can identify the front location as the subset of nodes Λ, where the prevalence is larger than some fixed reference value Θ relative to a maximal prevalence $j^\star$:

$$\Lambda = \{n \mid j_n(t)/j^\star > \Theta\} \tag{S46}$$

Once the front-nodes Λ are identified, we can compute the effective distance $d_n(m)$ with $n \in \Lambda$ from the candidate outbreak location m of the set of nodes that make up the front. This gives a set of distances

$$d_\Lambda(m) = \{d_n(m) \mid n \in \Lambda\} \tag{S47}$$

We can then compute the mean and variance of this set,

$$\langle d_\Lambda(m)\rangle = \frac{1}{N_\Lambda}\sum_{n\in\Lambda} d_n(m) \quad \text{and} \quad \left\langle (d_\Lambda(m) - \langle d_\Lambda(m)\rangle)^2 \right\rangle = \frac{1}{N_\Lambda}\sum_{n\in\Lambda} d_n(m)^2 - \langle d_\Lambda(m)\rangle^2 , \tag{S48}$$

where N_Ω is the number of nodes in Λ. If the candidate source nodes m exhibits a small mean $\langle d_F(m)\rangle$ and if the variance is small, the distance to the candidate is small and uniform, corresponding to a concentric ring, this is evidence that m is the correct outbreak location. For an alternative way to determine the outbreak location see Ref.[34].

Application to the 2009 H1N1 Pandemic

As a proof of concept we apply the method of outbreak reconstruction to data of a real global pandemic, the 2009 worldwide spread of the H1N1 influenza A virus (also known as the "swine flu" pandemic). This pandemic emerged most likely in Mexico with the first confirmed cases in the spring of 2009 and subsequent cases in the United States. The epidemic attained pandemic status shortly after many cases were confirmed in the Americas, Europe and Asia. During the time course of 2009 more than 100 countries confirmed cases. The time course of total prevalence as a function of week of 2009 is shown in Fig. S14, showing peaks in the summer and fall of 2009. The time course of the epidemic in a sample of countries is depicted in Fig. S15. To apply our outbreak reconstruction method we compiled temporal data on the weekly numbers of new incidences per country and using the air-traffic flux of passengers between 189 countries in a total of 252 countries. Arrival times of the epidemic in various countries are listed in Tab. S4. We performed the OL reconstruction treating every possible country as a potential candidate and computing the effective distance to the subset of countries that have a prevalence above some fixed value at a given amount of time. The threshold value was chosen such that a sufficient number of countries are always in that set and sufficiently small in order to have well defined fronts (in effective distance). The results are robust against variations of these parameters.

Validation using the global epidemic and mobility model (GLEAM)

In order to approximate a scenario of a controlled experiment and at the same time investigate a system with high dynamical complexity we applied our technique to simulated dynamics obtained using the global epidemic and mobility model (GLEAM) [21]. GLEAM is one of the most sophisticated computer simulation frameworks for the simulation of worldwide pandemics. Unlike the simplified metapopulation model used in the main text, GLEAM incorporates not only global mobility by air transportation but also short to intermediate commuter traffic. Most importantly, GLEAM incorporates stochastic effects in mobility events and allows the simulation of entire ensembles of virtual global epidemics, which can be used to obtain reliable statistics. In addition to the simple SIR compartmental model, more realistic scenarios can be captured with GLEAM, consisting of dozens of different compartments that govern local disease dynamics and differ in mobility behavior, clinical state, and antiviral treatment. To validate our approach and check the robustness our claims, we adapted a substantially more complex disease dynamics, consisting of seven different infectious state compartments, susceptible, latent, four different infected compartments and recovered. Details of this particular model are described in Ref. [21]. The four different classes of infecteds differ in their propensity to interact with susceptibles, their clinical state, their mobility patterns and antiviral treatment. The reaction scheme of this model is outlined in Fig. S17, see also Ref. [21]. The local dynamics is governed by 12 different parameters.

With this setup, we ran simulations with GLEAM for periods of 365 days. We averaged the outcome across 20 stochastic simulations for identical initial conditions (seed outbreak location in Mexico City). Based on this ensemble we tested the key hypothesis of the manuscript, that in terms of effective distance, the components of arrival time decouple into the effective distance that only depends on the underlying transport matrix, and the effective speed that depends on local disease specific parameters, only. The results of this analysis are shown in Fig. S18. The figure depicts different temporal snapshots of the GLEAM-simulated pandemic. The overall patterns is circular. Fluctuations, however, on the leading edge of the wave exist, which is also typical for stochastic ordinary reaction-diffusion systems that tend to display ragged wave fronts.

In analogy to the results presented in Fig. 2 of the main text the expected arrival time vs. effective distance for the ensemble of GLEAM-simulations is shown in Fig. S19. Despite fluctuations, the simulated data is consistent with the claim that effective distance and arrival time exhibit a linear functional relationship.

Predictability, outbreak similarity and shortest path trees

The GMN and related networks are highly redundant in the possible paths travelers can take from a given origin to their destination. The results presented in the manuscript are evidence that the family of pathways a contagion phenomena can take are dominated by the most effective shortest path. This strong dominance of most effective routes over the abundance of alternative routes can be explained in terms of a recently discovered property of shortest paths in strongly heterogeneous networks [29] based on the concept of *link salience*. Given a complex network with weighted links and the set of all shortest path trees (each rooted at one of the network's nodes), link salience is the fraction of shortest path trees a given link is part of. Intuitively, high salience means that a link is important to many nodes in terms of connecting them to the rest of the network. As is shown in Fig. S20, the distribution of link saliences in the GMN is bimodal. This implies that all links either participate in almost all shortest path trees, or in almost none. In combination with the observation that contagion processes predominantly spread through shortest path tree links, this implies that typically only a small fraction of links shape the overall spreading pattern (those links with $s \approx 1$) irrespective of the initial condition. This is why even stochastic epidemic dynamics are surprisingly predictable in strongly heterogeneous networks.

Bibliography

[40] A Barrat, M Barthelemy, R Pastor-Satorras, Alessandro Vespignani. The architecture of complex weighted networks. 101(11):3747–3752, Marzec 2004.

[41] Alain Barrat, Marc Barthelemy, Alessandro Vespignani. The effects of spatial constraints on the evolution of weighted complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(05):P05003, Maj 2005.

[42] Rafael Brune, Christian Thiemann, Dirk Brockmann. Predicting the origin of contagion processes on complex, multi-scale networks. *http://meetings.aps.org/link/BAPS.2012.MAR.H54.5*, 2012.

[43] Guido Caldarelli. *Scale-Free Networks: Complex Webs in Nature and Technology*. Oxford University Press, 2007.

[44] Luca Dall'Asta, Alain Barrat, Marc Barthelemy, Alessandro Vespignani. Vulnerability of weighted networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2006(04):P04006, 2006.

[45] OAG Worldwide Ltd., 2007.

Fig. S1.

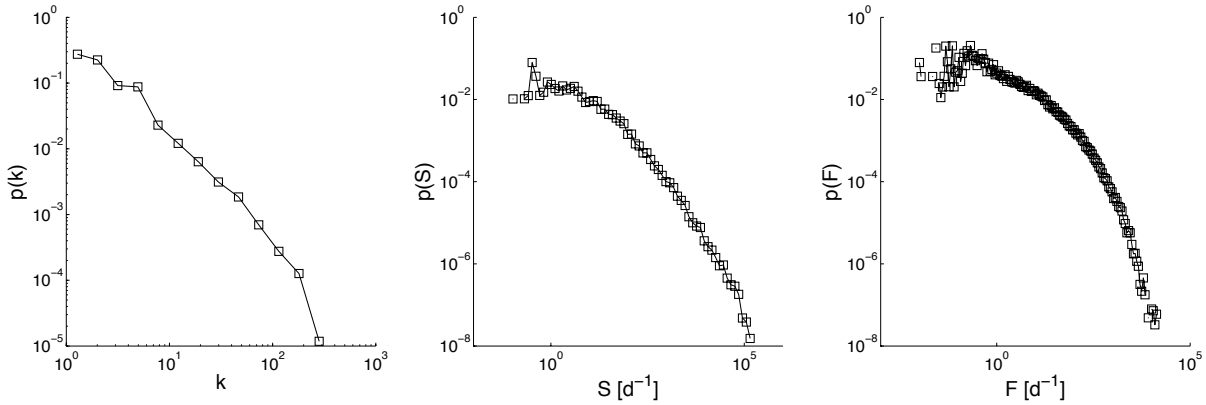

Figure S1: Statistical characteristics of the GMN. From left to right, estimated probability density functions of degree k, node strength S and link flux F in the network. All distributions range across several orders of magnitude, see also Refs. [5,39,40,41,44].

Fig. S2.

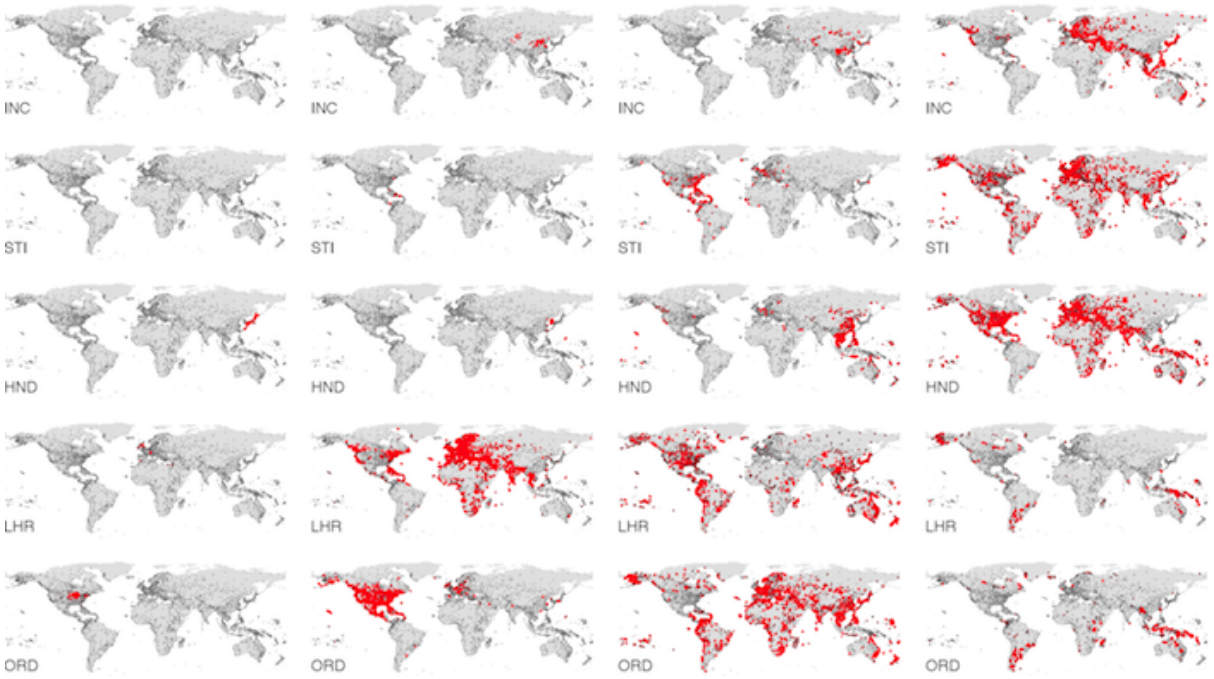

Figure S2: Time course of simulated pandemics. Each row depicts temporal snapshots of a simulated pandemic with different initial outbreak locations labeled by the three letter code of the corresponding airport at times $T = 53, 87, 123, 158$ days. The model parameters are the same in all simulations: $R_0 = 1.5$, $\beta = 0.28\,\mathrm{d}^{-1}$, $\gamma = 2.8 \times 10^{-3}\mathrm{d}^{-1}$, and $\varepsilon = 10^{-6}$.

