Chapter 4
Modeling and Predicting Human Infectious Diseases

Nicola Perra and Bruno Gonçalves

Abstract The spreading of infectious diseases has dramatically shaped our history and society. The quest to understand and prevent their spreading dates more than two centuries. Over the years, advances in Medicine, Biology, Mathematics, Physics, Network Science, Computer Science, and Technology in general contributed to the development of modern epidemiology. In this chapter, we present a summary of different mathematical and computational approaches aimed at describing, modeling, and forecasting the diffusion of viruses. We start from the basic concepts and models in an unstructured population and gradually increase the realism by adding the effects of realistic contact structures within a population as well as the effects of human mobility coupling different subpopulations. Building on these concepts we present two realistic data-driven epidemiological models able to forecast the spreading of infectious diseases at different geographical granularities. We conclude by introducing some recent developments in diseases modeling rooted in the big-data revolution.

4.1 Introduction

Historically, the first quantitative attempt to understand and prevent infectious diseases dates back to 1760 when Bernoulli studied the effectiveness of inoculation against Smallpox [1]. Since then, and despite some initial lulls [2], an intense research activity has developed a rigorous formulation of pathogens' spreading. In this chapter, we present different approaches to model and predict the spreading of infectious diseases at different geographical resolutions and levels of detail. We focus on airborne illnesses transmitted from human to human. We are the carriers of such diseases. Our contacts and mobility are the crucial ingredients to understand

N. Perra (✉)
Northeastern University, Boston, MA, USA
e-mail: n.perra@neu.edu

B. Gonçalves
Aix Marseille Université, Université de Toulon, CNRS, CPT, UMR 7332,
13288 Marseille, France
e-mail: bgoncalves@gmail.com

B. Gonçalves, N. Perra (eds.), *Social Phenomena*, Computational Social Sciences, DOI 10.1007/978-3-319-14011-7_4

and model their spreading. Interestingly, the access to large-scale data describing these human dynamics is a recent development in epidemiology. Indeed, for many years only the biological roots of transmission were clearly understood, so it is not surprising that classical models in epidemiology neglect realistic human contact structures or mobility in favor of more mathematically tractable and simplified descriptions of unstructured populations. We start our chapter with these modeling approaches that offer us an intuitive way of introducing the basic quantities and concepts in epidemiology.

Advances in technology are resulting in increased data on human dynamics and behavior. Consequently, modeling approaches in epidemiology are gradually becoming more detailed and starting to include realistic contact and mobility patterns. In Sects. 4.3 and 4.4 we describe such developments and analyze the effects of heterogeneities in contact structures between individuals and between cities/subpopulations.

With these ingredients in hand we then introduce state-of-the-art data-driven epidemiological models as examples of the modern capabilities in disease modeling and predictions. In particular, we consider GLEAM [3, 4], EpiSims [5], and FLuTE [6]. The first model is based on the metapopulation framework, a paradigm where the inter-population dynamics is modeled using detailed mobility patterns, while the intra-population dynamics is described by coarse-grained techniques. The other tools are, instead, agent-based model (ABM). This class of tools guarantees a very precise description of the unfolding of diseases, but need to be fed with extremely detailed data and are not computationally scalable. For these reasons their use so far has been limited to the study of disease spread within a limited numbers of countries. In comparison, metapopulation models include a reduced amount of data, while the approximated description of internal dynamics allows scaling the simulations to global scenarios.

Interestingly, the access to large-scale data on human activities has also started a new era in epidemiology. Indeed, the big-data revolution naturally results in real time data on the health related behavior of individuals across the globe. Such information can be obtained with tools that either require the active participation of individuals willing to share their health status or that is mined silently from individuals' health related data. Epidemiology is becoming digital [7, 8]. In Sect. 4.6 we introduce the basic concepts, approaches, and results in this new field of epidemiology. In particular, we describe tools that, using search queries, microblogging, or other web-based data, are able to predict the incidence of a wide range of diseases two weeks ahead respect to traditional surveillance.

