Effective distances for epidemics spreading on complex networks

Flavio Iannelli,[1,*] Andreas Koher,[2] Dirk Brockmann,[3,4] Philipp Hövel,[2] and Igor M. Sokolov[1]

[1]*Institute for Physics, Humboldt-University of Berlin, Newtonstraße 15, 12489 Berlin, Germany*

[2]*Institute for Theoretical Physics, Technische Universität Berlin, Hardenbergstraße 36, 10623 Berlin, Germany*

[3]*Robert Koch-Institute, Nordufer 20, 13353 Berlin, Germany*

[4]*Institute for Theoretical Biology and Integrative Research Institute of Life Sciences, Humboldt-University of Berlin, Philippstraße 13, Haus 4, 10115 Berlin, Germany*

(Received 22 August 2016; published 17 January 2017)

We show that the recently introduced logarithmic metrics used to predict disease arrival times on complex networks are approximations of more general network-based measures derived from random walks theory. Using the daily air-traffic transportation data we perform numerical experiments to compare the infection arrival time with this alternative metric that is obtained by accounting for multiple walks instead of only the most probable path. The comparison with direct simulations reveals a higher correlation compared to the shortest-path approach used previously. In addition our method allows to connect fundamental observables in epidemic spreading with the cumulant-generating function of the hitting time for a Markov chain. Our results provides a general and computationally efficient approach using only algebraic methods.

DOI: 10.1103/PhysRevE.95.012313

I. INTRODUCTION

Networks have received growing attention in the past decade particularly due to their applicability in describing a wide range of phenomena. Prominent examples are the mapping of the World Wide Web and structure of the Internet, social and financial networks, epidemiology, and language dynamics [1–4]. In the context of epidemic spreading the prediction of outbreaks has become particularly important for public health issues. The rapid growth in the velocity of transportation means and frequency of movements has further increased the risk that global emergent diseases such as H1N1 [5], SARS [6], or EBOV [7], and more recently ZIKV [8], will spread worldwide. The underlying mobility networks are usually scale-free [9]. This implies the absence of an epidemic threshold [10] that allows any transmittable disease to spread through the global population.

The large amount of traffic data both at the local and global scale, which became available in recent years, provides a new opportunity to understand such processes. On the one hand numerical simulations of infection spreading offer a practical tool for estimating the infection arrival time [11]. In this regard metapopulation models [12–14] provide a reasonable tool for maintaining a high level of complexity in the simulation of pandemics [11]. At the global scale the different subpopulations, defined by the nodes of the network, are cities that can be connected by directed or undirected fluxes of individuals, provided by the worldwide transportation network data. On the other hand algebraic methods give a solid foundation for drawing general conclusions and in many cases provide numerical instruments superior to direct simulations.

In this work we introduce a network-based measure that generalizes the concept of distance and that provides fundamental insights into the dynamics of disease transmission as well as an efficient numerical estimation of the infection arrival time. We compare this *effective distance* [16] with the numerical estimate of the transmission times using a metapopulation model to validate the method. A series of papers have already been devoted to this problem [17–21]. Most of them rely on the concept of most probable path, the shortest-path effective distance D_{ij}^{SP} for each source i and target j in the network. The latter can be defined, for both directed and undirected networks, as the geodesic graph distance with edge weights given by the first moment of a Gumbel distributed variable which depends only on the network topology and the infection rate. This shortest-path approach, however, significantly overestimates the infection arrival times [18,22]. A more realistic scenario takes into account all possible paths [19] yielding the multiple-path distance D_{ij}^{MP}, which is better suited to estimate the arrival times of the infection. This method allows, at least in principle, to take into account all possible directed transmission paths, although the computation becomes infeasible as their number grows exponentially with the the number of vertices in the network. The lack of a practical computational approach leads back to considering only the shortest path. A logarithmic *ad hoc* edge weight transformation was introduced in Ref. [20] by simply requiring that adding edges should translate to multiplying the associated probabilities. This follows the intuitive argument that a higher number of passengers reduces the separation distance between neighboring nodes. This logarithmic transformation can be viewed as a log-space reduction [23] and, as we will show later, it can be derived as a special case of the shortest-path effective distance defined previously in Ref. [18]. The estimated arrival time T_{ij}^{AR} from node i to node j, obtained from numerical simulations, correlates highly with the shortest-path effective distance. Using this metrics the complexity of disease transmission can be understood in terms of circular waves propagating at constant velocity from the infected node at time zero to all other nodes in the network [20]. The measure we introduce here aims to generalize previous definitions by including all walks that connect source and target.

*iannelli.flavio@gmail.com

II. EPIDEMIOLOGICAL MODEL

Let us consider a metapopulation susceptible-infected-removed dynamics $S \xrightarrow{\beta SI} I \xrightarrow{\mu I} R$, where β and μ are the infection and recovery rate, respectively. The nodes of the metapopulation network consists of subpopulations of constant size N_j, which divides into susceptible (S_j), infected (I_j), and removed (R_j) individuals

$$N_j = S_j^{(t)} + I_j^{(t)} + R_j^{(t)}. \tag{1}$$

In the metapopulation in addition to the local SIR reaction dynamics, the movement of a host between populations i and j is governed by the master equation

$$\partial_t X_j^{(t)} = \sum_{i \neq j} \left(X_i^{(t)} Q_{ij} - X_j^{(t)} Q_{ji} \right). \tag{2}$$