Fig. S3.

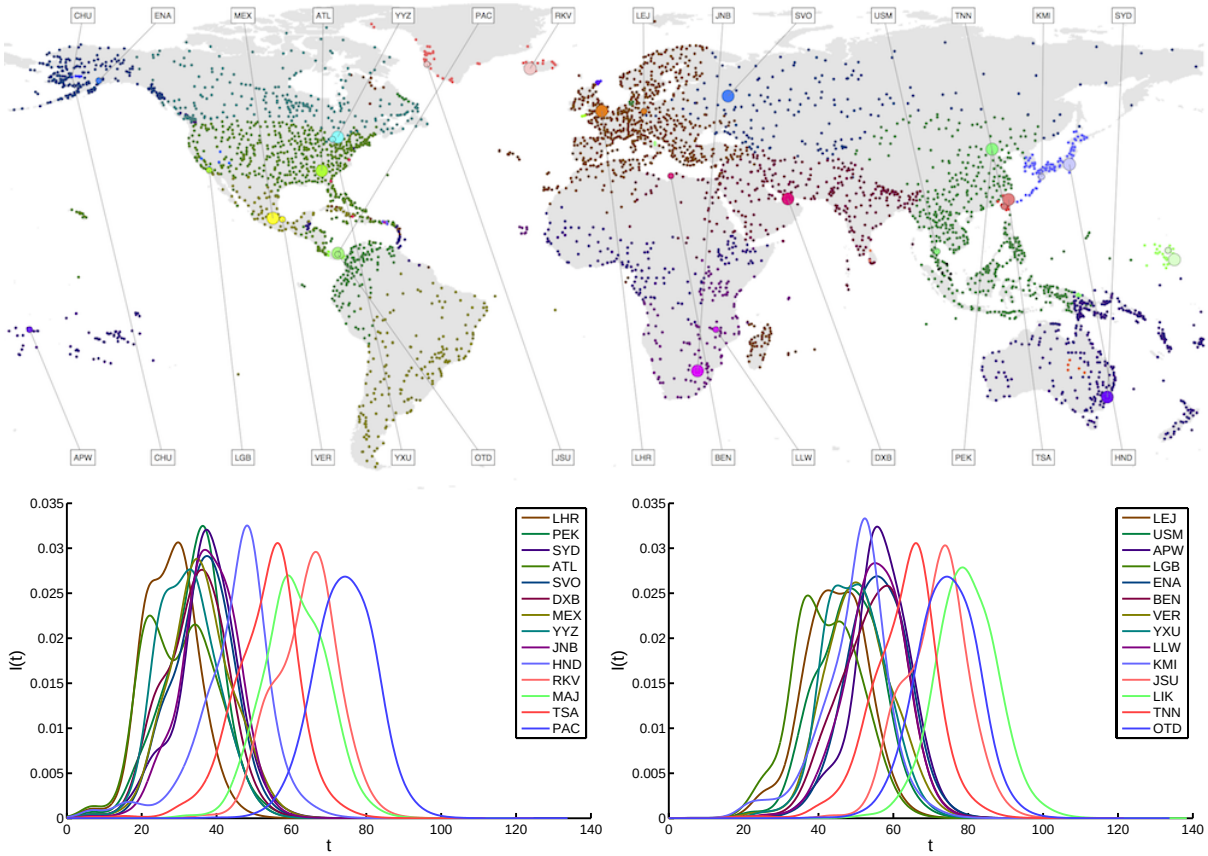

Figure S3: Characteristics of simulated pandemics. *Top:* The map depicts all locations (nodes) in the model, colored by region. Sample OLs were chosen such that each region contributes a high capacity hub (large symbol) and an average sized location (smaller symbol). For each of the 28 OLs, epidemics were simulated. The global prevalence curves $\mathscr{I}(t)$ for each simulation are shown in the bottom panels. *Left:* The 14 largest airports. *Right:* 14 airports of average capacity. Epidemic parameters were identical in each simulation: $R_0 = 1.5$, $\beta = 0.2857\,\mathrm{d}^{-1}$, $\gamma = 1.42 \times 10^{-3}\,\mathrm{d}^{-1}$ and $\varepsilon = 10^{-6}$.

Fig. S4.

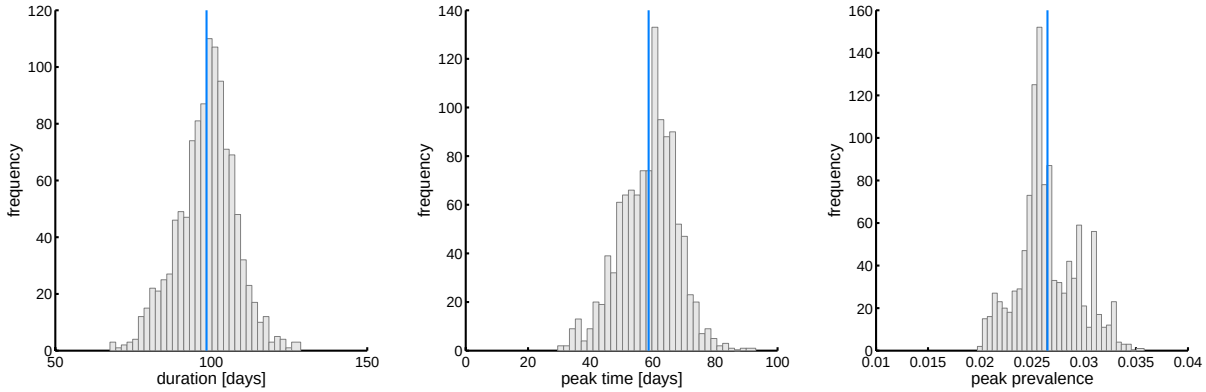

Figure S4: Distribution of global features of simulated epidemics. For each of the $N = 4069$ possible initial conditions an epidemic was simulated and duration T_{d} (left), peak time T_{p} (center), and peak fraction of infecteds $\mathscr{I}_{\max}$ (right) were determined. Frequency with respect to the ensemble of initial conditions are shown in each panel, respectively. Epidemic parameters R_0, β, γ and ε are identical to those in Fig. S3.

Fig. S5.

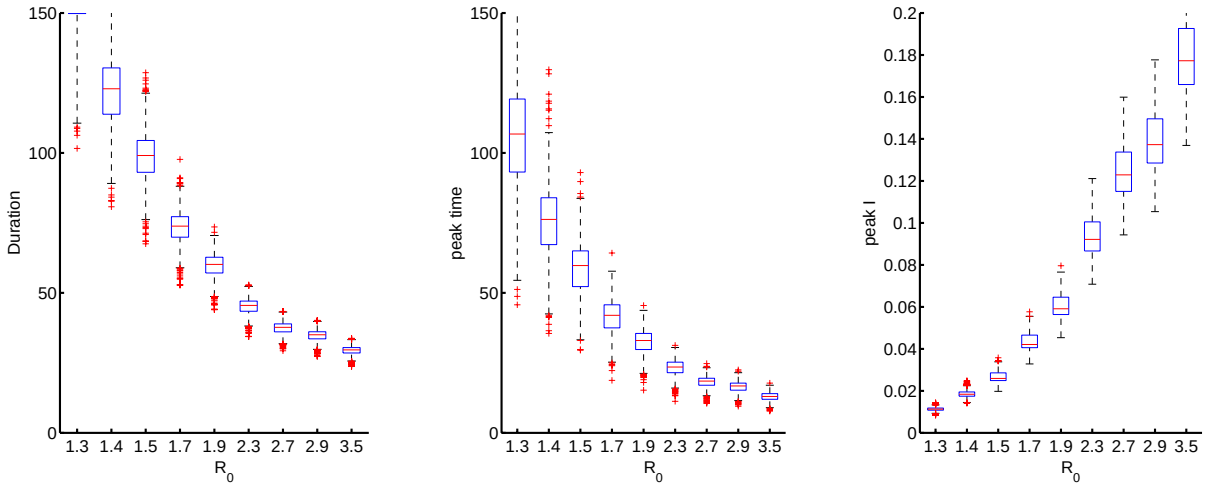

Figure S5: Epidemic properties as a function of the basic reproduction ratio R_0. Epidemic duration, peak time and maximum prevalence were computed for each value of R_0 and all possible OLs. The whiskers depict the distribution of results for fixed R_0. As expected, duration and peak time decrease with R_0 on average, whereas maximum prevalence increases. However, large deviations from the mean exist, in particular when R_0 is small. Other rate parameters in the simulations are $\beta = 0.2857\,\mathrm{d}^{-1}$, $\gamma = 2.8 \times 10^{-3}\mathrm{d}^{-1}$ and $\varepsilon = 10^{-5}$.

Fig. S6.

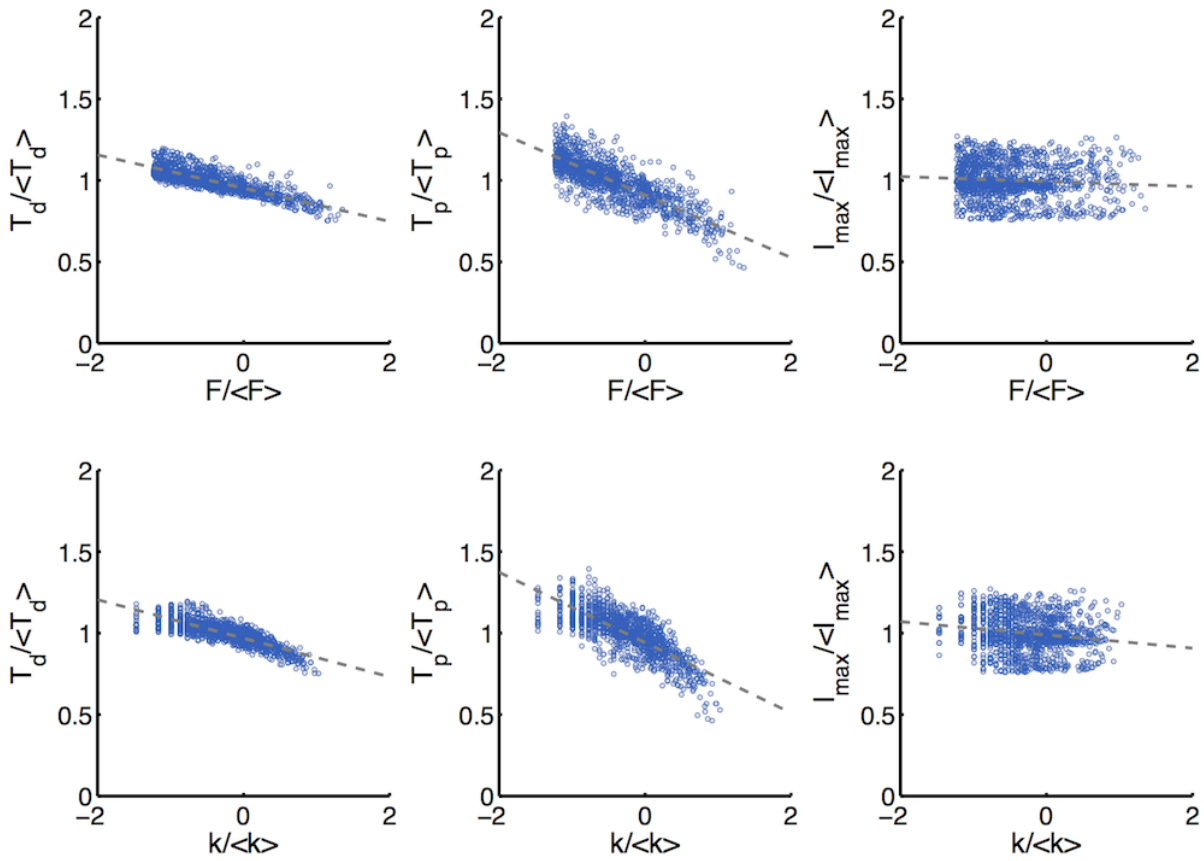

Figure S6: Epidemic properties as a function of node properties: *Top:* The scatterplots show the relationship between node capacity F_i (in units of the its mean $\langle F \rangle$) and duration T_d, peak time T_p, and $\mathscr{I}_\mathrm{max}$ (from left to right). *Bottom:* The relationship of node degree k_i (in units of its mean $\langle k \rangle$) and the same epidemic properties as above. Duration and peak time decrease with F_i and k_i. Peak prevalence is relatively robust with respect to changes in outbreak location properties. Rate parameters were identical to those in Fig. S5.

