Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

LETTER

Arrival time statistics in global disease spread

Aurélien Gautreau[1], **Alain Barrat**[1,2] **and Marc Barthélemy**[3]

[1] LPT, CNRS, UMR 8627, and Université Paris-Sud, Orsay, F-91405, France
[2] Complex Networks Lagrange Laboratory, ISI Foundation, Turin, Italy
[3] CEA-Centre d'Etudes de Bruyères-le-Châtel, Département de Physique Théorique et Appliquée, BP 12, F-91680 Bruyères-Le-Châtel, France
E-mail: aurelien.gautreau@th.u-psud.fr, Alain.Barrat@th.u-psud.fr and marc.barthelemy@gmail.com

Received 20 July 2007
Accepted 3 September 2007
Published 19 September 2007

Online at stacks.iop.org/JSTAT/2007/L09001
doi:10.1088/1742-5468/2007/09/L09001

Abstract. Metapopulation models describing cities with different populations coupled by the travel of individuals are of great importance in the understanding of disease spread on a large scale. An important example is the Rvachev–Longini model which is widely used in computational epidemiology. Few analytical results are, however, available and, in particular, little is known about paths followed by epidemics and disease arrival times. We study the arrival time of a disease in a city as a function of the starting seed of the epidemics. We propose an analytical ansatz, test it in the case of a spread on the worldwide air-transportation network, and show that it predicts accurately the arrival order of a disease in worldwide cities.

Keywords: random graphs, networks, interacting agent models, stochastic processes, new applications of statistical mechanics

J. Stat. Mech. (2007) L09001

©2007 IOP Publishing Ltd and SISSA

In modern societies, individuals can easily travel over a wide range of spatial and temporal scales. The interconnections of areas and populations through various means of transport have important effects on the geographical spread of epidemics. In particular, the structure and the different complexity levels of the air-transportation network are responsible for the heterogeneous and seemingly erratic outbreak patterns observed in the worldwide propagation of diseases [1] as recently documented for SARS [2, 3]. In order to describe such a complex phenomenon and to obtain powerful numerical forecasting tools, different levels of description are possible, ranging from a simple global mean-field to detailed agent-based simulations [4]–[10] that recreate entire populations and their dynamics at the scale of the single individual [10].

At large scale, such as the worldwide level, a very important class of models in modern epidemiology are the so-called metapopulation models which use a description at two levels by dividing the global population into interconnected subpopulations. Within each subpopulation, a mean-field-like model of epidemic spreading is used, while the spread from one subpopulation to another is due to the travel of individuals. Agents of each subpopulation can be in various states (healthy, infectious, recovered, etc), change state by contact with other agents and diffuse on the transportation network between subpopulations. Metapopulation models can thus be considered as reaction–diffusion processes, which opens up very interesting perspectives and issues [11] within the global framework of dynamical phenomena occurring on complex networks [12]–[15]. For the description of worldwide epidemic spreading, the subpopulations are cities connected by a transportation network in which links correspond to the existence of passenger flows described by the worldwide air-transportation network (WAN). The WAN represents a major channel for the worldwide spread of infectious diseases [3, 1] and its complex, heterogeneous features at various levels (degree distribution, traffic, populations) have recently been characterized [17, 16].

In this letter, we focus, in the framework of such metapopulation models, on the issue of the arrival time in a city of the first infectious individual. In particular, we study how this time depends on the origin of the disease and on the network characteristics. This problem is more complex than the one of random walks on complex networks [18], since the number of infectious individuals diffusing on the network is constantly evolving due to the inner-city epidemic dynamics. We also note that references [19, 20] were also concerned with the arrival time problem for an epidemic spreading on a complex network, but in a different framework: each network node was an individual (susceptible or infectious), while in our case each node represents a whole subpopulation. After the precise definition of the model, we will first consider the simple case of a one-dimensional topology for the transportation network in order to gain analytical insights into this problem. This will allow us to propose an analytical form for the arrival time in arbitrary networks. We then test this form in the case of the WAN by simulating numerically a stochastic spreading phenomenon on the network, and show that we can indeed predict with good accuracy the spreading phenomenon and the arrival order of a disease in various cities at a worldwide level.