4.2 Basic Concepts in Mathematical Epidemiology

Epidemic models divide the progression of the disease into several states or compartments, with individuals transitioning compartments depending on their health status. The natural history of the disease is represented by the type of

compartments and the transitions from one to another, and naturally varies from disease to disease. In some illnesses, Susceptible individuals (S) become infected and Infectious when coming in contact with one or more Infectious (I) persons and remain so until their death. In this case the disease is described by the so-called *SI* (susceptible-infected) model. In other diseases, as is the case for some sexual transmitted diseases, infected individuals recover becoming again Susceptible to the disease. These diseases are described by the *SIS* (susceptible-infected-susceptible) model. In the case of influenza like illnesses (ILI), on the other hand, infected individuals Recover becoming immune to future infections from the same pathogen. ILIs are described by the *SIR* (susceptible-infected-recovered) model. These basic compartments provide us with the fundamental description of the progression of an idealized infection in several general circumstances. Further compartments can be added to accurately describe more realistic illnesses such as Smallpox, Chlamydia, Meningitis, and Ebola [2, 9, 10]. Keeping this important observation in mind, here we focus on the *SIR* model.

4.2.1 Modeling Transitions Between Compartments

Epidemic models are often represented using chart such as the one seen in Fig. 4.1. Such illustrations are able to accurately represent the number of compartments and the disease's behavior in a concise and easily interpretable form. Mathematically, models can also be accurately represented as reaction equations as we will see below.

In general, epidemic models include two type of transitions, "interactive" and "spontaneous." Interactive transitions require the contact between individuals in two different compartments, while spontaneous transitions occur naturally at a fixed rate per unit time. For example, in the transition between S to I, Susceptible individuals become Infected due to the interaction with Infected individuals, i.e. $S+I \rightarrow 2I$. The transition is mediated by individuals in the compartment I, see Fig. 4.1. On the other hand, an Infectious individual can naturally recover from infection after a certain

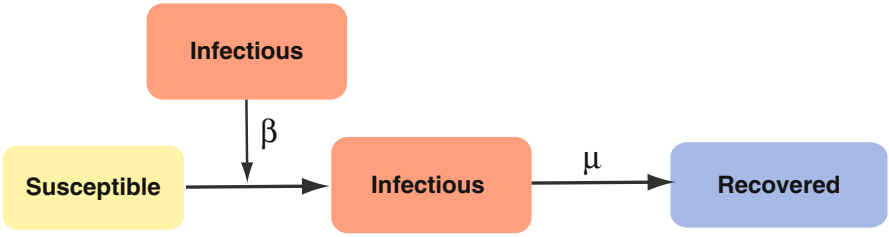

Fig. 4.1 Schematic representation of the SIR model. The transition from S to I is due to the interaction between susceptible and infectious individuals. The transition from I to R is instead spontaneous. The transition rates are β and μ, respectively

amount of time and become Recovered, i.e. $I \to R$. Individuals are considered to have a fixed recovery rate, μ, defined as the inverse of the average time τ spent in the infected compartment, $\mu = \tau^{-1}$

But how can we model the infection process? Intuitively we expect that the probability of single individual becoming infected must depend on (1) the number of infected individuals in the population, (2) the probability of infection given a contact with an infectious agent and, (3) the number of such contacts. In this section we neglect the details of who is in contact with whom and consider instead individuals to be part of a homogeneously mixed population where everyone is assumed to be in contact with everyone else (we tackle heterogeneous contacts in Sect. 4.3). In this limit, the per capita rate at which susceptible contract the disease, the force of infection λ, can be expressed in two forms depending on the type of population. In the first, often called mass-action law, the number of contacts per individual is independent of the total population size, and determined by the transmission rate β and the probability of randomly contacting an infected individual, i.e. $\lambda = \beta I/N$ (where N is the population size). In the second case, often called pseudo mass-action law, the number of contacts is assumed to scale with the population size, and the transmission rate β, i.e. $\lambda = \beta I$. Without loss of generality, in the following we focus on the first kind of contact.

4.2.2 The SIR Model

The *SIR* framework is the crucial pillar to model ILIs. Think, for example, at the H1N1 pandemic in 2009, or the seasonal flu that every year spread across the globe. The progression of such diseases, from the first encounter to the recovery, happens in matters of days. For this reason, birth and death rates in the populations can be generally neglected, i.e. $d_t N \equiv 0$ for all times t.

Let us define the fraction of individuals in the susceptible, infected, and recovered compartments as s, i, and r. The *SIR* model is then described by the following set of differential equations:

$$\begin{cases} d_t s = -s\lambda \\ d_t i = s\lambda - \mu i \\ d_t r = \mu i \end{cases} \tag{4.1}$$

where $\lambda = \beta i \equiv \beta \frac{I}{N}$ is the force of infection, and $d_t \equiv \frac{d}{dt}$. The first equation describes the infection process in a homogeneous mixed population. Susceptible individuals become infected through random encounters with Infected individuals. The second equation describes the balance between the in-flow (infection process, first term), and the out-flow (recovery process, second term) in compartment i. Finally, the third equation accounts for the increase of the recovered population due to the recovery process. Interestingly, the *SIR* dynamical equations, although

apparently very simple, due to their intrinsic non-linearity cannot be solved analytically. The description of the evolution of the disease can be obtained only through numerical integration of the system of differential equations. However, crucial analytic insight on the process can be obtained for early $t \sim t_0$ and late times $t \to \infty$.