Fig. S7.

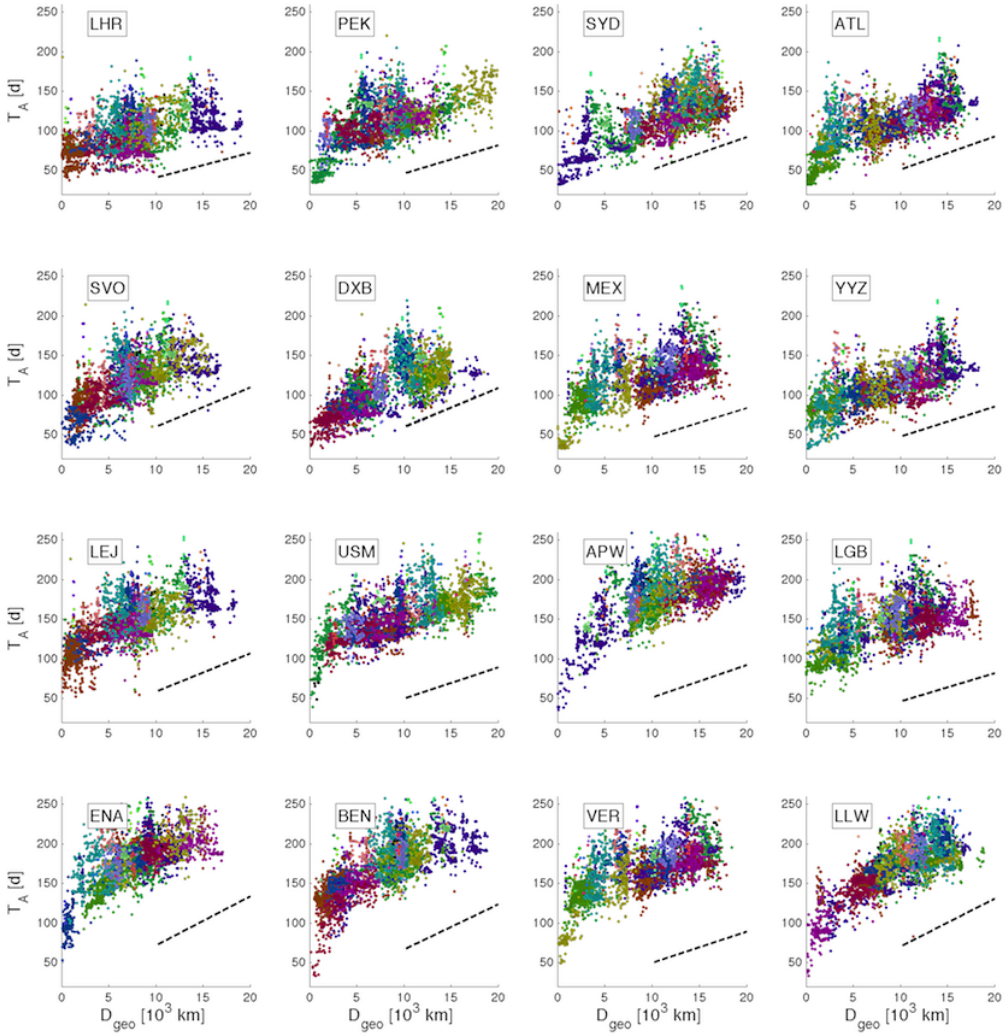

Figure S7: Arrival times vs. geographic distance. The panels depict scatter plots of arrival times T_a in days against geographic distance in km to the initial outbreak location. Symbols are colored according to geographic region. The OL in each panel is denoted by its three letter code, ref. Tab. S2. The dashed lines indicate the slope obtained from a linear regression. The inverse slope can be used as an estimate for the approximate propagation speed, see also Tab. S2. The top two rows are regional hubs (largest capacity nodes in the corresponding region), the bottom two rows are locations that with capacity closest to the regional mean. Parameters of the simulation are the same as for the simulations in Fig. 1.

Fig. S8.

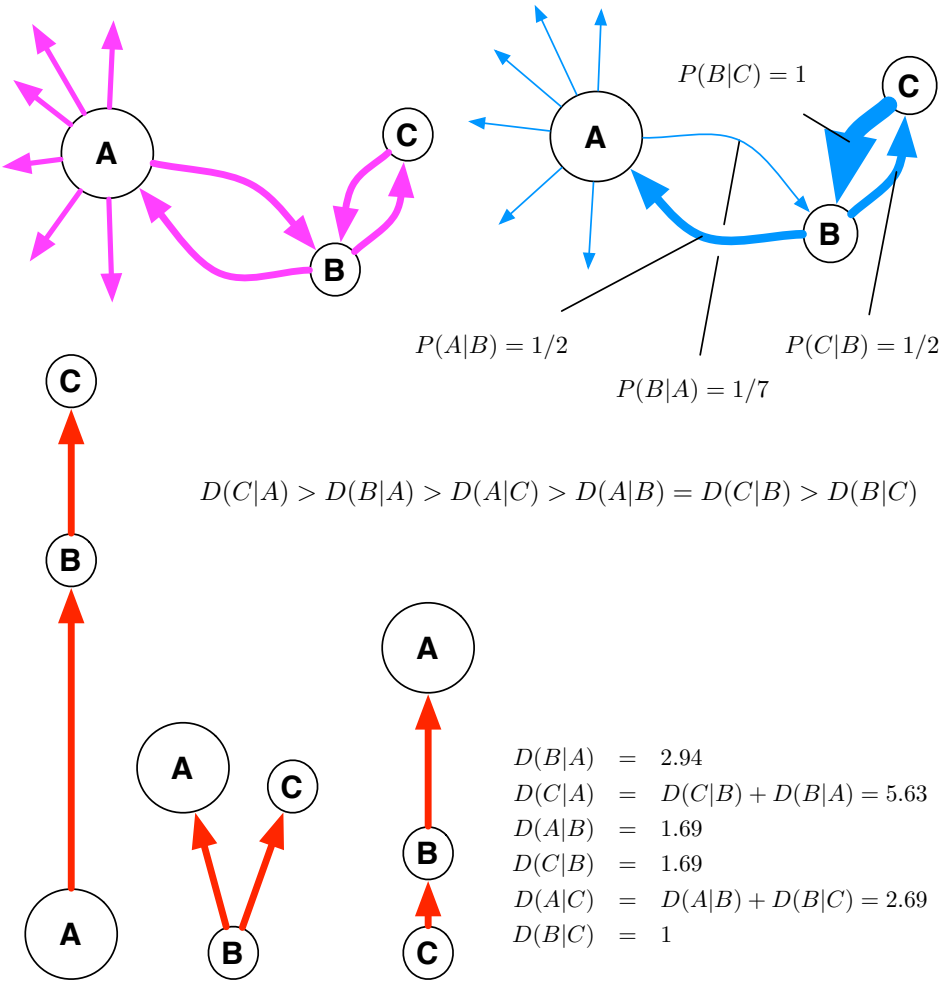

Figure S8: Graphical explanation of shortest paths. *Top left:* Consider a three node network with traffic flux quantified by the purple arrows. For simplicity we assume that the amount of traffic is the same along all links. *Top right:* Based on this we can compute the probabilities $P(n|m)$ of a random walker at node m of moving to n. The magnitude of these probabilities is indicated by line width. For example a random walker starting at A has a smaller probability of going to B than vice versa. *Bottom:* For a network of three nodes, 6 different shortest paths exist. The longest among those is the path from A to B. This is intuitive because a random walker that starts at the hub node A only reaches the remote small node C with a small probability. The path from C to A is smaller, because a random walker placed there will jump to B with certainty and with probability $1/2$ move on to C.

Fig. S9.

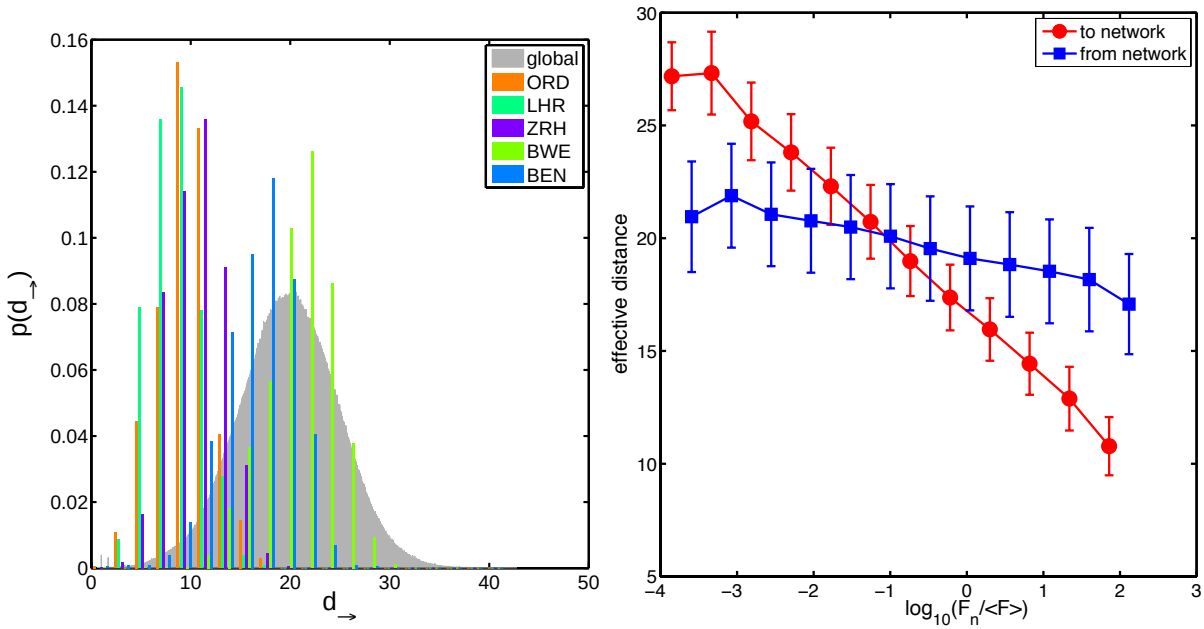

Figure S9: Effective distances in the GMN: *Left:* For the ensemble of all reference nodes, the pdf $p(d_{\rightarrow})$ of effective distances to other nodes is shown (gray), the distribution for a few selected nodes n in color. Note that these distributions are distances *to* other nodes from a reference node n. *Right:* Average (and standard deviation of) effective distance to (red) other nodes and from (blue) other nodes conditioned on the node flux of the reference node. As expected the distance *to* the rest of the network decreases with node capacity F_n. However the typical distance to a node is fairly independent of node capacity. This means: The larger the reference node the easier it is to get somewhere in the network. However, getting to a node from somewhere is independent of the destination node's properties.