While the precise model describing the epidemic spreading at the subpopulation level could be refined at will in order to describe a particular disease, we are here interested in generic and fundamental aspects of the metapopulation modeling approach. We therefore restrict our study to a simple SI disease model in which individuals are either healthy

J. Stat. Mech. (2007) L09001

(susceptible, S) or can become infectious (I) if in contact with an infectious individual. The Rvachev–Longini SI model [21] describes the evolution of the number of infectious $I_i(t)$ individuals (and also of $S_i(t)$) in each city i through

$$\partial_t I_i = K(\{X_i\}) + \Omega(\{I_j\}), \tag{1}$$

where the first term K on the right-hand side describes the (epidemic) reaction process inside each subpopulation (city), due to the interaction of individuals in the various possible states. In our case $X \in \{S, I\}$ (we have checked that more involved models, such as SIS or SIR, give consistent results [22]) and the standard homogeneous mixing assumption in each city gives [4]: $K(\{X_i\}) = \lambda I_i(N_i - I_i)/N_i$, where N_i is the population of city i and λ the spreading rate. The second term Ω represents the evolution due to the arrival or departure of infectious individuals from or to other cities and is determined by passenger flows on the transportation network. This model therefore considers a simplified mechanistic approach with a widely used Markovian assumption in which individuals are not labeled according to their original subpopulation, and at each time step the same traveling probability applies to all individuals in the subpopulation, without memory of their origin [21, 3, 1]. Denoting by w_{ij} the average number of passengers traveling from i to j per unit of time ($w_{ij} = 0$ if there is no direct connection), the probability per unit time that an individual travels from city i to city j is then given by w_{ij}/N_i. The full metapopulation model is therefore described by

$$\partial_t I_i = \lambda I_i(t)\frac{N_i - I_i(t)}{N_i} + \sum_j \frac{w_{ji}}{N_j} I_j - \sum_j \frac{w_{ij}}{N_i} I_i. \tag{2}$$

This original formulation considers only expectation values, which can take continuous values, so that 'fractions' of infectious individuals can travel and infect neighboring cities arbitrarily fast[4]. To investigate arrival times, one therefore needs to take into account the inherent stochasticity of the spreading. We thus consider in all our numerical simulations the stochastic generalization described in [1, 3] where the number of individuals traveling on each connection is an integer variable randomly extracted at each time step of length Δt, with average $\Delta t w_{ij} I_i/N_i$ (in the numerical simulations we will use $\Delta t = 1$day); for simplicity we keep the endogenous growth deterministic since we are mainly concerned with the effect of travel, but we have checked that inclusion of stochastic effects as in [1] do not change our results [22]. Note that, in real cases such as the WAN, most weights are symmetric ($w_{ij} = w_{ji}$) [16] but the probabilities of travel from one city to another are not since they depend on the populations of the various cities: the travel effectively occurs as a random diffusion with non-symmetric rates on the transportation network. The topological distance thus does not contain all the information needed to characterize such a process, nor does *a priori* the optimal weighted distance, which takes into account the weights [23] but not the populations nor the endogenous epidemic evolution. Moreover, since most transportation networks are small-world networks, many cities lie at the same topological distance from a given seed, but will potentially be reached at very different times.

Before turning to numerical simulations of the described model, we present an analytical approach to the determination of arrival times. Let us first consider the simple

[4] A numerical integration of (2) leads to an infection of all cities at time 0^+.