4.2.2.1 Epidemic Threshold

Under which conditions a disease starting from a small number, I_0, of individuals at time t_0 is able to spread in the population? To answer this question let us consider the early stages of the spreading, i.e. $t \sim t_0$. The equation for the infected compartment can be written as $d_t i = i(\beta s - \mu)$, indicating an exponential behavior for early times. It then follows that if the initial fraction of susceptible individuals, $s_0 = S_0/N$, is smaller than μ/β, the exponent becomes negative and the disease dies out. We call this value the epidemic threshold [11] of the *SIR* model. The fraction of susceptibles in the population has to be larger than a certain value, that depends on the disease details, in order to observe an outbreak.

Typically, the initial cluster of infected individuals is small in comparison with the population size, i.e. $s_0 \gg i_0$, or $s_0 \sim 1$. In this case, the threshold condition can be re-written as $\beta/\mu > 1$. The quantity:

$$R_0 \equiv \frac{\beta}{\mu} \tag{4.2}$$

is called the *basic reproductive number*, and is a crucial quantity in epidemiology and provides a very simple interpretation of the epidemic threshold. Indeed, the disease is able to spread if and only if each infected individual is able to infect, on average, more than one person before recovering. The meaning of R_0 is then clear: it is simply the average number of infections generated by an initial infectious seed in a fully susceptible population [10].

4.2.2.2 Disease-Free Equilibrium

For any value of $\mu > 0$, the *SIR* dynamics will eventually reach a stationary, disease-free, state characterized by $i = d_t i = 0$. Indeed, infected individuals will keep recovering until they all reach the R compartment. What is the final number of recovered individuals? Answering this apparently simple question is crucial to quantify the impact of the disease. We can tackle such conundrum dividing the first equation with the third equation in the system 4.1. We obtain $d_r s = -R_0 s$ which in turn implies $s_t = s_0 e^{-R_0 r_t}$. Unfortunately, this transcendent equation cannot be solved analytically. However, we can use it to gain some important insights on the *SIR* dynamics. We note that for any $R_0 > 1$, in the limit $t \to \infty$, we must have

$s_\infty > 0$. In other words, despite R_0, the disease-free equilibrium of an SIR model is always characterized by some finite fraction of the population in the Susceptible compartment, or, in other words, some individuals will always be able to avoid the infection. In the limit where $R_0 \sim 1$ we can obtain an approximate solution for r_∞ (or equivalently for $s_\infty = 1 - r_\infty$) by expanding $s_\infty = s_0 e^{-R_0 s_\infty}$ at the second order around $r_\infty \sim 0$. After a few basic algebraic manipulations we obtain $r_\infty = \frac{2(R_0-1)}{R_0^2}$ [9].

4.3 Beyond Homogeneous Mixing

In the previous sections we presented the basic concepts and models in epidemiology by considering a simple view of a population where individuals mix homogeneously. Although such approximation allows a simple mathematical formulation, it is far from reality. Individuals do not all have the same number of contacts, and more importantly, encounters are not completely random [12–15]. Some persons are more prone to social interactions than others, and contacts with family members, friends, and co-workers are much more likely than interactions with any other person in the population.

Over the last decade the *network framework* has been particularly effective in capturing the complex features and the heterogeneous nature of our contacts [12–16]. In this approach, individuals are represented by nodes while links represent their interactions. As described in different chapters of the book (see Chaps. 3, 6, and 10), human contacts are not heterogeneous in both number and intensity [12–15, 17] but also change over time [18]. This framework naturally introduces two timescales, the timescale at which the network connections evolve, τ_G and the inherent timescale, τ_P, of the process taking place over the network. Although the dynamical nature of interactions might have crucial consequences on the disease spreading [19–24], the large majority of results in the literature deal with one of two limiting regimens [25, 26]. When $\tau_G \gg \tau_P$, the evolution of the network of contacts is much slower than the spreading of the disease and the network can be considered as static. On the other hand, when $\tau_P \gg \tau_G$, the links are said to be annealed and changes in networks structure are much faster than the spreading of the pathogen. In both cases the two time-scales are well separated allowing for a simpler mathematical description. Here we focus on the annealed approximation ($\tau_P \gg \tau_G$) that provides a simple stage to model and understand the dynamical properties of epidemic processes. We refer the reader to Chap. 3 Face-to-Face Interactions for recent approaches that relax this time-scale separation assumption.