Fig. S10.

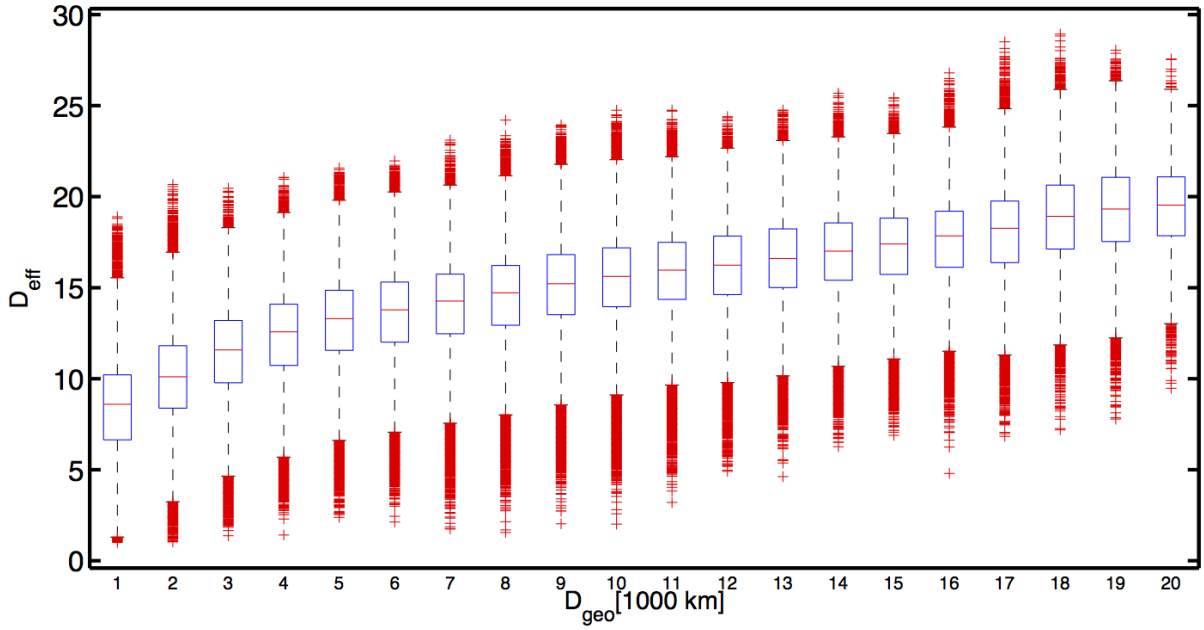

Figure S10: Effective vs. geographic distance. The distribution of effective distances D_{eff} for a sequence of intervals of geographic distances D_{geo} between pairs of nodes. For a given geographic distance interval, the distribution of effective distances almost spans the entire range of D_{eff}. Horizontal red lines are median, blue boxes limit the 25th and 75th percentile, outliers are marked in red.

Fig. S11.

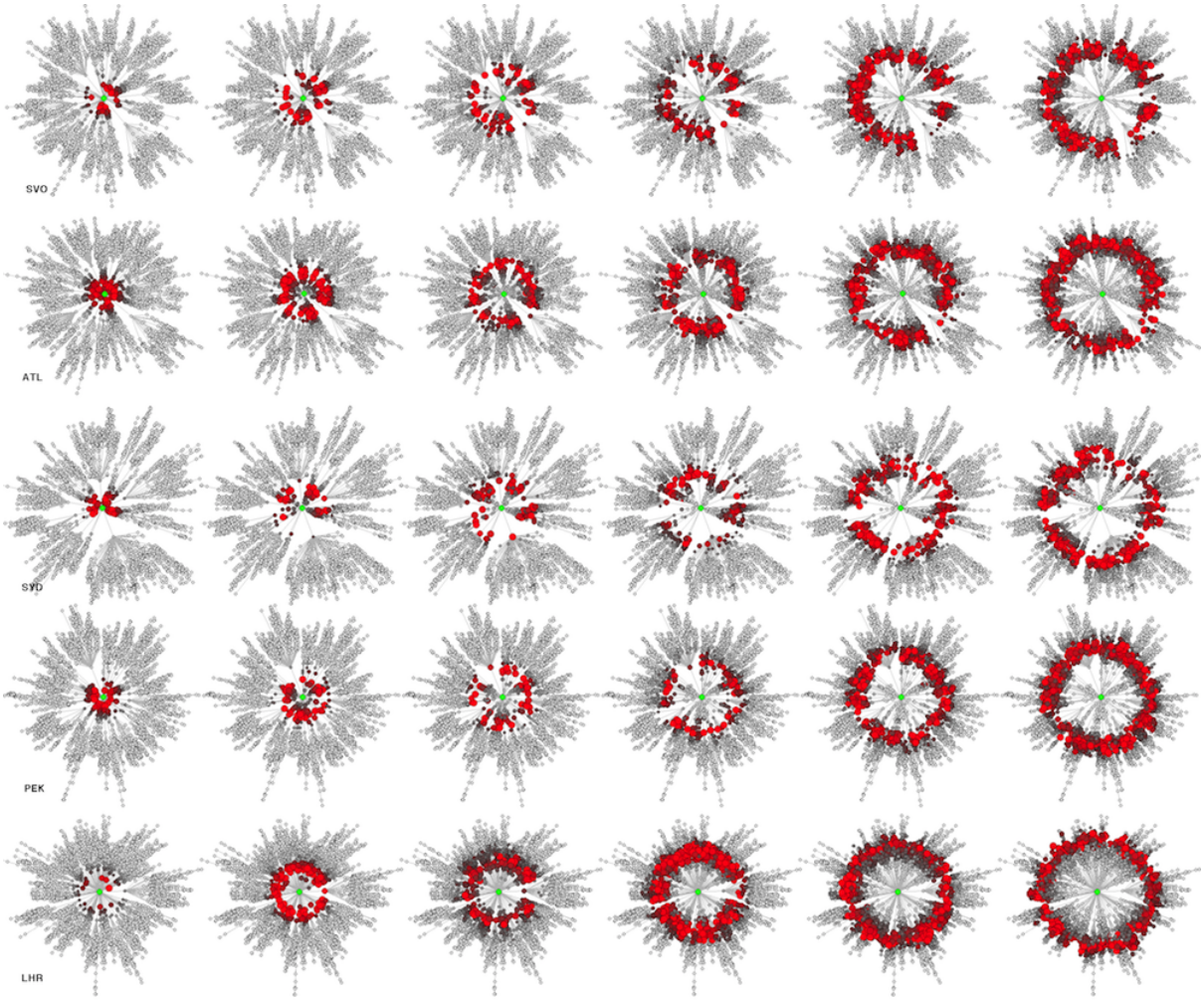

Figure S11: Global disease dynamics using effective distance and shortest path trees. For 5 regional hub OLs in the GMN (from top to bottom: Moscow, Atlanta, Sydney, Beijing, London) and times $T = 56$, 70, 84, 98, 112, and 126 (from left to right) the panels depict the nodes' disease prevalence (proportional to symbol size and redness) of a simulated epidemic. Parameters of the simulations are $R_0 = 1.5$, $\beta = 0.2857\,\mathrm{d}^{-1}$, $\gamma = 1.42 \times 10^{-3}\,\mathrm{d}^{-1}$ and $\varepsilon = 10^{-6}$).

Fig. S12.

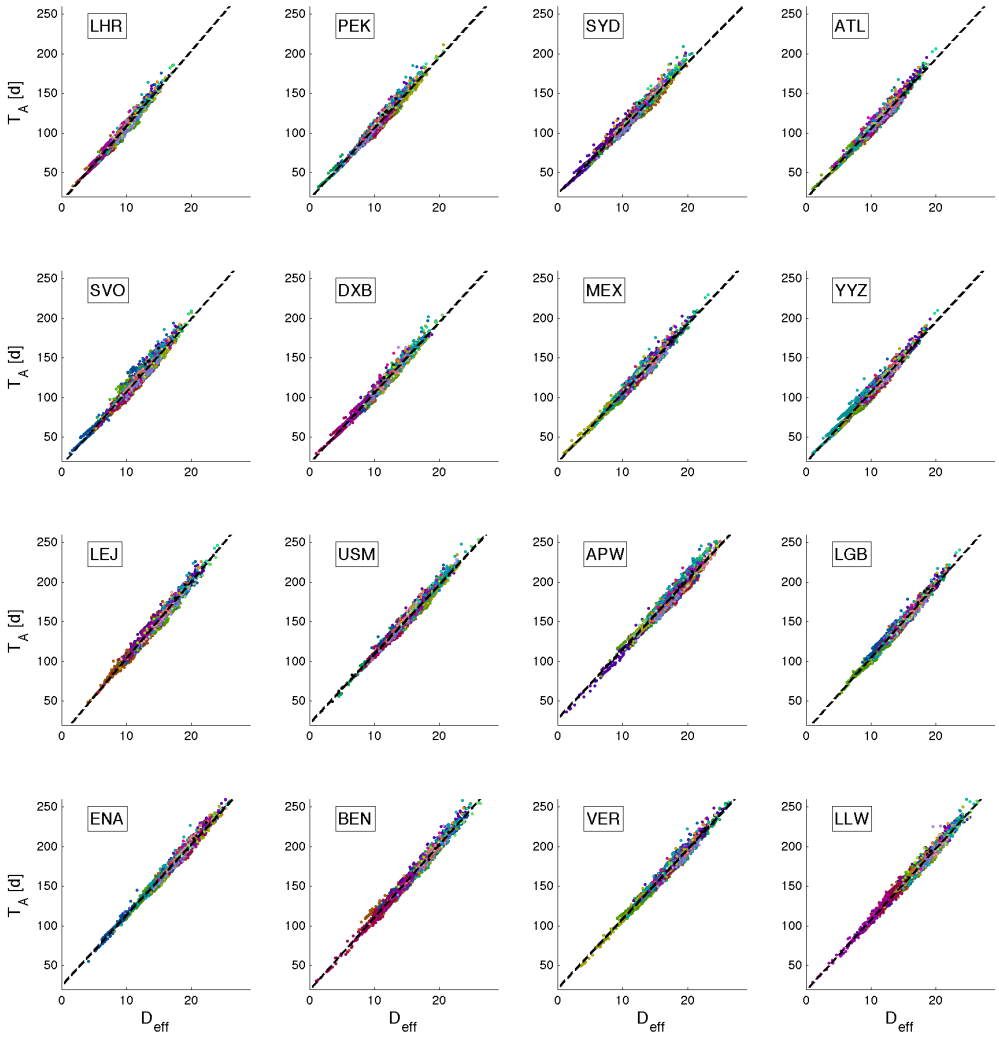

Figure S12: Correlation between arrival time and effective distance. For the same OLs and simulation parameters as in Fig. S7 each panel shows a scatter plot of epidemic arrival time vs. effective distance. Effective distance is a much better predictor of arrival times, see also Tab. S3.