J. Stat. Mech. (2007) L09001

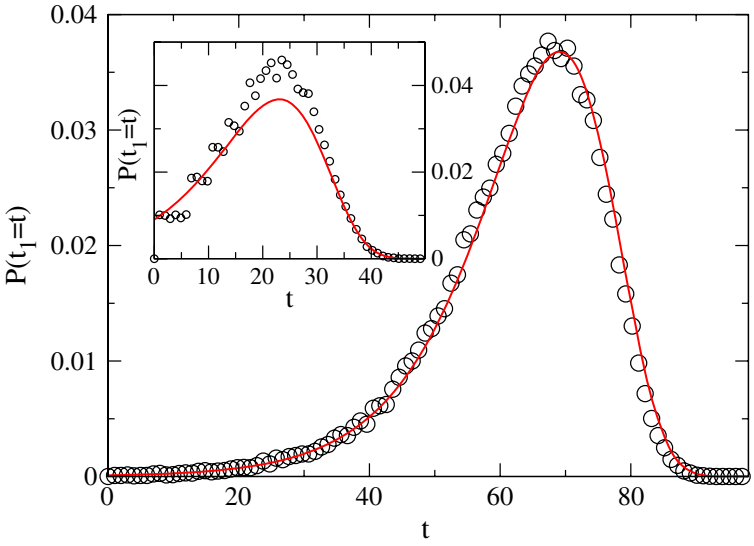

Figure 1. Two-cities model. Arrival time t_1 distribution computed by numerical simulation and compared with the result of equation (4) for $w/N_0\lambda = 10^{-2}$ (line). Inset: the same for $w/N_0\lambda = 10^{-1}$.

case of two cities (0 and 1), with populations N_0, N_1 which are connected by a passenger flux $w_{01} = w$. We assume that, at $t = 0$, there are $I^0 = 1$ infectious people in the city 0. Let us first consider that the travel events occur as instantaneous jumps of probability $p = (w/N_0)\Delta t$, at discretized times, in units of Δt. The probability that the time of arrival t_1 of the epidemic in the city 1 is equal to $t = n\Delta t$ is then

$$P_d(t_1 = n\Delta t) = [1 - (1-p)^{I^0(n\Delta t)}] \prod_{i=1}^{n-1} (1-p)^{I^0(\mathrm{i}\Delta t)}. \tag{3}$$

In order to obtain the density probability $P(t)$ of the arrival time in city 1 we consider the limit $\Delta t \to 0$, using the following assumptions: (i) $I^0(t) \ll N$, which is realistic for the usual diseases, in which only small fractions of the population are infectious; (ii) at t_1 the number of infected in city 0 is large enough so that the continuous limit for $I^0(t)$ can be used. This last assumption is satisfied if $1/\lambda \ll \langle t_1 \rangle$. Within these assumptions, we obtain

$$P(t)\,\mathrm{d}t = \frac{w}{N_0} \exp\left(\lambda t - \frac{w}{N_0\lambda} \mathrm{e}^{\lambda t}\right) \Theta(t)\,\mathrm{d}t \tag{4}$$

(the last assumption is then $1 \ll \ln(N_0\lambda/w)$). Here $\Theta(t)$ is the Heaviside function which ensures the positivity of the arrival time. We recognize in (4) a Gumbel distribution with average $\langle t_1 \rangle = (1/\lambda)[\ln(N_0\lambda/w) - \gamma]$, where γ is the Euler constant. The variance is $\mathrm{Var}(t_1) = \pi/\sqrt{6}\lambda$ and does not depend on w/N_0 (the contribution of the negative values of t in the Gumbel distribution has to be negligible which is satisfied if $\int_{-\infty}^{0} P(t)\,\mathrm{d}t = w/N_0\lambda \ll 1$). Within these assumptions, we obtain a good agreement between results of numerical simulations using discretized travel events [1] and the theory which uses continuous approximations (see figure 1).