Let us consider a network $G(N, E)$ characterized by N nodes connected by E edges. The number of contacts of each node is described by the degree k. The degree distribution $P(k)$ characterizes the probability of finding a node of degree k. Empirical observations in many different domains show heavy-tailed degree distributions usually approximated as power-laws, i.e. $P(k) \sim k^{-\alpha}$ [12, 13].

Furthermore, human contact networks are characterized by so-called *assortative mixing*, meaning a positive correlation between the degree of connected individuals. Correlations are encoded in the conditional probability $P(k'|k)$ that a node of degree k is connected with a node of degree k' [12, 13]. While including realistic correlations in epidemic models is crucial [27–29] they introduce a wide set of mathematical challenges that are behind the scope of this chapter. In the following, we consider the simple case of uncorrelated networks in which the interdependence among degree classes is removed.

4.3.1 The SIR Model in Networks

How can we extend the *SIR* model to include heterogeneous contact structures? Here we must take a step further than simply treating all individuals the same. We start distinguishing nodes by degree while considering all vertices with the same degree as statistically equivalent. This is known as the degree block approximation and is exact for annealed networks. The quantities under study are now $i_k = \frac{I_k}{N_k}$, $s_k = \frac{S_k}{N_k}$, and $r_k = \frac{R_k}{N_k}$, where the I_k, S_k, and R_k are the number of infected, susceptible, recovered individuals in the degree class k. N_k instead describes the total number of nodes in the degree class k. The global averages are given by $i = \sum_k P(k)\, i_k$, $s = \sum_k P(k)\, s_k$, $r = \sum_k P(k)\, r_k$. Using this notation and heterogeneous mean field (HMF) theory [26], the system of differential equations (4.1) can now be written as:

$$\begin{cases} d_t s_k = -s_k \lambda_k \\ d_t i_k = s_k \lambda_k - \mu i_k \\ d_t r_k = \mu i_k \end{cases} \tag{4.3}$$

The contact structure introduces a force of infection function of the degree. In particular, $\lambda_k = \gamma k \Theta_k$ where γ is the rate of infection per contact, i.e. $\beta = \gamma k$, and Θ_k describes the density of infected neighbors of nodes in the degree class k. Intuitively, this density is a function of the conditional probability that a node k is connected to any node k' and proportional to the number of infected nodes in each class k': $\Theta_k = \sum_{k'} P(k'|k)\, i_{k'}$. In the simple case of uncorrelated networks the probability of finding a node of degree k' in the neighborhood of a node in degree class k is independent of k. In this case $\Theta_k = \Theta = \sum_{k'} (k'-1)\, P(k')\, i_{k'} / \langle k \rangle$ where the term $k'-1$ is due to the fact that at least one link of each infected node points to another infected vertex [15].

4.3.1.1 Epidemic Threshold

In order to derive the epidemic threshold let us consider the early time limit of the epidemic process. As done in Sect. 4.2.2.1 let us consider that at $t \sim t_0$ the

population is formed mostly by susceptible individuals. In the present scenario this implies $s_k \gg i_k$ and $r_k \sim 0\ \forall k$. The equation for the infected compartment then becomes $d_t i_k = \gamma k \Theta - \mu i_k$. Multiplying both sides for $P(k)$ and summing over all values of k we obtain $d_t i = \gamma \langle k \rangle \Theta - \mu i$. In order to understand the behavior of i around t_0 let us consider an equation built by multiplying both sides of the last equation by $(k-1)P(k)/\langle k \rangle$ and summing over all degree classes. We obtain $d_t \Theta = \gamma(\frac{\langle k^2 \rangle - \langle k \rangle}{\langle k \rangle})\Theta - \mu\Theta$. The fraction of infected individuals in each value of k will increase if and only if $d_t \Theta > 0$. This condition is verified when [15]:

$$R_0 \equiv \frac{\beta}{\mu} > \frac{\langle k \rangle^2}{\langle k^2 \rangle - \langle k \rangle} \tag{4.4}$$

giving us the epidemic threshold of an *SIR* process unfolding on an uncorrelated network.