Here we introduced $X_j^{(t)} = \{S_j^{(t)}, I_j^{(t)}, R_j^{(t)}\}$ as a placeholder variable, and the transition rates Q_{ij} are defined as the conditional probability of a randomly chosen individual to jump from location i to location j within one time step

$$\mathbb{P}\left(X_j^{(t+\Delta t)} \middle| X_i^{(t)} \right) \approx Q_{ij} \Delta t, \quad i \neq j. \tag{3}$$

The transition rates Q_{ij} can equivalently be defined in terms of the weighted adjacency matrix A_{ij}^W as $Q_{ij} = A_{ij}^W / N_i \in [0,1]$. The latter is obtained from the actual passenger fluxes between any two airports in the global mobility network used in the simulations [15]. This network consists of $V = 3865$ vertices (airports) and $E = 51\,440$ directed edges (fluxes), with very broad degree and weight distributions (see Figs. 1 and 2), where the high heterogeneity in the network connectivity is graphically reproduced. For the network diameter we found $D = 16$ (connecting Stuart Island to Narsaq Kujalleq Heliport) and the global clustering coefficient is $c = 0.26 \pm 0.01$. A peculiar feature of this network is that the antisymmetric part of the fluxes $\chi = A_{ij}^W - A_{ji}^W$ is vanishing to a high degree of accuracy so that it can be considered as undirected [24]. The weighted adjacency matrix of the undirected air traffic network is then defined by $A_{ij}^W = A_{ij} W_{ij}$, where A_{ij} is the adjacency matrix element and $W_{ij} \geqslant 0$ the corresponding weight. The symmetry in the adjacency matrix implies that for adjacent populations

$$A_{ij}^W = Q_{ij} N_i = Q_{ji} N_j = A_{ji}^W. \tag{4}$$

The Markov transition matrix associated to the network can be written in terms of both the fluxes A_{ij}^W and the local transition rates Q_{ij}:

$$P_{ij} = \frac{A_{ij}^W}{\sum_j A_{ij}^W} = \frac{Q_{ij}}{\sum_j Q_{ij}}, \tag{5}$$

and it is row stochastic by construction. From (4) we also have the detailed balance

$$P_{ij} k_i^W = P_{ji} k_j^W, \tag{6}$$

where we have introduced the weighted degree $k_i^W = \sum_j A_{ij}^W$, sometimes denoted as strength [25], as the asymptotic probability distribution for the corresponding Markov chain. Thus using (6) we can rewrite (2), in terms of the compartment densities $x_j^{(t)} = X_j^{(t)} / N_j$ to obtain

$$\begin{aligned} \partial_t x_j^{(t)} &= \frac{1}{N_j} \sum_{i \neq j} \left(x_i^{(t)} A_{ij}^W - x_j^{(t)} A_{ji}^W \right) \\ &= \frac{k_j^W}{N_j} \sum_{i \neq j} P_{ji} \left(x_i^{(t)} - x_j^{(t)} \right). \end{aligned} \tag{7}$$

Furthermore, we can remove the dependence on the population size N_j by introducing a constant global mobility rate α. This parameter is defined as the ratio $\alpha = \Phi / N$, between the total daily passenger flux $\Phi = \sum_{ij} A_{ij}^W$ and the total population $N = \sum_i N_i$, i.e., the rate to leave a node for a randomly chosen individual. A local node dependent mobility rate can also be defined as

$$\alpha_i = \frac{\sum_j A_{ij}^W}{N_i} = \sum_j Q_{ij}. \tag{8}$$

In the global mobility network data the total traffic of each node, i.e., its weighted degree k_j^W, is with a good accuracy proportional to its population N_i via the global mobility rate

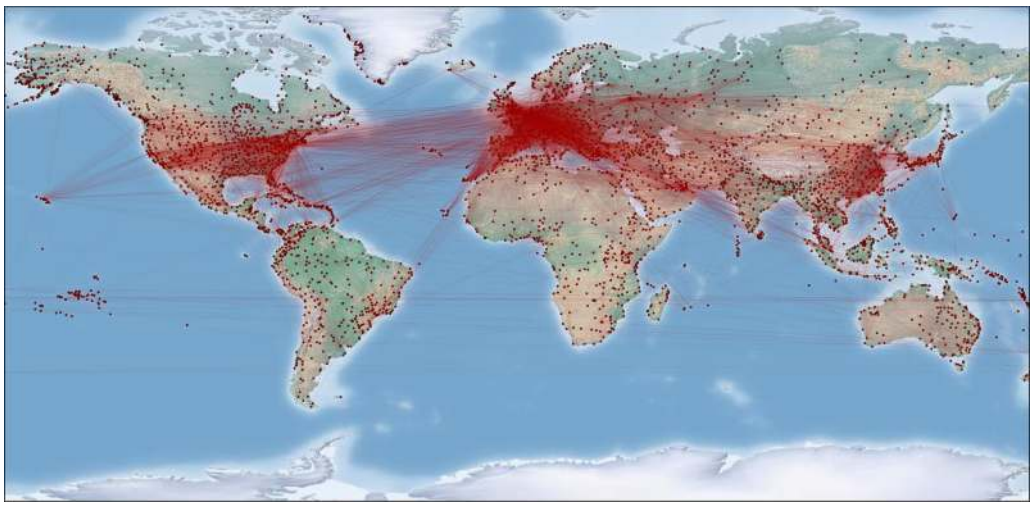

FIG. 1. The global mobility network used in the simulations consisting of $V = 3865$ airports and $E = 51\,440$ flights [15].