doi:10.1088/1742-5468/2007/09/L09001

J. Stat. Mech. (2007) L09001

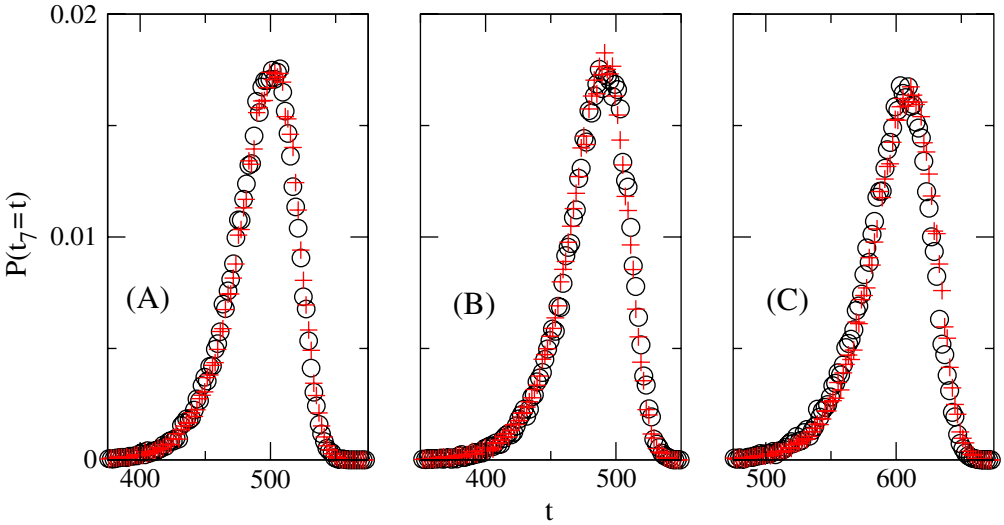

Figure 2. (A)–(C) Arrival time distribution on a line at city #7, from numerical simulations for a fixed random set of populations $\{N_i\}$ and weights $\{w_i\}$ (black circles). Red crosses: distributions for (A) uniform travel $w_i = \bar{w}$ and populations $\{N_i\}$; (B) uniform populations $N_i = \bar{N}$ and weights $\{w_i\}$; (C) uniform populations $N_i = \bar{N}$ and weights $w_i = \bar{w}$. We use a small value of n since most real complex networks are small-world: any node lies at a small distance from the seed.

We now consider the case of a one-dimensional line of cities connected by passenger fluxes of random intensity. We assume that the spreading process starts at city 0 and we denote by t_n the arrival time in city n. The quantities having the same unit as t_n are $1/\lambda$ and N_i/w_i, where w_i is the number of passengers traveling from i to $i+1$ per unit time. Dimensional analysis then implies that the probability distribution of the adimensional quantity λt_n must be a function of the other adimensional quantities which are the $w_i/(\lambda N_i)$: $P(\lambda t_n) = G_n(\lambda t_n, \{w_i/N_i\lambda\})$, where G_n is an unknown function. One can write t_n as a sum of random variables, $\Delta_i = t_i - t_{i-1}$ which are, however, correlated since each local infection process depends on the history of the epidemics in all previously infected cities. While a complete study of $P(\lambda t_n)$ is left for future work [22], numerical simulations of the spreading show (figure 2) that it obeys important invariance properties. For heterogeneous populations and travels (w_i and N_i are distributed uniformly in $[10, 2000]$ and $[10^5, 2\times10^7]$, respectively), the whole distribution is invariant when one replaces (i) all the random weights by their geometrical mean $\overline{w} = (\prod_{i=0}^{n-1} w_i)^{1/n}$; (ii) all the random populations by their geometrical mean $\overline{N} = (\prod_{i=0}^{n-1} N_i)^{1/n}$; (iii) all weights by $\overline{w}$ and all populations by $\overline{N}$. The ratios of the average times for these different sets stay very close to 1, with deviations at most of the order of 5%.

The average arrival time can thus be written as $\lambda\langle t_n\rangle = F(\{\frac{w_i}{N_i\lambda}\})$, where $F(x_1, \ldots, x_n)$ is a symmetric function of its variables which depends only on the product $\prod x_i$, and such that $\langle t_1\rangle$ is the average of the Gumbel distribution (4). This leads to the following ansatz:

$$\lambda\langle t_n\rangle \approx \chi(n) \equiv \ln\left[\prod_{i=0}^{n-1} \frac{N_i\lambda \mathrm{e}^{-\gamma}}{w_i}\right]. \tag{5}$$

J. Stat. Mech. (2007) L09001

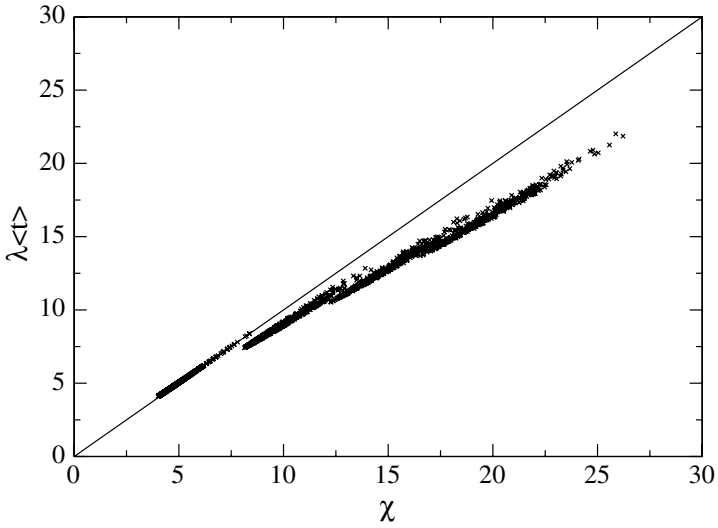

Figure 3. $\lambda\langle t_n\rangle$ versus $\chi(n)$ for five cities connected on a line, with 100 different random sets $\{w_i, N_i\}$. Each point is an average over 1000 epidemics for each realization of the random weights.