Remarkably, due to their broad-tailed nature, real contact networks display fluctuations in the number of contacts (large $\langle k^2 \rangle$) that are significantly larger than the average degree $\langle k \rangle$ resulting in very small thresholds. Large degree nodes (hubs) facilitate an extremely efficient spreading of the infection by directly connecting many otherwise distant nodes. As soon as the hubs become infected diseases are able to reach a large fraction of the nodes in the network. Real interaction networks are extremely fragile to disease spreading. While this finding is somehow worrisome, it suggests very efficient strategies to control and mitigate the outbreaks. Indeed, hubs are *central* nodes and play a crucial role in the network connectivity [12] and by vaccinating a small fraction of them one is able to quickly stop the spread of the disease and protect the rest of the population. It is important to mention that in realistic settings the knowledge of the networks' structure is often limited. Hubs might not be easy to easily known and other indirect means must be employed. Interestingly, the same feature of hubs that facilitates the spread of the disease also allows for their easy detection. Since high degree nodes are connected to a large number of smaller degree nodes, one may simply randomly select a node, A, from the network and follow one of its links to reach another node, B. With high probability, node B has higher degree than A and is likely a hub. This effect became popularized as the *friend paradox*: on average your friends have more friends than you do [12]. Immunizing node B is then much more effective than immunizing node A. Remarkably, as counter-intuitive as this methodology might seem, it works extremely well even in the case of quickly changing networks [30–32].

4.4 Metapopulation Models

The next step in the progression towards more realistic modeling approaches is to consider the internal structure of the nodes. If each node in the network represents a homogeneously mixed sub-population instead of a single individual and we

consider the edges to represent interactions or mobility between the different sub-populations, then we are in the presence of what is known as meta-population. This concept was originally introduced by R. Levins in 1969 [33] for the study of geographically extended ecological populations where each node represents one of the ecological niches where a given population resides.

The metapopulation framework was later extended for use in epidemic modeling by Sattenspiel in 1987. In a landmark paper [34] Sattenspiel considered two different types of interactions between individuals, local ones occurring within a given node, and social ones connecting individuals originating from different locations on the network. This idea was later expanded by Sattenspiel and Dietz to include the effects of mobility [35] and thus laying the foundations for the development of epidemic models at the global scale.

Metapopulation epidemic models are extremely useful to describe particle reaction-diffusion models [36]. In this type of model each node is allowed to have zero or more individuals that are free to diffuse among the nodes constituting the network. In our analysis, as done in the previous section, we follow the HMF approach and consider all nodes of degree k to be statistically equivalent and write all quantities in terms of the degree k. To start, let us define the average number of individuals in a node of degree k to be $W_k = \frac{1}{N_k} \sum_i W_i \delta (k_i - k)$, where N_k is the number of nodes with degree k and the sum is taken over all nodes i. The mean field dynamical equation describing the variation of the average number of individuals in a node of degree k is then:

$$\frac{\mathrm{d}W_k(t)}{\mathrm{d}t} = -p_k W_k(t) + k \sum_{k'} P\left(k'|k\right) p_{k'k} W_{k'}(t) \tag{4.5}$$

where p_k and $p_{kk'}$ represent, respectively, the rate at which particles diffuse out of a node of degree k and diffuse from a node of degree k to one of degree k'.

With these definitions, the meaning of each term of this equation becomes intuitively clear: the negative term represents individuals leaving the node, while the positive term accounts for individuals originating from other nodes arriving at this particular class of node. The conditional probability $P(k'|k)$ encodes all the topological correlations of the network. By imposing that the total number of particles in the system remains constant, we obtain:

$$p_k = k \sum_{k'} P\left(k|k'\right) p_{kk'} \tag{4.6}$$

that simply states that the number of particles arriving at nodes of degree k' coming from nodes of degree k must be the same as the number of particles leaving nodes of degree k. The probabilities p_k and $p_{kk'}$ encode the details of the diffusion process [37]. In the simplest case, the rate of movement of individuals is independent of the degree of their origin $p_k = p$ for all values of the degree.

Furthermore, if individuals that are moving simply select homogeneously among all of their connections, then we have $p_{kk'} = p/k$. In this case, the diffusion process will reach a stationary state when:

$$W_k = \frac{k}{\langle k \rangle} \overline{W} \tag{4.7}$$

where $\overline{W} = W/N$, W is the total number of walkers in the system, and N the total number of nodes. The simple linear relation between W_k and k serves as a strong reminder of the importance of network topology. Nodes with higher degree will acquire larger populations of particles while nodes with smaller degrees will have proportionally smaller populations. However, even in the steady state, the diffusion process is ongoing, so individuals are continuously arriving and leaving any given node but are doing so in a way that maintains the total number of particles in each node constant.