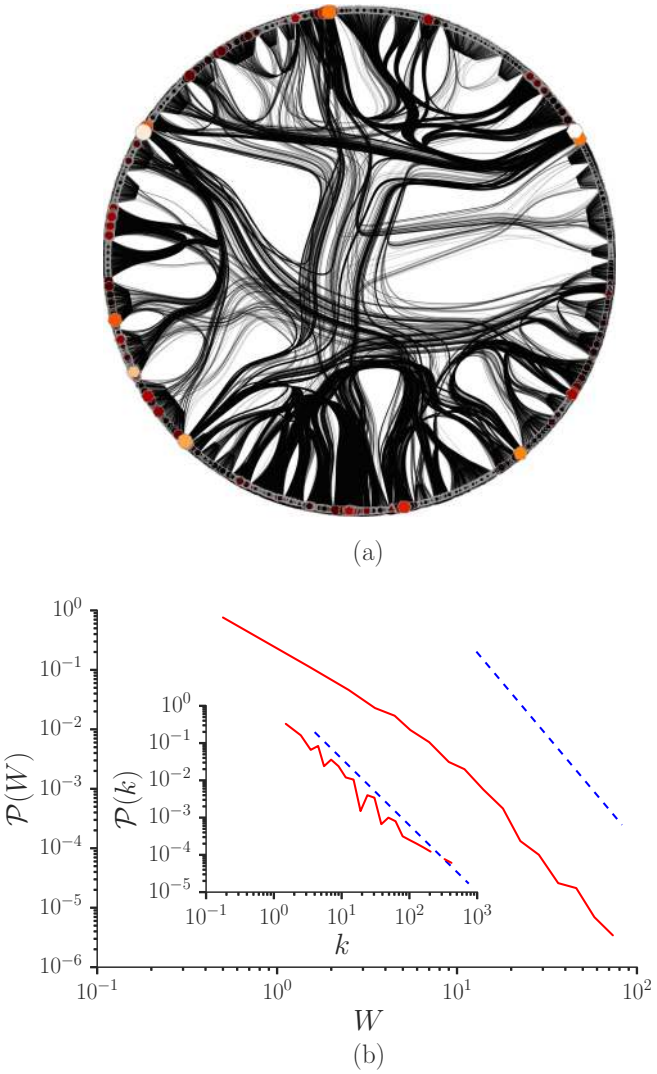

FIG. 2. (a) Circular representation of the global mobility network with vertex colors and size corresponding to its strength k_i^W. (b) Normalized weight distribution $\mathcal{P}(W) \sim W^{-\gamma_w}$ with scaling exponent $\gamma_w = 3.60 \pm 0.14$. Inset: Unweighted degree distribution $\mathcal{P}(k) \sim k^{-\gamma_k}$ with scaling exponent $\gamma_k = 1.79 \pm 0.10$. Scaling exponents are obtained using the method presented in Ref. [27].

α, thus in our case $\alpha_i = \alpha \ \forall i$. The complete SIR dynamics of the Rvachev-Longini model [20,26] becomes

$$\begin{aligned} \partial_t s_j^{(t)} &= \Omega\left(\{s_j^{(t)}\}\right) - \beta s_j^{(t)} i_j^{(t)} \\ \partial_t i_j^{(t)} &= \Omega\left(\{i_j^{(t)}\}\right) + \beta s_j^{(t)} i_j^{(t)} - \mu i_j^{(t)} \\ \partial_t r_j^{(t)} &= \Omega\left(\{r_j^{(t)}\}\right) + \mu i_j^{(t)} \end{aligned} , \tag{9}$$

where the mobility function for each compartment density $x_j = X_j/N_j$,

$$\Omega(\{x_j\}) = \alpha \sum_{i \neq j} P_{ji}\left(x_i^{(t)} - x_j^{(t)}\right), \tag{10}$$

accounts for diffusion along the nodes.

Integrating system (9) we obtain the contagion dynamics on the transportation network A_{ij}^W with the global mobility rate α and the infection parameters β and μ. Finally, the arrival time T_{ij}^{AR} for each source-target pair in the global mobility network is computed by considering a node j infected as soon as one infected individual is present. After introducing the effective distance we use T_{ij}^{AR} to compare the goodness of the different measures.

III. EFFECTIVE NETWORK-BASED MEASURES

The fundamental metric on a network is given by the (geodesic) shortest-path length over all paths Γ_{ij} connecting node i to node j. In a weighted network for each edge $(k,l) \in \Gamma_{ij}$ no node is visited more than once and contributes to the total length with its corresponding weight [1]

$$D_{ij} = \min_{\{\Gamma_{ij}\}} \sum_{(k,l) \in \Gamma_{ij}} \frac{1}{A_{kl}^W}, \tag{11}$$

where the inverse is used because in our case a higher flux of passengers reduces the distance between two nodes. Starting from this definition it is possible to extend the notion of distance by replacing the adjacency matrix weight with an effective quantity that captures the essence of the problem of predicting when the disease will arrive at a certain place. In Ref. [18] the authors defined the shortest-path distance D_{ij}^{SP} by first considering the susceptible-infected model with only two cities, by then generalizing to arbitrary topologies. We first show how their analytical approach leads to an equivalent definition of effective distance used in Ref. [20], and then we generalize it to more realistic propagation scenarios where all paths are taken into account.