Figure 3 shows that the average arrival time in a city is indeed determined by χ to a very good extent (while the arrival time at a given topological distance from the seed can vary a lot). More quantitatively, χ is approximately proportional to $\lambda\langle t\rangle$, which it slightly overestimates since we neglect the flow of infectious individuals from $n-2$ to $n-1$ with respect to the endogenous increase of I_{n-1} during $[t_{n-1}; t_n]$ [22].

We now consider a generic transportation network between the cities. The quantity (5) can easily be computed on any path of length n on the network. While the spread can *a priori* follow multiple paths from one city to another, we can reasonably assume that the most probable path is the one which minimizes the value of χ computed on it, leading to the smallest arrival time possible (a more refined ansatz taking into account multiple paths does not lead to strong differences in the final results [22]). We thus obtain the following ansatz for the arrival time at a city t of a disease starting at node s:

$$\chi(s,t) = \min_{\{P_{st}\}} \sum_{(k,l)\in P_{st}} \left[\ln\left(\frac{N_k\lambda}{w_{kl}}\right) - \gamma\right], \tag{6}$$

where $\{P_{st}\}$ is the set of all possible paths connecting s to t, and the sum is over the links (k,l) on the paths. In other terms, we have introduced a new (non-symmetric) weight $\ln(N_i\lambda/w_{ij}) - \gamma$ on each oriented link (i,j) of the network (note that, since the weights are real-valued, it is highly improbable that two different paths with the same sum of weights exist).

We have simulated, using the model developed in [1], and summarized above, a spreading phenomenon on a subnetwork of the WAN, composed of the 2400 nodes for which the populations are larger than 10 000 inhabitants and which corresponds to 98% of the total traffic[5]. The arrival times are computed by solving numerically the equations

[5] Similar results are obtained for simulations on various network models, but we present results for the WAN which contains additional correlations which may not be present in many models [16].