In more realistic settings, the traffic of individuals between two nodes is function of their degree [37]:

$$p_{kk'} = w_0 \frac{(kk')^\theta}{T_k} \tag{4.8}$$

In this expression θ modulates the strength of the diffusion flow between degree classes (empirical values are in the range $-0.5 \leq \theta \leq 0.5$ [3]), where w_0 is a constant and $T_k = w_0 \langle k^{1+\theta} \rangle / \langle k \rangle$ is the proper normalization ensured by the condition in Eq. (4.6). In these settings, the diffusion process reaches a stationary state when:

$$W_k = \frac{k^{1+\theta}}{\langle k^{1+\theta} \rangle} \overline{W} \tag{4.9}$$

Note that for $\theta = 0$ this solution coincides with the case of homogeneous diffusion [Eq. (4.7)].

Combining this diffusion process with the (epidemic) reaction processes described above we finally obtain the full reaction-diffusion process. To do so we must simply write Eq. (4.5) for each state of the disease (e.g., Susceptible, Infectious, and Recovered for a simple *SIR* model) and couple the resulting equations using the already familiar epidemic equations. The full significance of Eq. (4.7) now becomes clear: nodes with higher degree have higher populations and are visited by more travelers, making them significantly more likely to also receive an infected individual that can act as the seed of a local epidemic.

In a metapopulation epidemic context we must then consider two separate thresholds, the basic reproductive ratio, R_0, that determines whether or not a disease can spread within one population (node) and a critical diffusion rate, p_c, that determines if individual mobility is sufficiently large to allow the disease to spread from one population to another. It is clear that if $p = 0$ particles are completely

unable to move from one population to another so the epidemic cannot spread across subpopulations and that if $p = 1$ all individuals are in constant motion and the disease will inevitably spread to every subpopulation on the network with a transition occurring at some critical value p_c.

In general, the critical value p_c cannot be calculated analytically using our approach as it depends non-trivially on the detailed structure of the network and the fluctuations of the diffusion rate of single individuals. However, in the case of uncorrelated networks a closed solution can be easily found for different mobility patterns. Indeed, in the case where the mobility is regulated by Eq. (4.8) we obtain:

$$p_c = \frac{1}{\overline{W}} \frac{\langle k^{1+\theta} \rangle^2}{\langle k^{2+2\theta} \rangle - \langle k^{1+2\theta} \rangle} \frac{\mu R_0^2}{2(R_0 - 1)^2} \tag{4.10}$$

Interestingly, the critical value of p is inversely proportional to the degree heterogeneity in the network, so that broad tailed networks have very low critical values. This simple fact explains why simply restricting travel between populations is a highly ineffective way to prevent the global spread of an epidemic.

The mobility patterns considered so far are so-called Markovian: individuals move without remembering where they have been nor they have a home where they return to after each trip. Although this is a rough approximation of individuals behavior, Markovian diffusion patterns are allowed to analytically describe the fundamental dynamical properties of many systems. Recently, new analytic results have been proposed for non-Markovian dynamics that include origin-destination matrices and realistic travel routes that follow shortest paths [38]. In particular, the threshold within such mobility schemes reads as:

$$p_c = \frac{1}{\overline{W}} \frac{\langle k^{\eta} \rangle}{\langle k^{1+\eta} \rangle} \frac{\langle k \rangle \mu R_0^2}{2 (R_0 - 1)^2} \tag{4.11}$$

The exponent η, typically close to 1.5 in heterogeneous networks, emerges from the shortest paths routing patterns [38]. Interestingly, for values of $\theta \leq 0.2$, fixing $\eta = 1.5$, p_c in the case of Markovian mobility patterns is larger than the critical value in a system subject to non-Markovian diffusion. The presence of origin-destination matrices and shortest paths mobility lower the threshold facilitating the global spreading of the disease. Instead, for values of $\theta > 0.2$ the contrary is true.

In these models the internal contacts rate is considered constant across each subpopulation. Interestingly, recent longitudinal studies on phone networks [39] and Twitter mention networks [40] point to the evidence that contacts instead scale super-linearly with the subpopulation sizes. Considering the heterogeneity in population sizes observed in real metapopulation networks, the scaling behavior entails deep consequence in the spreading dynamics. A recent study generalized the metapopulation framework considering such observations. Interestingly, the critical mobility thresholds, in the case of mobility patterns described by Eq. (4.8), changes significantly being lowered by such scaling features of human contacts [40].