Let us consider two susceptible populations i and j and place an infected individual in i at time $t_i = 0$. Assuming $Q_{ij}\Delta t \ll 1$ we can derive a probability density function for the infection hitting time h_j to city j of the Gumbel type [18]. The first moment of this distribution is given by

$$\langle h_j \rangle_i = \frac{1}{\beta}\left(\ln \frac{\beta}{Q_{ij}} - \gamma_e\right), \tag{12}$$

where $\gamma_e \approx 0.5772$ is the Euler-Mascheroni constant, and the average $\langle \ldots \rangle_i$ is taken over times starting at $t_i = 0$. Using (5) we find

$$\begin{aligned} \beta \langle h_j \rangle_i &= \ln \frac{\beta}{\sum_j Q_{ij}} - \gamma_e - \ln \frac{Q_{ij}}{\sum_j Q_{ij}} \\ &= \delta - \ln P_{ij}, \end{aligned} \tag{13}$$

where we assume the mobility rate (8) to be node independent, i.e., $\alpha_i = \alpha \ \ \forall i$ and $\delta = \ln(\beta/\alpha) - \gamma_e$ is a dimensionless parameter. This result can easily be generalized to the SIR model and a network with an arbitrary number of nodes simply by minimizing the quantity $\langle h_l \rangle_k$ for each consecutive link (k,l) that belongs to the path Γ_{ij} connecting source i to target j. For the arbitrary heterogenous population and network topology with an arbitrary number of nodes by taking the minimum over all paths yields the shortest-path effective distance

$$D_{ij}^{\mathrm{SP}}(\delta) \equiv \min_{\{\Gamma_{ij}\}} \sum_{(k,l) \in \Gamma_{ij}} (\delta - \ln P_{kl}) \approx (\beta - \mu) \langle h_j \rangle_i , \tag{14}$$

where for the SIR metapopulation dynamics we have

$$\delta = \ln\left(\frac{\beta-\mu}{\alpha}\right) - \gamma_e. \tag{15}$$

Since each term in the sum is strictly positive, one can obtain the complete shortest-path distance matrix for each source and target pair using the Dijkstra algorithm [28] in a time $O(VE + V^2 \log V)$, where E and V are the graph size and order, respectively. The effective distance defined in Ref. [20] can then be recovered as a special case of (14) simply by setting the scale parameter δ to unity. This fact allows for a deeper and more complete understanding of this effective distance and offers a more solid explanation to the extremely high correlation with the infection arrival time found in Ref. [20]. The most important limitation of (14) is that only one path is considered, namely, the path that in addition to minimizing the topological length also maximizes the probability associated to that. Thus in this scenario the contribution to the disease spread comes only from a single route. The effective infection arrival time $D_{ij}^{\mathrm{SP}}(\delta)/\mathcal{V}^{\mathrm{EF}}$, where $\mathcal{V}^{\mathrm{EF}} \approx \beta - \mu$ is the linearized effective speed of the infection [18,20], is in fact overestimated with respect to the simulations result T_{ij}^{AR} [22].

To take into account the most realistic disease spread scenario one has to consider instead the multiplicity of transmission routes. For two distinct paths Γ_{ij} and Γ'_{ij} connecting i with j, the authors in Ref. [19] found that a two-path distance D_{ij}^{2P} satisfies the equation

$$e^{-D_{ij}^{\mathrm{2P}}} = e^{-D^{\Gamma_{ij}}} + e^{-D^{\Gamma'_{ij}}}, \tag{16}$$

where

$$D^{\Gamma_{ij}}(\delta) = \ln\left(\prod_{(k,l)\in\Gamma_{ij}} \frac{e^{\delta}}{P_{kl}}\right) \tag{17}$$

is the effective distance associated to the path Γ_{ij} of arbitrary length, which corresponds to a Gumbel distributed hitting time. Relation (16) can then be easily generalized to an arbitrary number of multiple paths as

$$\exp\left(-D_{ij}^{\mathrm{MP}}(\delta)\right) = \sum_{\{\Gamma_{ij}\}} \exp[-D^{\Gamma_{ij}}(\delta)], \tag{18}$$

so that we obtain

$$D_{ij}^{\mathrm{MP}}(\delta) = -\ln\left[\sum_{\{\Gamma_{ij}\}} e^{-n_{ij}\delta} F_{ij}(\Gamma_{ij})\right]. \tag{19}$$

Here we have defined the total probability associated to the path Γ_{ij} as

$$F_{ij}(\Gamma_{ij}) = \prod_{(k,l)\in\Gamma_{ij}} P_{kl}. \tag{20}$$

By grouping all probabilities associated to paths of the same length in the quantity $F_{ij}(n) = \sum_{|\Gamma_{ij}|=n} F_{ij}(\Gamma_{ij})$, we can replace in (19) the sum over all paths connecting i to j with a sum over integer path lengths to get

$$D_{ij}^{\mathrm{MP}}(\delta) = -\ln\left[\sum_{n=1}^{n_{\max}} e^{-n\delta} F_{ij}(n)\right], \tag{21}$$

where $n_{\max}$ is the maximum path length in the network. If we select the path $\widetilde{\Gamma_{ij}}$ of length $\widetilde{n}$ that is associated to the dominant contribution, i.e., the path that maximize its associated probability and minimizes the topological path length, one recovers the shortest-path effective distance of (14)

$$\widetilde{D_{ij}^{\mathrm{MP}}}(\delta) = \widetilde{n}\delta - \ln F_{ij}(\widetilde{n}) = D_{ij}^{\mathrm{SP}}(\delta). \tag{22}$$

Therefore the multiple-path distance gives a more accurate estimate of the infection arrival time, as it allows to take into account the most probable route as well as all possible alternative transmission routes. However, since the total number of paths between i and j can scale as $O(V!)$, the measure D_{ij}^{MP} becomes computationally infeasible for large graphs.