Fig. S13.

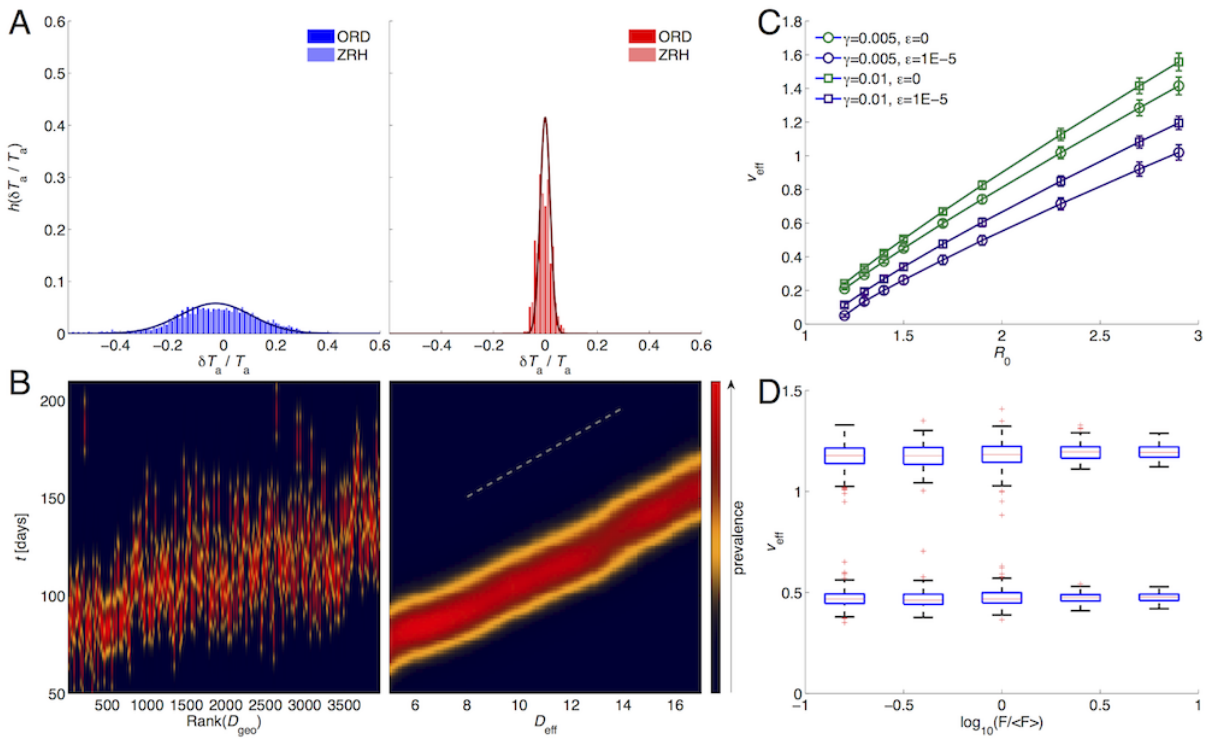

Figure S13: Effective wavefront propagation. (**A**) Distribution of relative residuals $r = \delta Ta/Ta$ for two sample OLs (Zurich and Chicago), using geographic distances (left) and effective distances (right). Solid lines indicate normal distributions for reference. The increase in fidelity of the linear scaling can be quantified by the ratio of the variance of the relative residuals, which is $\sigma^2(r_{\text{geo}})/\sigma^2(r_{\text{eff}}) = 60.42$ and 65.75 for ORD and ZRH, respectively. Using effective distances increases fidelity more than 50-fold. (**B**) Spatiotemporal dynamics in geographic (left) and effective (right) representations for a simulated pandemic with OL Zurich (parameters are $R_0 = 1.5$, $\gamma = 1.41 \times 10^{-3}\text{d}^{-1}$ and $\varepsilon = 10^{-5}$). In the conventional view, the abscissa represents the rank of geographic distance from the OL. Note the strong variability of prevalence profiles as a function of geographic distance rank. Using effective distances as the spatial coordinate, in contrast, results in a strikingly smooth propagating wave shape that is reminiscent of ordinary reaction diffusion dynamics. (**C**) Effective speed v_{eff} as a function of epidemic parameters R_0 , γ and ε. For each parameter combination and each of the 4069 possible OLs, the speed was computed and is shown as a function of the basic reproduction ratio R_0 for a fixed mobility rate and invasion threshold. Note that the error bars are surprisingly small, which implies that effective speeds, unlike conventionally estimated speeds, are independent of the OL. (**D**) The above is also supported by grouping OLs according to the nodes' traffic capacity F_n. The figure shows that the effective speeds of pandemics are roughly independent of airport capacity. The top row of whiskers corresponds to $R_0 = 2.9$, the bottom to $R_0 = 1.7$, and for both runs we set $\gamma = 2.8 \times 10^{-3}$ and $\varepsilon = 10^{-5}$.

Fig. S14.

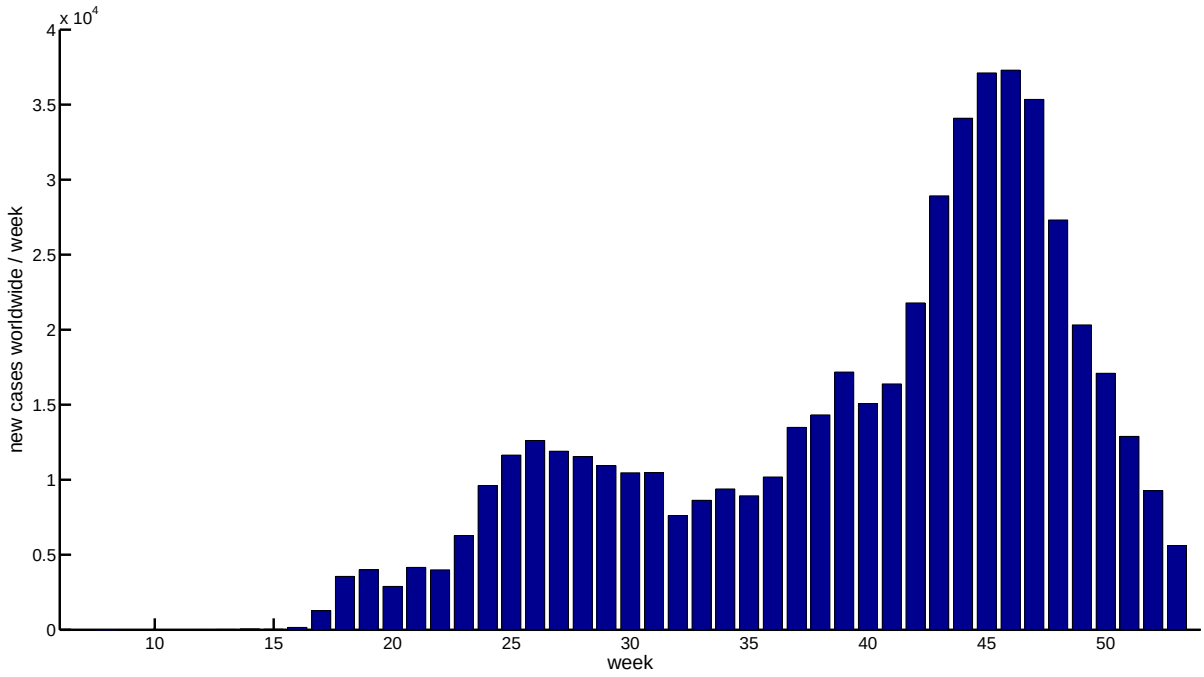

Figure S14: Global prevalence of the H1N1 influenza pandemic worldwide. Bars depict the number of new, confirmed cases in each week of 2009. In total, 103 countries were affected in this time period. Global prevalence was computed from data provided by FluNet [36].

Fig. S15.

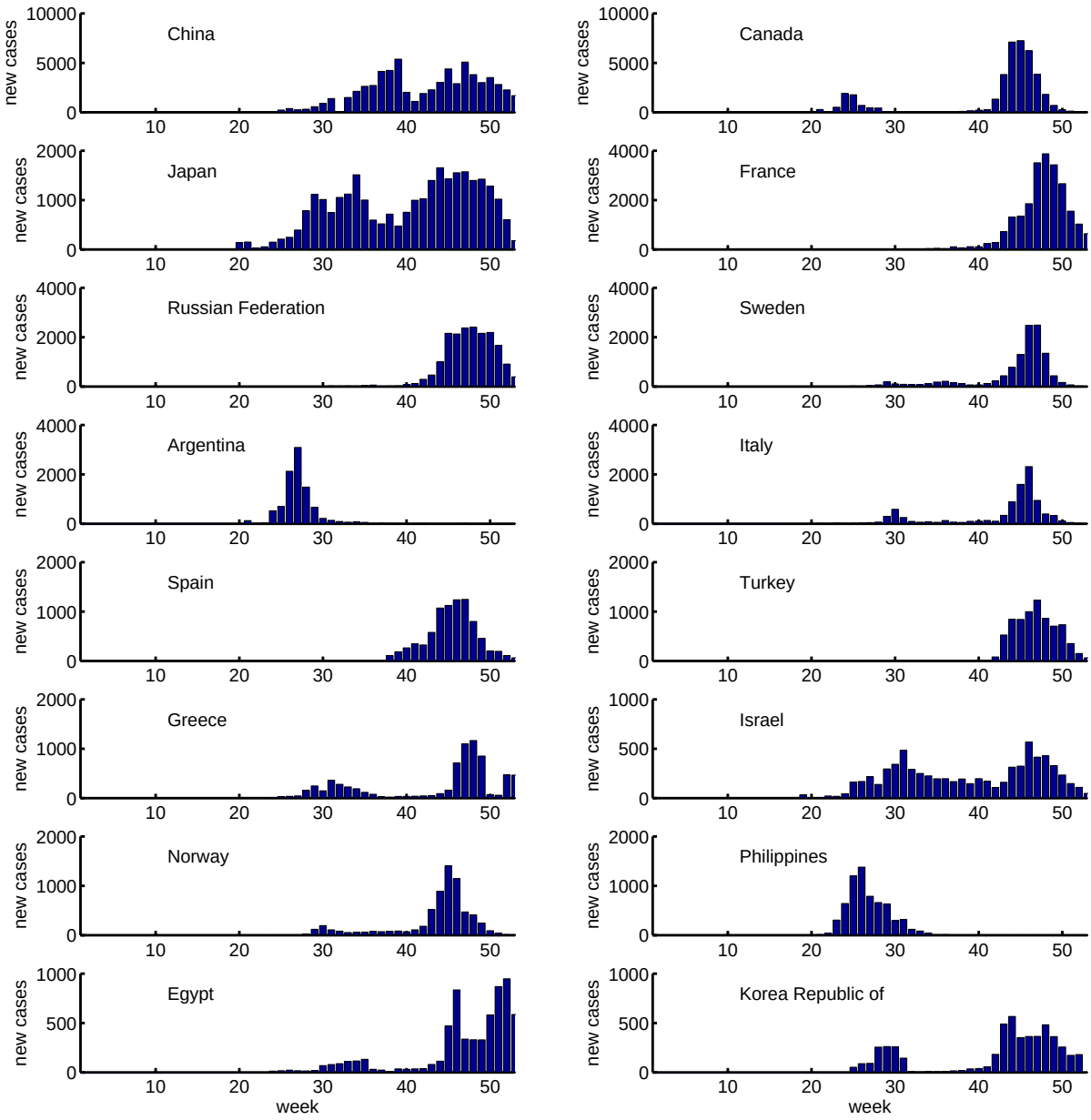

Figure S15: H1N1 prevalence curves for individual countries. Countries with highest total prevalence (except Mexico and the United States) are shown. Prevalence was computed from data provided by FluNet [36].