Despite their simplicity, metapopulation models are extremely powerful tools in large scale study of epidemics. They easily lend themselves to large scale numerical stochastic simulations where the population and state of each node can be tracked and analyzed in great detail and multiple scenarios as well as interventions can be tested.

The state of the art in the class of metapopulation approaches is currently defined by the global epidemic and mobility model (*GLEAM*) [3, 4]. *GLEAM* integrates worldwide population estimates [41, 42] with complete airline transportation and commuting databases to create a world wide description of mobility around the world that can then be used as the substrate on which the epidemic can spread. *GLEAM* divides the globe into 3362 transportation basins. Each basin is defined empirically around an airport and the area of the basin is determined to be the region within which residents would likely use that airport for long distance travel. Each basin represents a major metropolitan area such as New York, London, or Paris. Information about all civilian flights can be obtained from the International Air Transportation Association (*IATA*) [43] and the Official Airline Guide (*OAG*) [44] that are responsible for compiling up-to-date databases of flight information that airlines use to plan their operations. By connecting the population basins with the direct flight information from these databases we obtain the network that acts as a substrate for the reaction diffusion process.

While most human mobility does not take place in the form of flights, the flight network provides the fundamental structure for long range travel that explains how diseases such as SARS [45], Smallpox [46], or Ebola [47] spread from country to country. To capture the finer details of within country mobility further information must be considered. GLEAM uses census information to create a commuting network at the basin level that connects neighboring metropolitan areas proportionally to the number of people who live in one are but work in the other.

Short-term short-distance mobility such as commuting is fundamentally different from medium-term long-distance airline travel. In one case, the typical timescale is work-day ($8h$) while in the other it is 1 day. This timescale difference is taken into account in *GLEAM* in an effective, mean-field, manner instead of explicitly through a reaction process such as the one described above. This added layer is the final piece of the puzzle that brings the whole together and allows *GLEAM* to describe accurately the spread from one country to the next but also the spread happening within a given country [48].

In Fig. 4.2 we illustrate the progression in terms of detail that we have undergone since our initial description of simple homogeneously mixed epidemic models in a single population. With all these ingredients in place we have a fine grained description of mobility on a world wide scale on top of which we can finally build an epidemic model.

Within each basin, *GLEAM* still uses the homogeneous mixing approximation. This assumption is particularly suited for diseases that spread easily from person to person through airborne means such as ILI. *GLEAM* describes influenza through an *SEIR* model as illustrated in Fig. 4.3. *SEIR* models are a modification of the *SIR* model described above that includes a further compartment, Exposed, to represent

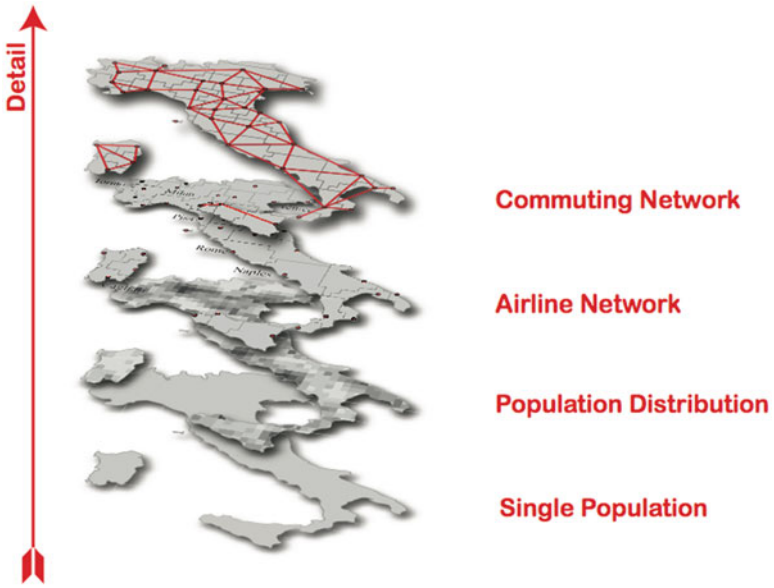

Fig. 4.2 The multilayer structure of *GLEAM*. Each layer increases the level of detail with respect to the previous ones