Both measures D_{ij}^{SP} and D_{ij}^{MP} rely on the fact that the epidemic will spread along simple paths, i.e., routes that do not cross themselves. Here we follow instead a different approach and introduce a distance that includes all possible random walks from source to target. Relaxing the assumption of directed spread is equivalent to effectively erasing the memory from the system at each time step. This is achieved by including in (19) all walks Ξ_{ij}, which contrary to the paths Γ_{ij}, allow also already visited nodes. We define the random-walk effective distance by generalizing (19) as

$$D_{ij}^{\mathrm{RW}}(\delta) = -\ln\left[\sum_{\{\Xi_{ij}\}} e^{-n_{ij}\delta} H_{ij}(\Xi_{ij})\right], \tag{23}$$

where $H_{ij}(\Xi_{ij})$ is the probability associated to a walk that starts in i and arrives to j. As for the probabilities F_{ij} we can group the probabilities associated to walks of the same length into $H_{ij}(n) = \sum_{|\Xi_{ij}|=n} H_{ij}(\Xi_{ij})$. The latter is precisely the hitting time probability for a Markov chain defined recursively as [29]

$$H_{ij}(n) = \sum_{k\neq j} P_{ik} H_{kj}(n). \tag{24}$$

Thus $H_{ij}(n)$ is simply the nth power of the subtransition probability matrix obtained by removing the jth row and column. Contrary to the multiple-path scenario now the walks are unbounded and so becomes $n_{\max}$. Furthermore since each term in the sum (23) is positive, assuming the convergence of the sum, we can rearrange it as

$$D_{ij}^{\mathrm{RW}}(\delta) = -\ln\left[\sum_{n=1}^{\infty} e^{-n\delta} H_{ij}(n)\right]. \tag{25}$$

In Fig. 3 we use the Pearson correlation coefficient R^2 for quantifying the accuracy of the different measures using São Paulo airport as the source of the infection. Each dot in the scatter plot corresponds to an airport, which is labeled infected in the simulations when the the infection density is greater than zero. The high correlation with the infection arrival time found in Ref. [20] using a shortest-path approach (light blue) is improved when considering the random-walk effective distance (violet). The points on the dashed diagonal indicate a perfect agreement between the simulation and the effective distance. The correlation distribution considering all nodes in the network as initial infected seed shows that not only the

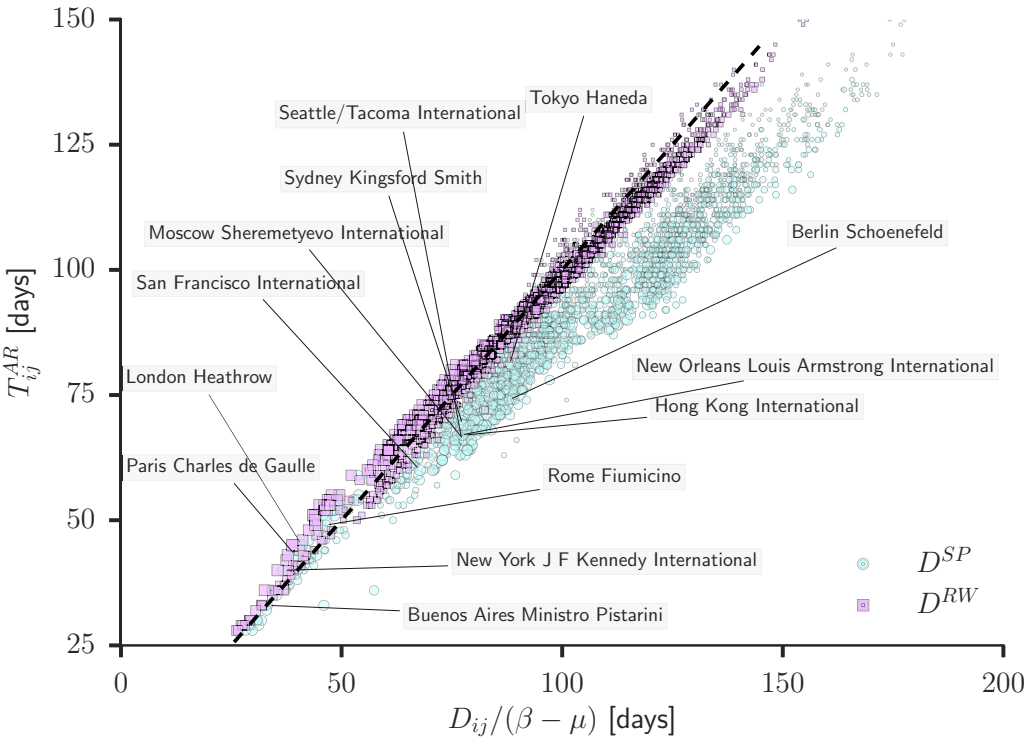

FIG. 3. Correlation of the infection arrival times in days obtained from simulation of (9) with the shortest-path (light blue circles) and the random-walk effective distance (violet squares). The source infected node i is São Paulo Guarulhos International Airport, and each point in the scatter plot corresponds to an airport j in the global mobility network, with size proportional to its strength k_j^W. The mobility and epidemiological parameters are respectively $\alpha = 0.028$ d^{-1}, $\beta = 0.407$ d^{-1}, and $\mu = 0.271$ d^{-1} resulting in $\delta = 1$. The Pearson correlation coefficients are $R^2_{\rm SP} = 0.96$ and $R^2_{\rm RW} = 0.99$.

measure proposed here possesses a higher averaged correlation but it is also more peaked around it (see Fig. 4).