Fig. S16.

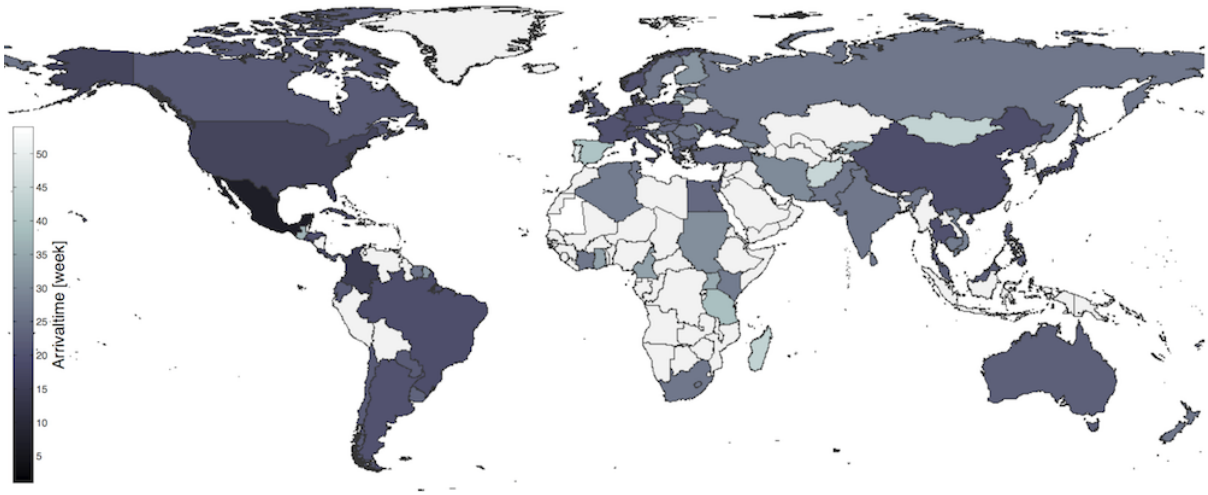

Figure S16: H1N1 arrival times. The map color codes the arrival time of the 2009 H1N1 pandemic in units "week of 2009". Arrival time is defined as the time of the first confirmed case of H1N1 in the country. The map depicts 252 countries, 189 are part of the largest connected component of GMNc.

Fig. S17.

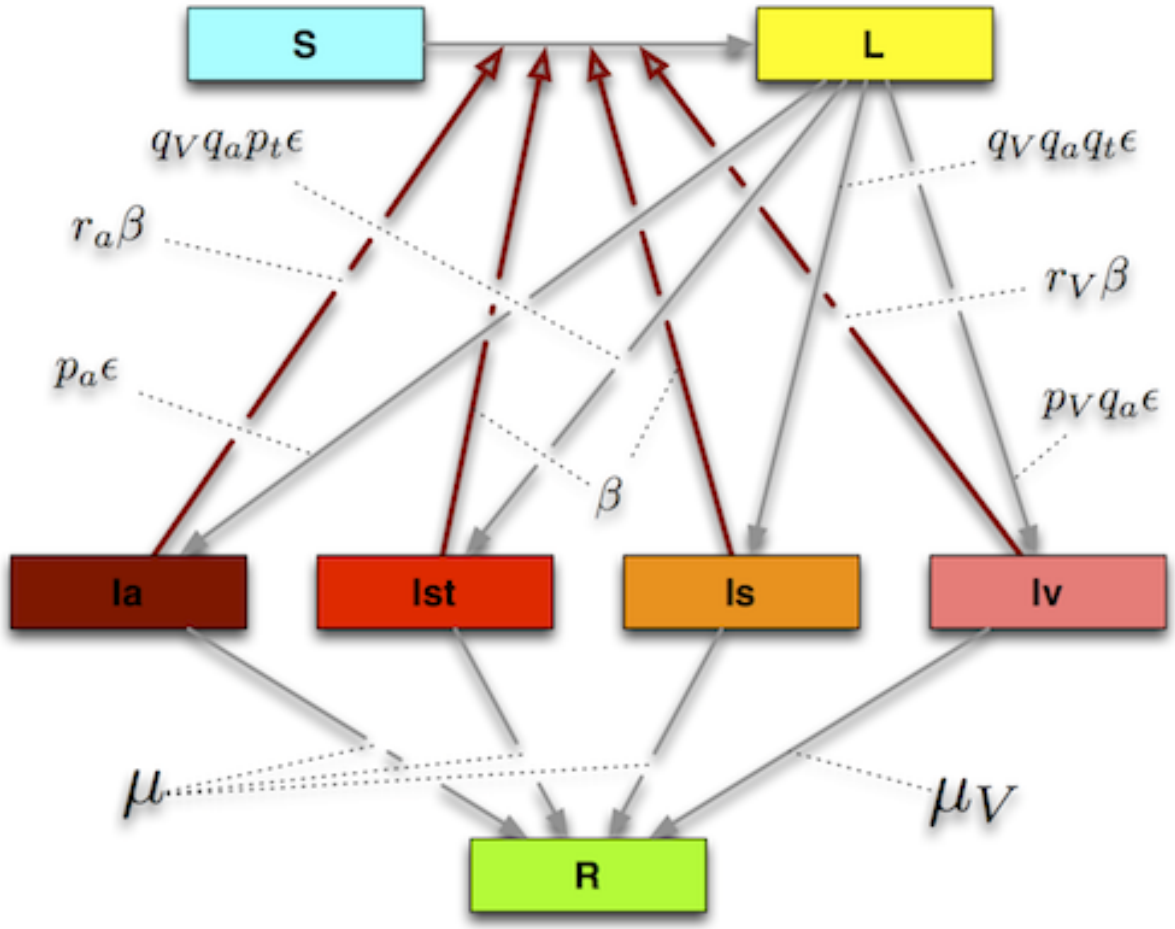

Figure S17: Compartmental structure of the modified Susceptible-Latent-Infected-Recovered model, adapted from Ref. [21]. Susceptibles enter the latent class by interacting with one of the infected subclasses. The base infection rate is β. This infection rate is modified by factors each of which depends on the class of infecteds a susceptible interacts with. Latent individuals enter one of the infected classes at different rates (base rate is ε), modified by parameters of the system. Altogether the dynamics is shaped by 10 parameters.

Fig. S18.

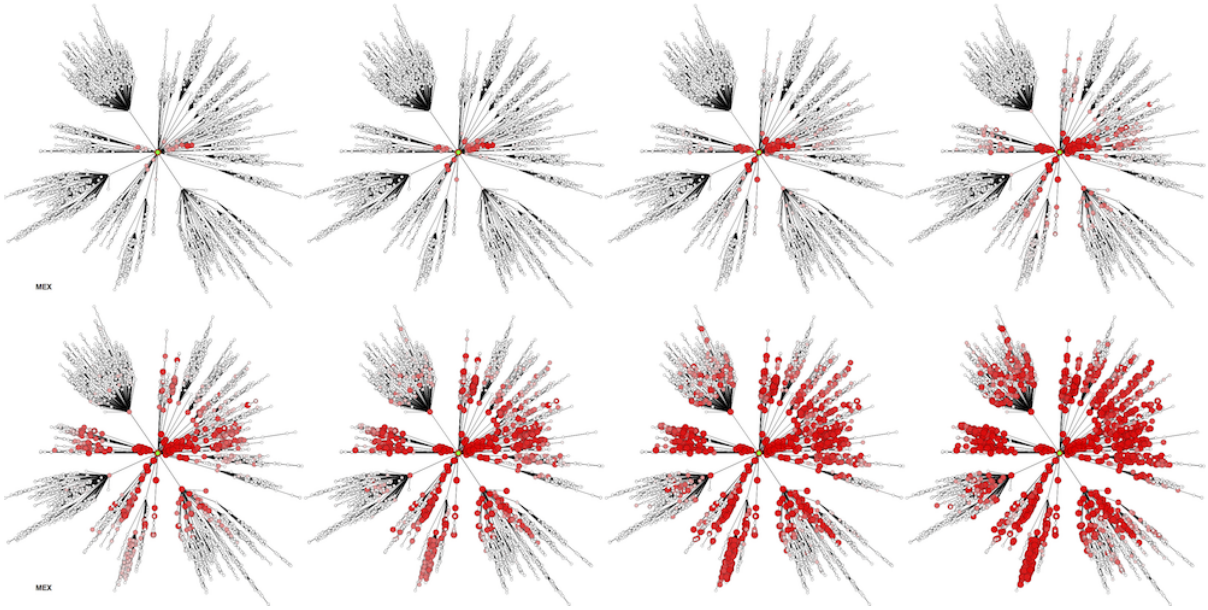

Figure S18: Epidemic Simulations using GLEAM. For the system described in Ref. [21] and outlined in Fig. S17 each panel from top left to bottom right depicts a temporal snapshot of the epidemic wave front. Color encodes the relative concentration of recovered individuals. The pattern has a pronounced circular shape, evidence that the decoupling of effective distance and effective speed is true also in complex, highly sophisticated stochastic simulations.

Fig. S19.