J. Stat. Mech. (2007) L09001

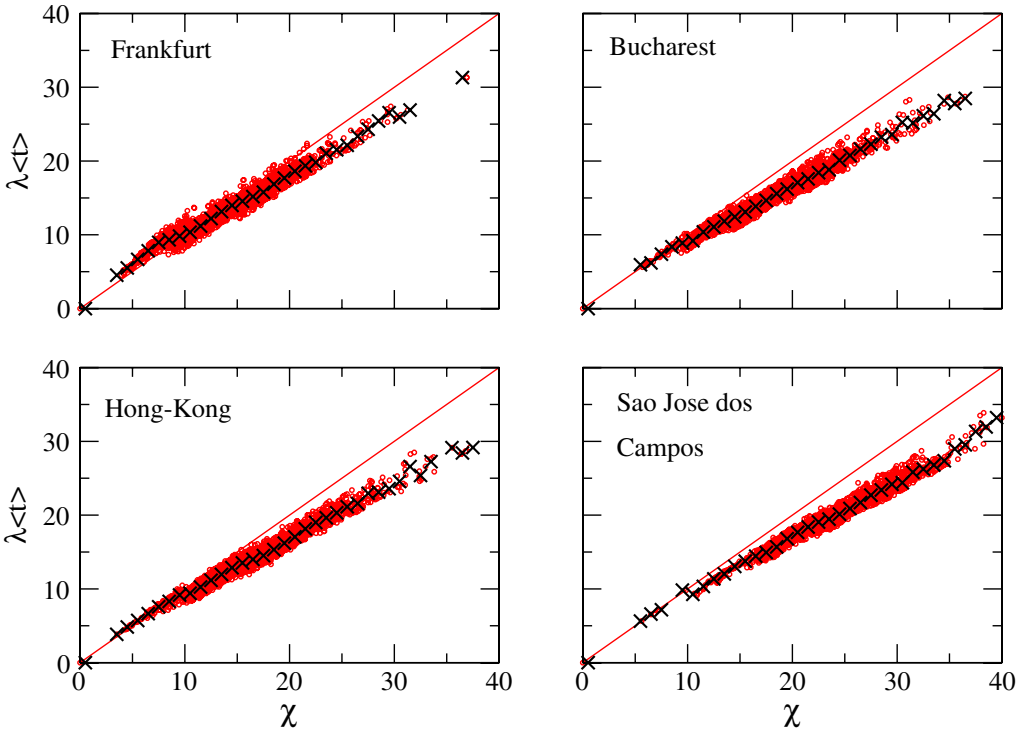

Figure 4. $\lambda\langle t\rangle$ versus χ on the WAN for diseases starting in different cities (whose name is specified in each graph). Each red circle corresponds to a city and averages are done over 1000 realizations of the spreading. Crosses are an average over cities with the same χ. When the initial seed is a hub, the average arrival time is larger than χ in the first cities reached, due to the multiplicity of possible paths [22].

of the Rvachev–Longini model with discretized random travel events and averaging over 1000 realizations of the spreading with the same seed (one infectious individual in a given city). Figure 4 shows the obtained values of $\lambda\langle t\rangle$ versus χ for various initial seeds. We observe that the average arrival time is indeed determined by the value of χ in a given city: various cities with the same χ are reached at the same time by the disease propagation. While χ quantitatively overestimates the arrival time, the two quantities are correlated strongly enough, in order to obtain with a good confidence the order of arrival of the disease in different cities. More precisely, if we denote $\Delta\chi(i,j) = |\chi(j) - \chi(i)|$, we show in figure 5 the probability $f_c(\Delta\chi)$ that the arrival times *in one realization of the spread* $t^{(i)}$ and $t^{(j)}$ follow the same order as given by $\chi(i)$ and $\chi(j)$ [i.e. $(t^{(i)} - t^{(j)})(\chi(i) - \chi(j)) > 0$]. In other words, f_c is the probability that the disease arrival rank for the two cities i and j is correctly predicted by χ. If $\Delta\chi(i,j)$ is equal to 0, no prediction is possible and we indeed obtain $f_c(0) = 0.5$. For $\Delta\chi > 10$, almost all node couples are correctly ranked. This result has, however, to be weighted by the number of couples with such a large $\Delta\chi$. We thus plot on the same figure the cumulative distribution $p_>(\Delta\chi)$ of the number of couples of nodes with a given value of $\Delta\chi$. We see, for example, that approximately 80% of the couples of cities have a $\Delta\chi > 2$ and more than 70% of these couples are correctly sorted (instead of just 50% on average if no information is available).

J. Stat. Mech. (2007) L09001

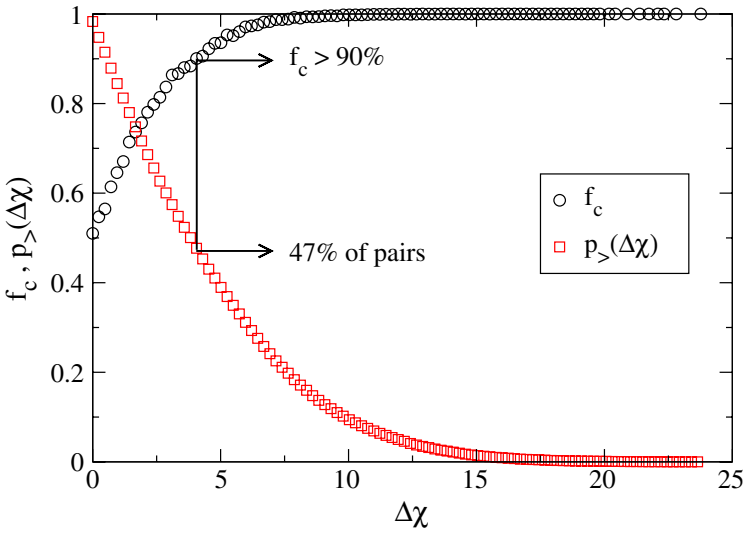

Figure 5. Fraction of couples of nodes correctly ranked as a function of their $\Delta\chi$ (circles), in each realization of the spread, and cumulative distribution (squares) of the values of $\Delta\chi$ (i.e. fraction of couples of cities (i, j) with $\Delta\chi(i, j) > \Delta\chi$).