A remarkable interpretation of the random-walk effective distance can be found by noticing that by definition $D_{ij}^{\rm RW}(\delta) = -\ln\langle e^{-\delta h_j}\rangle_i$, where h_j is the hitting time to node j [30]. Thus, since $H_{ij}(0) = 0$ for $i \neq j$, we have the correspondence

$$D_{ij}^{\rm RW}(\delta) = -C_{ij}(-\delta), \tag{26}$$

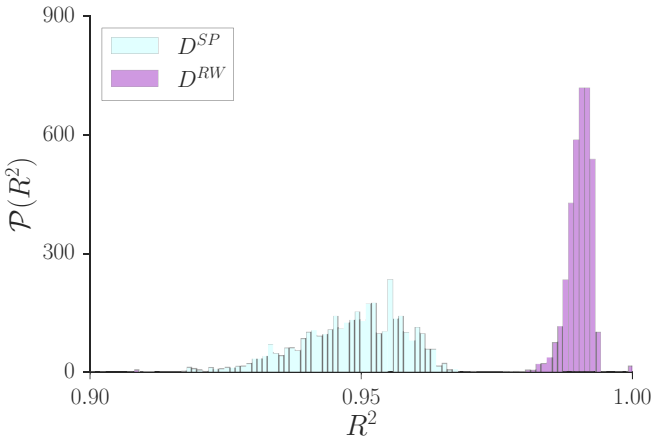

FIG. 4. Distribution of the Pearson coefficient R^2 considering all possible infection sources in the global mobility network for same parameters as in Fig. 3.

with the logarithm of the moment-generating function, i.e., the cumulant-generating function of the hitting time in a Markov chain

$$C_{ij}(s) = \ln\left[\sum_{n=0}^{\infty} e^{ns}\ H_{ij}(n)\right] = \ln\langle e^{sh_j}\rangle_i\,. \tag{27}$$

Hence one obtains the cumulants of the hitting time by differentiating the random-walk effective distance with respect to δ at $\delta = 0$. This interesting correspondence allows one to rigorously relate epidemiological quantities such as the arrival time and the speed of infection in a reaction-diffusion model to the fluctuations of the hitting time. Then one can interpret $D_{ij}^{\rm RW}(\delta)$ as a generalized free energy in a statistical physics perspective [31] and providing a more profound theoretical framework than the *ad hoc* measure proposed in Ref. [20].

From the computational side, in order to evaluate the infinite sum in (25), we can restrict ourselves to the first nonvanishing contributions, which dominate due to the decreasing exponential in the walk length n. However, we can also solve the complete expression by rewriting (25) into a geometric series. This requires to vectorize $D_{ij}^{\rm RW}$ with respect to the arrival node j to obtain

$$d_{i(j)}^{\rm RW}(\delta) = -\ln\{[e^{\delta}\mathbf{I}(j|j) - \mathbf{P}(j|j)]^{-1}\mathbf{p}(j)\}_i, \tag{28}$$

where $\mathbf{P}(j|j)$ and $\mathbf{I}(j|j)$ are the transition and identity submatrices obtained by deleting row and column j [32], while $\mathbf{p}(j)$ is the jth column of $\mathbf{P}$ with element j removed. To obtain the previous expression we have used that for $\delta > 0$, all

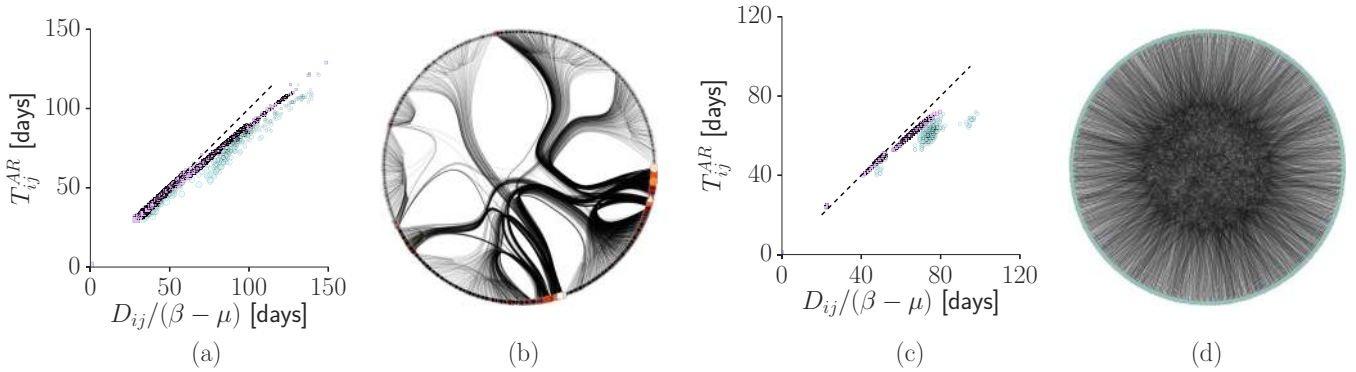

FIG. 5. Correlation of the shortest-path (light blue) and random-walk (violet) approach with the simulations arrival time for the the U.S. airport network used in Ref. [13] (a) and its randomized (Erdős-Rényi) version (c). In (b) and (d) the corresponding networks visualization consisting of $V = 500$ nodes and $E = 5960$ edges. Parameters as in Fig. 3.

eigenvalues of the matrix $e^{-\delta}\mathbf{P}(j|j)$ are strictly smaller than unity. For each arrival node the random-walk effective distance can then be obtained in polynomial time $O(V^{3.4})$ using, for instance, the Coppersmith-Winograd algorithm for matrix inversion [33], making the problem of parallel transmission routes feasible even for large networks as the one used in our simulations.