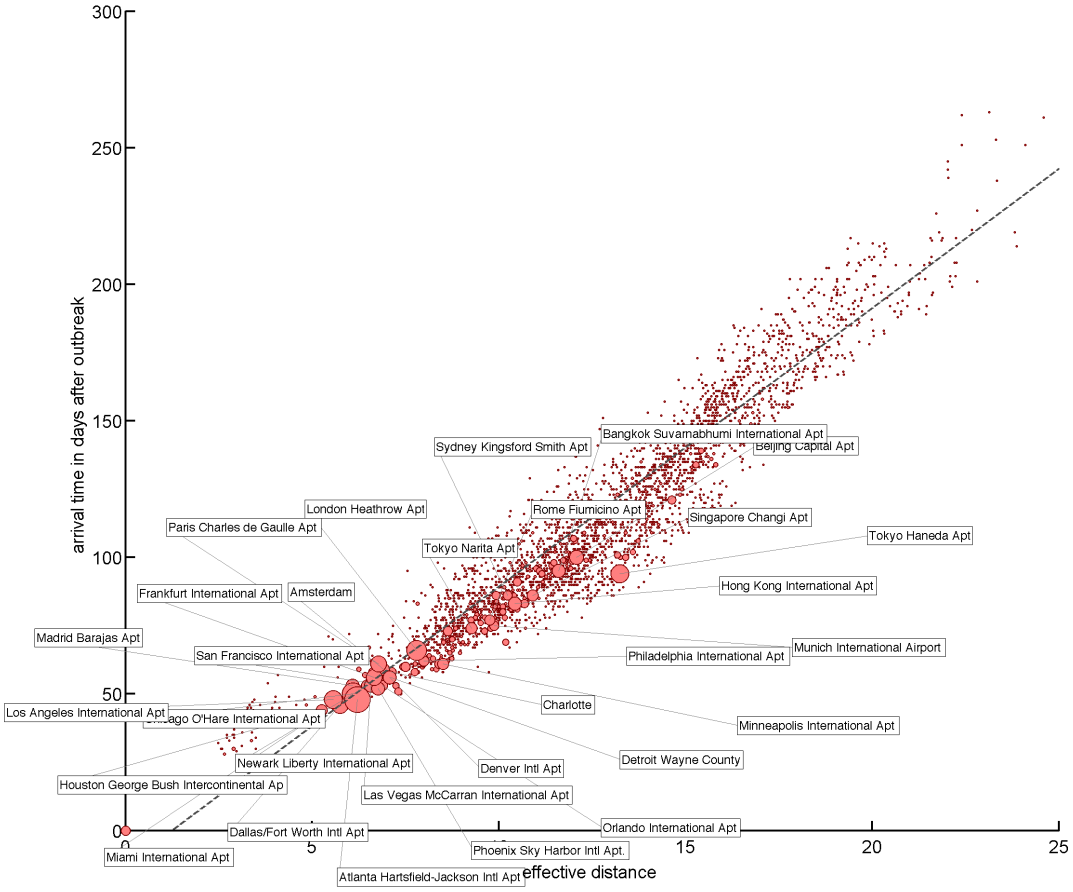

Figure S19: Arrival time vs. effective distance in GLEAM simulations. Based on 20 stochastic simulations of the model discussed in Ref. [21].

Fig. S20.

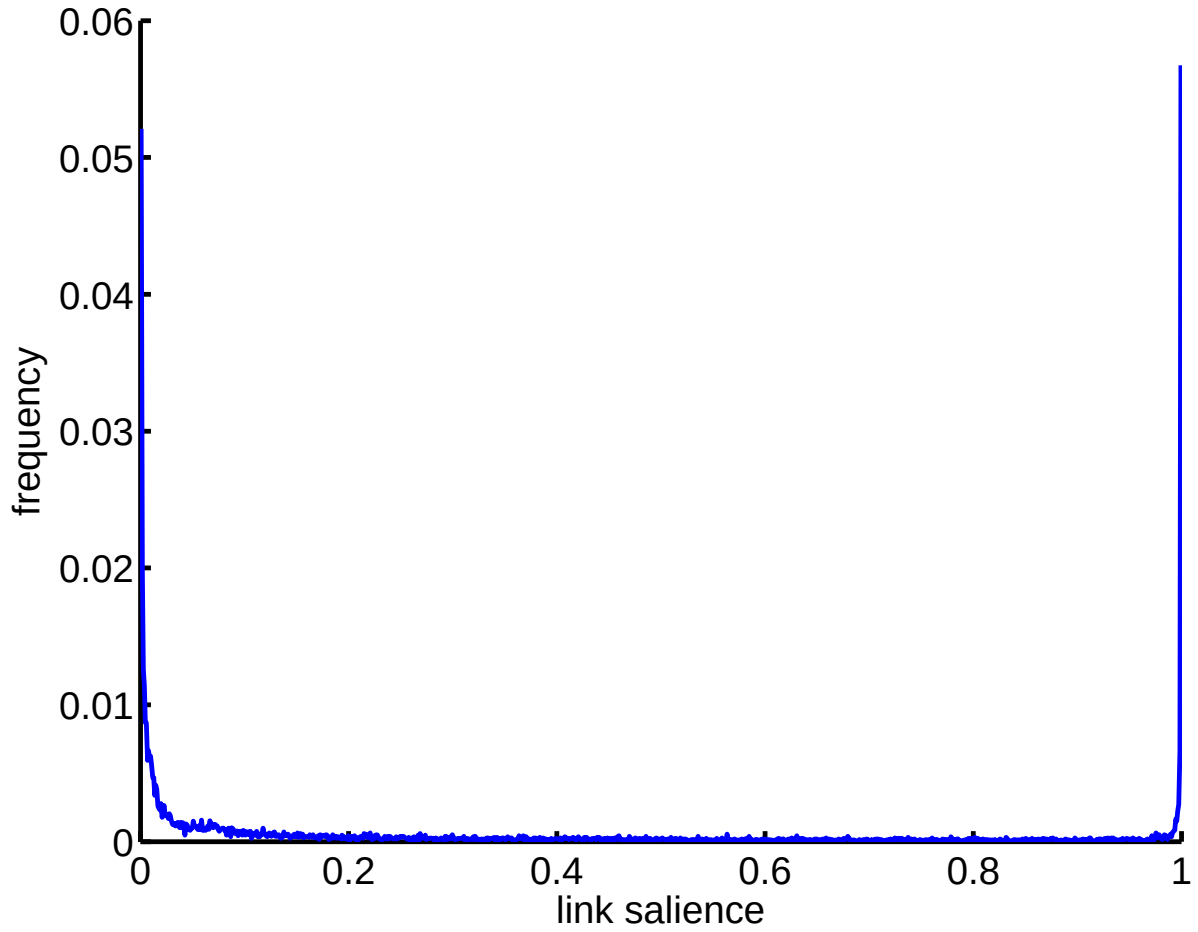

Figure S20: Salience distribution in the GMN. The salience s of a link is the fraction of shortest path trees the link is part of. Thus, it is a measure of how essential a link is to spreading processes that predominantly follow shortest paths. The salience distribution is bimodal in strongly heterogeneous networks. Thus, links either always participate in spreading processes or never. This is the reason why disease dynamics on strongly heterogeneous networks are more predictable than the redundancy of paths in such networks would suggest. The number of links that are almost certainly active in spreading processes are much fewer than the full links set of the network [29].

Tab. S1.

	N	L	$\rho[10^{-3}]$	$\Phi[\mathrm{d}^{-1}]$	$\langle F\rangle\ [\mathrm{d}^{-1}]$	$c_v(F)$	$\langle k\rangle$	$c_v(k)$	$\langle S\rangle\ [10^3\,\mathrm{d}^{-1}]$	$c_v(S)$
GMN	4069	25453	3.10	8.9×10^6	175	2.60	12.51	2.15	2.18	4.08
GMNc	189	5004	140.83	8.9×10^6	636	3.70	26.47	0.82	16.40	2.34

Table S1: Properties of the airport and country based global mobility networks GMN and GMNc, respectively. The columns from left to right are, number of nodes, links, link density $\rho = 2L/N^2$, total traffic flux in the network, mean flux per link, its coefficient of variation, mean node degree, its coefficient of variation, mean node strength in passengers per day and its coefficient of variation.

Code	Airport	Lon.	Lat.	$F\,[10^5/\mathrm{d}]$	R^2_{geo}	v_{geo} [km/d]
LHR	London	$-0.461°$	$51.477°$	1.2574	0.281	323.02
PEK	Beijing	$116.583°$	$40.067°$	0.8284	0.283	279.68
SYD	Sydney	$151.167°$	$-33.933°$	0.5405	0.413	244.31
ATL	Atlanta	$-84.418°$	$33.652°$	1.5817	0.466	241.90
SVO	Moscow	$37.417°$	$55.969°$	0.2595	0.402	200.65
DXB	Dubai	$55.350°$	$25.250°$	0.4845	0.448	201.95
MEX	Mexico City	$-99.067°$	$19.433°$	0.5058	0.352	272.14
YYZ	Toronto	$-79.633°$	$43.686°$	0.5329	0.428	266.14
LEJ	Leipzig/Halle	$12.233°$	$51.417°$	0.0316	0.446	206.49
USM	Koh Samui	$100.067°$	$9.550°$	0.0228	0.445	253.11
APW	Apia Faleolo	$-172.000°$	$-13.817°$	0.0059	0.339	243.78
LGB	Long Beach	$-118.150°$	$33.817°$	0.0521	0.306	279.36
ENA	Kenai	$-151.267°$	$60.550°$	0.0042	0.468	161.61
BEN	Benghazi	$20.267°$	$32.100°$	0.0154	0.509	176.19
VER	Veracruz	$-96.183°$	$19.133°$	0.0166	0.393	253.56
LLW	Lilongwe	$33.783°$	$-13.783°$	0.0068	0.553	165.37

Table S2: Properties of sample OLs. Each row contains information on a subset of OLs shown in Fig. S3 and corresponds to one of the panels in Fig.S7. The estimated speed of expansion from the correlograms v_{geo}. The degree of correlation of arrival time and geographic distance is quantified by the correlation coefficient R^2_{geo}.

Tab. S2.

Tab. S3.

Code	Airport	R^2_{geo}	R^2_{eff}	σ^2_{geo}	σ^2_{eff}	$\sigma^2_{\text{geo}}/\sigma^2_{\text{geo}}$
LHR	London	0.281	0.976	0.0550	0.0015	37.3759
PEK	Beijing	0.283	0.976	0.0693	0.0013	54.0128
SYD	Sydney	0.413	0.975	0.0509	0.0014	36.0350
ATL	Atlanta	0.466	0.978	0.0596	0.0015	40.9225
SVO	Moscow	0.402	0.969	0.0606	0.0018	33.9176
DXB	Dubai	0.448	0.983	0.0509	0.0013	39.9663
MEX	Mexico City	0.352	0.980	0.0695	0.0011	63.3636
YYZ	Toronto	0.428	0.973	0.0480	0.0017	27.8716
LEJ	Leipzig/Halle	0.446	0.978	0.0456	0.0009	51.2036
USM	Koh Samui	0.445	0.978	0.0396	0.0007	53.8615
APW	Apia Faleolo	0.339	0.961	0.0555	0.0010	54.0747
LGB	Long Beach	0.306	0.979	0.0430	0.0008	55.2007
ENA	Kenai	0.468	0.985	0.0457	0.0005	88.0279
BEN	Benghazi	0.509	0.983	0.0548	0.0007	74.8667
VER	Veracruz	0.393	0.981	0.0498	0.0006	83.3615
LLW	Lilongwe	0.553	0.979	0.0395	0.0007	55.2757

Table S3: Correlation between arrival time and effective distance. Each row corresponds to one of the OLs also listed in Tab. S2. Using the data depicted in Fig. S12 as a basis we computed the correlation coefficient of arrival time with effective distance R^2_{eff} and compare it to the correlation coefficient of arrival time with geographic distance R^2_{geo}. Effective distance correlates substantially better with arrival times for all OLs, supplementing the results for the two sample OLs discussed in the main text. In order to quantify the fluctuations around the linear regression we computed the variance of the residuals using either effective or geographic distance, described in detail in the caption of Fig. 2. These variances are denoted by σ^2_{eff} and σ^2_{geo}, respectively. Their ratio (last column) quantifies the increase in predictability when using effective instead of geographic distance.

Tab. S4.