From a theoretical point of view, metapopulation models go far beyond classical random walks and deserve many further theoretical investigations. In this letter, we have proposed an ansatz for the arrival time of a disease in a city, knowing the starting point of the spread. This ansatz is a good approximation and predicts with accuracy the arrival order of the disease in the different cities, even if they are at the same topological distance from the seed [22]. Containment strategies could use such information to target the cities most at risk of rapid infection, and therefore deploy limited supplies of vaccine or antivirals in an efficient way. Further developments could include more sophisticated compartmental or metapopulation models, and the systematic investigation of various structures of complex networks [22]. Finally, it would be interesting to extend this study to other scales, like the urban scale, where nodes are locations such as homes, offices or malls [10].

We thank V Colizza and A Vespignani for discussions and comments. AG and MB also thank the School of Informatics, Indiana University where part of this work was performed. We are very grateful to V Colizza for providing the data on city populations and we thank IATA (http://www.iata.org) for making their database available to us. AG and AB are partially supported by EU contract 001907 (DELIS).

References

[1] Colizza V, Barrat A, Barthélemy M and Vespignani A, 2006 *Proc. Natl Acad. Sci. USA* **103** 2015
Colizza V, Barrat A, Barthélemy M and Vespignani A, 2006 *Bull. Math. Biol.* **68** 1893
[2] http://www.who.int/csr/sars/en
[3] Hufnagel L, Brockmann D and Geisel T, 2004 *Proc. Natl Acad. Sci. USA* **101** 15124
[4] Anderson R M and May R M, 1992 *Infectious Diseases in Humans* (Oxford: Oxford University Press)
[5] Hethcote H W and Yorke J A, 1984 *Lect. Notes Biomath.* vol 56 (Berlin: Springer)
[6] Keeling M J, 1999 *Proc. R. Soc.* B **266** 859
[7] Pastor-Satorras R and Vespignani A, 2001 *Phys. Rev. Lett.* **86** 3200

J. Stat. Mech. (2007) L09001

[8] Lloyd A L and May R M, 2001 *Science* **292** 1316
[9] Ferguson N M, Keeling M J, Edmunds W J, Gani R, Grenfell B T, Anderson R M and Leach S, 2003 *Nature* **425** 681
[10] Eubank S, Guclu H, Anil Kumar V S, Marathe M V, Srinivasan A, Toroczkai Z and Wang N, 2004 *Nature* **429** 180
[11] Colizza V, Pastor-Satorras R and Vespignani A, 2007 *Nat. Phys.* **3** 276
[12] Albert R and Barabási A-L, 2000 *Rev. Mod. Phys.* **74** 47
[13] Dorogovtsev S N and Mendes J F F, 2003 *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford: Oxford University Press)
[14] Pastor-Satorras R and Vespignani A, 2003 *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge: Cambridge University Press)
[15] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U, 2006 *Phys. Rep.* **424** 175
[16] Barrat A, Barthélemy M, Pastor-Satorras R and Vespignani A, 2004 *Proc. Natl Acad. Sci. USA* **101** 3747
[17] Guimerà R and Amaral L A N, 2004 *Eur. Phys. J.* B **38** 381
[18] Noh J D and Rieger H, 2004 *Phys. Rev. Lett.* **92** 118701
[19] Barthélemy M, Barrat A, Pastor-Satorras R and Vespignani A, 2004 *Phys. Rev. Lett.* **92** 178701
[20] Crépey P, Alvarez F P and Barthélemy M, 2006 *Phys. Rev.* E **73** 046131
[21] Rvachev L A and Longini I M, 1985 *Math. Biosci.* **75** 3
[22] Gautreau A, Barrat A and Barthélemy M, 2007 in preparation
[23] Wu Z, Braunstein L A, Colizza V, Cohen R, Havlin S and Stanley H E, 2006 *Phys. Rev.* E **74** 056104

J. Stat. Mech. (2007) L09001