For highly heterogeneous topologies, such as the air-traffic network [24], only a small number of paths contributes to D_{ij}^{MP}. Taking the limit of the dominant contribution in (23), which corresponds to selecting the dominant path in (19), allows one to neglect the sum over the walks (paths), and it yields as for (22)

$$\widetilde{D_{ij}^{\mathrm{RW}}}(\delta) = -\ln[e^{-\tilde{n}\delta}H_{ij}(\tilde{n})] = D_{ij}^{\mathrm{SP}}(\delta). \tag{29}$$

In Fig. 5 the comparison between the shortest-path and random-walk approach for the U.S. air-traffic network [13] shows that the results presented here are robust also for Erdős-Rényi networks. The correlation coefficients are respectively $R^2_{\mathrm{SP}} = 0.99$ and $R^2_{\mathrm{RW}} = 1.00$ for the U.S. air-traffic network and $R^2_{\mathrm{SP}} = 0.94$ and $R^2_{\mathrm{RW}} = 1.00$ for the randomized Erdős-Rényi network. An higher correlation and stability of the random-walk approach is also observed in the case of artificial networks, as for unweighted Barabási-Albert [9] and lattice models (see Fig. 6). The correlation coefficients for the latter are $R^2_{\mathrm{SP}} = 1.00$ and $R^2_{\mathrm{RW}} = 1.00$ for the Barabási-Albert network and $R^2_{\mathrm{SP}} = 0.99$ and $R^2_{\mathrm{RW}} = 1.00$ for the lattice.

IV. CONCLUSIONS

In summary we have presented a generalization of the concept of effective distance by overcoming the restriction of simple path propagation of a disease. The proposed random-walk effective distance includes the previously defined shortest-path measure as a particular case. The remarkable correlation found with the infection arrival time can be explained as follows. The contribution of looped trajectories in the propagation of physical information is neglected because of the decreasing exponential in the walk length. The latter serves as damping such contributions for long walks, and in particular allows us to neglect infinite loop contributions. In scenarios where multiple parallel paths are important, for instance, in Erdős-Rényi graphs or regular lattices, the assumption of a single dominant path breaks down, and the measure proposed here can be used as an efficient alternative. The predictive power of the random-walk effective distance can be used for containment strategies and estimation of arrival times for real global pandemics from the underlying networks topology. The random-walk metric can in fact be generally applied to

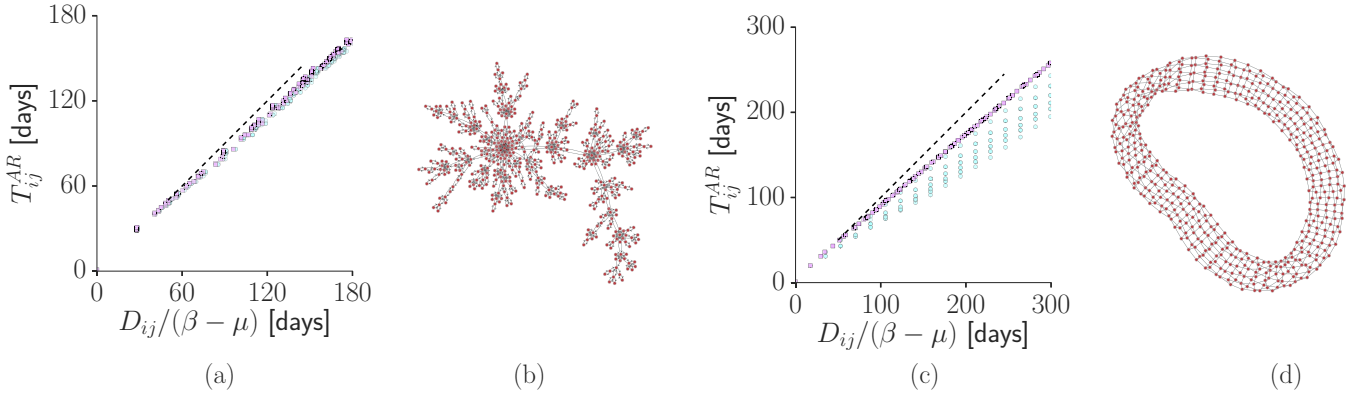

FIG. 6. Correlation of the shortest-path (light blue) and random-walk (violet) approach with the simulations arrival time for an unweighted Barabási-Albert network (a) and a two-dimensional lattice embedding (c) both consisting of $V = 500$ nodes. In (b) and (d) the corresponding networks visualization. The number of edges is, respectively, $E = 998$ and $E = 1000$. Parameters as in Fig 3.

any weighted and directed network besides the transportation ones, for instance, in the context of social interactions and rumor spreading. For unweighted locally treelike networks both the shortest-path and random-walk effective distances yield maximum correlation with the simulated arrival time, as the shortest path tends to dominate.