Country	T_a	**Country**	T_a	**Country**	T_a
Mexico	6	Czech Republic	21	Tunisia	25
Colombia	14	Malaysia	21	Algeria	26
USA	15	Panama	21	Cambodia	26
Costa Rica	16	Bulgaria	22	Kenya	26
El Salvador	17	Estonia	22	Luxembourg	26
Brazil	18	Greece	22	New Caledonia	26
China	18	Slovakia	22	Macedonia Former Yugoslav Republic of	26
Cuba	18	Ukraine	22	Croatia	27
Denmark	18	Uruguay	22	Guadeloupe	27
Germany	18	Egypt	23	Viet Nam	27
Ireland Republic of	18	Hungary	23	Iran Islamic Republic of	28
Italy	18	Martinique	23	Latvia	28
Netherlands	18	Turkey	23	Georgia	29
Singapore	18	Fiji	24	Sudan	29
Switzerland	18	Iraq	24	Finland	30
Argentina	19	Morocco	24	French Guiana	31
France	19	Suriname	24	Cameroon	32
Honduras	19	Belgium	25	Ghana	32
Israel	19	Cape Verde	25	Lithuania	32
Japan	19	Cote d'Ivoire	25	Kyrgyzstan	34
Norway	19	India	25	Uganda	34
Poland	19	New Zealand	25	Guatemala	36
Thailand	19	Pakistan	25	Malta	37
Canada	20	Korea Republic of	25	Moldova Republic of	37
Chile	20	Romania	25	Tanzania United Republic of	37
Ecuador	20	Russian Federation	25	Spain	38
Jamaica	20	Serbia	25	Madagascar	41
Paraguay	20	Slovenia	25	Mongolia	41
Philippines	20	South Africa	25	Afghanistan	42
United Kingdom	20	Sri Lanka	25	Portugal	47
Australia	21	Sweden	25		

Table S4: H1N1 arrival time by country. The table lists the arrival time (first confirmed case) in 83 affected countries in the first 50 weeks of 2009. Arrival times are given as week of 2009. See also Fig. S16 for a geographical representation of arrival times.

Movie 1.

Effective vs. geographic distance. The clip depicts the time course of a simulated pandemic with initial outbreak in Atlanta, USA. The panels on the left and right depict the same simulation, only in different representations. On the left one observes a concentric, expanding wave front in the effective distance representation. The same simulation exhibits more complex spatio-temporal structure in the conventional geographic representation on the right.

Movie 2.

Effective vs. geographic distance. The clip depicts the time course of a simulated pandemic with initial outbreak in Mexico City. The panels on the left and right depict the same simulation, only in different representations. On the left one observes a concentric, expanding wave front in the effective distance representation. The same simulation exhibits more complex spatio-temporal structure in the conventional geographic representation on the right.

Movie 3.

Effective vs. geographic distance. The clip depicts the time course of a simulated pandemic with initial outbreak in Paphos, Greece. The panels on the left and right depict the same simulation, only in different representations. On the left one observes a concentric, expanding wave front in the effective distance representation. The same simulation exhibits more complex spatio-temporal structure in the conventional geographic representation on the right.

References and Notes

1. R. M. Anderson, R. M. May, *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford and New York, 1991).

2. World Health Organization, Global Alert and Response; www.who.int/csr/en/.

3. A. R. McLean, R. M. May, J. Pattison, R. A. Weiss, *SARS: A Case Study in Emerging Infections* (Oxford Univ. Press, Oxford, 2005).

4. C. Fraser, C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, T. Jombart, W. R. Hinsley, N. C. Grassly, F. Balloux, A. C. Ghani, N. M. Ferguson, A. Rambaut, O. G. Pybus, H. Lopez-Gatell, C. M. Alpuche-Aranda, I. B. Chapela, E. P. Zavala, D. M. Guevara, F. Checchi, E. Garcia, S. Hugonnet, C. Roth, WHO Rapid Pandemic Assessment Collaboration, Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science* **324**, 1557–1561 (2009). Medline doi:10.1126/science.1176062

5. L. Hufnagel, D. Brockmann, T. Geisel, Forecast and control of epidemics in a globalized world. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 15124–15129 (2004). Medline doi:10.1073/pnas.0308344101

6. D. Brockmann, L. Hufnagel, T. Geisel, The scaling laws of human travel. *Nature* **439**, 462–465 (2006). Medline doi:10.1038/nature04292

7. M. Moore, P. Gould, B. S. Keary, Global urbanization and impact on health. *Int. J. Hyg. Environ. Health* **206**, 269–278 (2003). Medline doi:10.1078/1438-4639-00223

8. A. Vespignani, Predicting the behavior of techno-social systems. *Science* **325**, 425–428 (2009). Medline doi:10.1126/science.1171990

9. V. Colizza, A. Barrat, M. Barthélemy, A. Vespignani, The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2015–2020 (2006). Medline doi:10.1073/pnas.0510525103

10. B. S. Cooper, R. J. Pitman, W. J. Edmunds, N. J. Gay, Delaying the International Spread of Pandemic Influenza. *PLOS Med.* **3**, e212 (2006). doi:10.1371/journal.pmed.0030212

11. T. D. Hollingsworth, N. M. Ferguson, R. M. Anderson, Frequent travelers and rate of spread of epidemics. *Emerg. Infect. Dis.* **13**, 1288–1294 (2007). Medline doi:10.3201/eid1309.070081

12. J. M. Epstein, D. M. Goedecke, F. Yu, R. J. Morris, D. K. Wagener, G. V. Bobashev, Controlling pandemic flu: The value of international air travel restrictions. *PLOS ONE* **2**, e401 (2007). Medline doi:10.1371/journal.pone.0000401

13. V. Colizza, A. Barrat, M. Barthelemy, A.-J. Valleron, A. Vespignani, Modeling the Worldwide Spread of Pandemic Influenza: Baseline Case and Containment Interventions. *PLOS Med.* **e13**, 95 (2007). doi:10.1371/journal.pmed.0040013

14. R. Fisher, The wave of advance of advantageous genes. *Ann. Eugen.* **7**, 355–369 (1937). doi:10.1111/j.1469-1809.1937.tb02153.x

15. J. V. Noble, Geographic and temporal development of plagues. *Nature* **250**, 726–729 (1974). Medline doi:10.1038/250726a0

16. J. D. Murray, *Mathematical Biology* (Springer, Berlin, 2005).

17. D. Brockmann, T. Geisel, Lévy flights in inhomogeneous media. *Phys. Rev. Lett.* **90**, 170601 (2003). Medline doi:10.1103/PhysRevLett.90.170601

18. D. Brockmann, L. Hufnagel, Front propagation in reaction-superdiffusion dynamics: Taming Lévy flights with fluctuations. *Phys. Rev. Lett.* **98**, 178301 (2007). doi:10.1103/PhysRevLett.98.178301

19. D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, A. Vespignani, Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21484–21489 (2009). Medline doi:10.1073/pnas.0906910106

20. V. Colizza, R. Pastor-Satorras, A. Vespignani, Reaction–diffusion processes and metapopulation models in heterogeneous networks. *Nat. Phys.* **3**, 276–282 (2007). doi:10.1038/nphys560

21. W. Van den Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, A. Vespignani, The GLEaMviz computational tool, a publicly available software to explore realistic epidemic spreading scenarios at the global scale. *BMC Infect. Dis.* **11**, 37 (2011). Medline doi:10.1186/1471-2334-11-37

22. D. Balcan, B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, A. Vespignani, Modeling the spatial spread of infectious diseases: The GLobal Epidemic and Mobility computational model. *J. Comput. Sci* **1**, 132–145 (2010). Medline doi:10.1016/j.jocs.2010.07.002

23. N. M. Ferguson, D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, D. S. Burke, Strategies for mitigating an influenza pandemic. *Nature* **442**, 448–452 (2006). Medline doi:10.1038/nature04795

24. L. A. Rvachev, I. M. Longini Jr., *Math. Biosci.* **75**, 3 (1985). doi:10.1016/0025-5564(85)90064-1

25. S. Eubank, H. Guclu, V. S. Anil Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, N. Wang, Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180–184 (2004). Medline doi:10.1038/nature02541

26. M. Tizzoni, P. Bajardi, C. Poletto, J. J. Ramasco, D. Balcan, B. Gonçalves, N. Perra, V. Colizza, A. Vespignani, Real-time numerical forecast of global epidemic spreading: Case study of 2009 A/H1N1pdm. *BMC Med.* **10**, 165 (2012). Medline doi:10.1186/1741-7015-10-165

27. M. Ajelli, B. Gonçalves, D. Balcan, V. Colizza, H. Hu, J. J. Ramasco, S. Merler, A. Vespignani, Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. *BMC Infect. Dis.* **10**, 190 (2010). Medline doi:10.1186/1471-2334-10-190

28. R. M. May, Uses and abuses of mathematics in biology. *Science* **303**, 790–793 (2004). Medline doi:10.1126/science.1094442

29. D. Grady, C. Thiemann, D. Brockmann, Robust classification of salient links in complex networks. *Nat. Commun.* **3**, 864 (2012). doi:10.1038/ncomms1847

30. V. Colizza, A. Vespignani, Invasion threshold in heterogeneous metapopulation networks. *Phys. Rev. Lett.* **99**, 148701 (2007). Medline doi:10.1103/PhysRevLett.99.148701

31. V. Belik, T. Geisel, D. Brockmann, Natural Human Mobility Patterns and Spatial Spread of Infectious Diseases. *Physical Review X* **1**, 011001 (2011). doi:10.1103/PhysRevX.1.011001

32. E. Brunet, B. Derrida, Shift in the velocity of a front due to a cutoff. *Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics* **56**, 2597–2604 (1997). doi:10.1103/PhysRevE.56.2597

33. A. Y. Lokhov, M. Mézard, H. Ohta, L. Zdeborová, Inferring the origin of an epidemy with dynamic message-passing algorithm; http://arxiv.org/abs/1303.5315 (2013).

34. P. C. Pinto, P. Thiran, M. Vetterli, Locating the source of diffusion in large-scale networks. *Phys. Rev. Lett.* **109**, 068702 (2012). Medline doi:10.1103/PhysRevLett.109.068702

35. E. J. Abbott, F. A. Firestone, *Mech. Eng.* **55**, 569–572 (1933).

36. World Heath Organization, FluNet (2013).

37. J. Anderson, *Am. Econ. Rev.* **69**, 106–116 (1979).

38. Robert-Koch-Institute, Survstat (2012).

39. O. Woolley-Meza, C. Thiemann, D. Grady, J. J. Lee, H. Seebens, B. Blasius, D. Brockmann, Complexity in human transportation networks: A comparative analysis of worldwide air transportation and global cargo-ship movements. *Eur. Phys. J. B* **84**, 589–600 (2011). doi:10.1140/epjb/e2011-20208-9

40. A. Barrat, M. Barthélemy, R. Pastor-Satorras, A. Vespignani, The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 3747–3752 (2004). Medline doi:10.1073/pnas.0400087101

41. A. Barrat, M. Barthelemy, A. Vespignani, The effects of spatial constraints on the evolution of weighted complex networks. *J. Stat. Mech.* **2005**, P05003 (2005). doi:10.1088/1742-5468/2005/05/P05003

42. Rafael Brune, Christian Thiemann, and Dirk Brockmann. Predicting the origin of contagion pro- cesses on complex, multi-scale networks; *http://meetings.aps.org/link/BAPS.2012.MAR.H54.5* (2012).

43. G. Caldarelli, *Scale-Free Networks: Complex Webs in Nature and Technology* (Oxford Univ. Press, 2007).

44. L. Dall'Asta, A. Barrat, M. Barthelemy, A. Vespignani, Vulnerability of weighted networks. *J. Stat. Mech.* **2006**, P04006 (2006). doi:10.1088/1742-5468/2006/04/P04006

45. OAG Worldwide Ltd., 2007.