From a theoretical point of view our results show that the average infection arrival time in a metapopulation model can be approximated by the cumulant-generating function of the hitting time for a Markov chain. In fact, the generating function approach can also be used to formally derive the latter from the first moments of the Gumbel distribution [18]. The connection with the cumulant-generating function allows for an interpretation within statistical physics. In particular this would explain how the different approaches are connected in terms of the entropy associated to paths of fixed length [31,34]. This observation links disease spreading on complex networks with a generic diffusion process. Further developments and extensions of our results include the generalization to temporal networks by considering a set of transition matrices, one for each time step [35,36].

ACKNOWLEDGMENTS

We thank T. Isele for insightful discussions and technical assistance. This work was developed within the scope of the IRTG 1740/TRP 2015/50122-0 and funded by the DFG/FAPESP. A.K. and P.H. acknowledge support by Deutsche Forschungsgemeinschaft in the framework of SFB910.

F.I. and A.K. contributed equally to this work.

[1] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks* (Cambridge University Press, Cambridge, 2008).
[2] R. Pastor-Satorras and A. Vespignani, *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, Cambridge, 2004).
[3] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).
[4] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).
[5] Y. Yang, J. D. Sugimoto, M. E. Halloran, N. E. Basta, D. L. Chao, L. Matrajt, G. Potter, E. Kenah, and I. M. Longini, Science **326**, 729 (2009).
[6] V. Colizza, A. Barrat, M. Barthélemy, and A. Vespignani, BMC Med. **5**, 1 (2007).
[7] C. Poletto, M. F. Gomes, A. P. y Piontti, L. Rossi, L. Bioglio, D. L. Chao, I. M. Longini, M. E. Halloran, V. Colizza, and A. Vespignani, Eurosurveillance **19**, 42 (2014).
[8] S. Ioos, H.-P. Mallet, I. L. Goffart, V. Gauthier, T. Cardoso, and M. Herida, Med. Maladies Infect. **44**, 302 (2014).
[9] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[10] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).
[11] W. V. d. Broeck, C. Gioannini, B. Gonçalves, M. Quaggiotto, V. Colizza, and A. Vespignani, BMC Infect. Dis. **11**, 1 (2011).
[12] H. H. K. Lentz, T. Selhorst, and I. M. Sokolov, Phys. Rev. E **85**, 066111 (2012).
[13] V. Colizza, R. Pastor-Satorras, and A. Vespignani, Nat. Phys. **3**, 276 (2007).
[14] V. Colizza and A. Vespignani, J. Theor. Biol. **251**, 450 (2008).
[15] Global mobility data of air traffic was provided by OAG (www.oag.com).
[16] See http://barabasi.com/networksciencebook/
[17] L. A. Braunstein, S. V. Buldyrev, R. Cohen, S. Havlin, and H. E. Stanley, Phys. Rev. Lett. **91**, 168701 (2003).
[18] A. Gautreau, A. Barrat, and M. Barthélemy, J. Stat. Mech. (2007) L09001.
[19] A. Gautreau, A. Barrat, and M. Barthelemy, J. Theor. Biol. **251**, 509 (2008).
[20] D. Brockmann and D. Helbing, Science **342**, 1337 (2013).
[21] G. Lawyer, BMC Infect. Dis. **16**, 1 (2016).
[22] P. Crepey, F. P. Alvarez, and M. Barthélemy, Phys. Rev. E **73**, 046131 (2006).
[23] M. Roosta, J. Math. Anal. Appl. **88**, 341 (1982).
[24] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, Proc. Natl. Acad. Sci. USA **101**, 3747 (2004).
[25] G. Caldarelli, *Scale-Free Neworks* (Oxford University Press, Oxford, 2007).
[26] L. A. Rvachev and I. M. Longini, Math. Biosci. **75**, 3 (1985).
[27] A. Clauset, C. R. Shalizi, and M. E. J. Newman, SIAM Rev. **51**, 661 (2009).
[28] E. W. Dijkstra, Numer. Math. **1**, 269 (1959).
[29] J. R. Norris, *Markov Chains* (Cambridge University Press, Cambridge, 1998).
[30] J. G. Kemeny, J. L. Snell *et al.*, *Finite Markov Chains*, Vol. 356 (van Nostrand, Princeton, NJ, 1960).
[31] I. Kivimki, M. Shimbo, and M. Saerens, Physica A **393**, 600 (2014).
[32] F. Zhang, *Matrix Theory: Basic Results and Techniques* (Springer Science & Business Media, New York, 2011).
[33] D. Coppersmith and S. Winograd, in *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing* (ACM, New York, 1987), pp. 1–6.
[34] F. Bavaud and G. Guex, in *Social Informatics: Proceedings of the 4th International Conference, SocInfo 2012, Lausanne, Switzerland, December 5–7, 2012*, edited by K. Aberer, A. Flache, W. Jager, L. Liu, J. Tang, and C. Guéret (Springer, Berlin, 2012), pp. 68–81.
[35] H. H. K. Lentz, T. Selhorst, and I. M. Sokolov, Phys. Rev. Lett. **110**, 118701 (2013).
[36] A. Koher, H. H. K. Lentz, P. Hövel, and I. M. Sokolov, PLoS ONE **11**, 1 (2016).