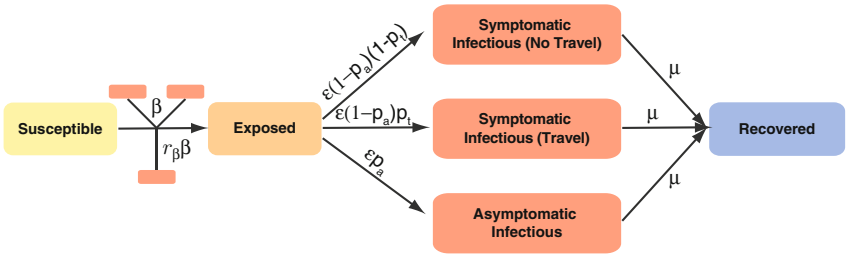

Fig. 4.3 SEIR Epidemic structure used in GLEAM

individuals in the incubation phase of the disease that are already infected but not yet Infectious. *GLEAM* further expands on this model by distinguishing three classes of Infectious individuals based on the severity of the disease. One third of the infectious individuals are asymptotic individuals do not display any symptoms and continue to behave normally while having an infectiousness reduced by a factor $r_\beta = 0.5$. Of the remaining symptomatic individuals, one half is sick enough to decide to not travel or commute while the remaining half continue to travel normally.

Despite their apparent complexity, large scale models such as *GLEAM* are controlled by just a small number of parameters and ultimately, it's the proper setting of these few parameters that is responsible for the proper calibration of the model and validity of the results obtained. Most of the disease and mobility parameters are

set directly from the literature or careful testing so that as little as possible remains unknown when it is time to apply it to a new outbreak.

GLEAM was put to the test during the 2009 H1N1 pandemic with great success. During the course of the epidemic, researchers were able to use official data as it was released by health authorities around the world. In the early days of the outbreak there was a great uncertainty about the correct value of the R_0 for the 2009/H1N1 pdm strain in circulation so a methodology to determine it had to be conceived.

One of the main advantages of epidemic metapopulation models is their computational tractability. It was this feature what proved invaluable when it came to determine the proper value of R_0. By plugging in a given set of parameters one is able to generate several hundreds or thousands of *in silico* outbreaks. Each outbreak contains information not only about the number of cases in each city or country as a function of time but also information about the time when the first case occurs within a given country. In general, each outbreak will be different due to stochasticity and by combining all outbreaks generated for a certain parameter set we can calculate the probability distribution of the arrival times. The number of times that an outbreak generated the seeding of a country, say the UK, in the same day as it occurred in reality provides us with a measure of how likely the parameter values used are. By multiplying this probability for all countries with a known arrival time we can determine the overall *Likelihood* of the simulation:

$$\mathscr{L} = \prod_c P_c\left(t_c\right) \tag{4.12}$$

where the product is taken over all countries c with known arrival time t_c and the probability distribution of arrival times, $P_c\left(t\right)$ is determined numerically for each set of input values. The set of parameters that maximizes this quantity is then the one whose values are the most likely to be correct. Using this procedure the team behind *GLEAM* determined that the mostly likely value of the basic reproductive ratio was $R_0 = 1.75$ [49], a value that was later confirmed by independent studies [50, 51].

Armed with an empirical estimate of the basic reproductive ratio for an ongoing pandemic, they then proceeded to use this value to estimate the future progression of the pandemic. Their results predicting that the full peak of the pandemic would hit in October and November 2009 were published in early September 2009 [49]. A comparison between these predictions and the official data published by the health authorities in each country would be published several years later [52] clearly confirming the validity of *GLEAM* for epidemic forecasting in real time. Indeed, the model predicted, months in advance, the correct peak week in 87 % of countries in the north hemisphere for which real data was accessible. In the rest of cases the maximum error reported has been 2 weeks. *GLEAM* can also be further extended to include age-structure [53], interventions and travel reductions.

4.5 Agent-Based Models

The next logical step in the hierarchy of large scale epidemic models is to take the description of the underlying population all the way down to the individual level with what are known as ABM. The fundamental idea behind this class of model is a deceptively simple one: treat each individual in the population separately, assigning it properties such as age, gender, workplace, residence, family structure, etc... These added details give them a clear edge in terms of detail over metapopulation models but do so at the cost of much higher computational cost.

The first step in building a model of this type is to generate a synthetic population that is statistically equivalent to the population we are interested in studying. Typically this is in a hierarchical way, first generating individual households, aggregating households into neighborhoods, neighborhoods into communities, and communities into the census tracts that constitute the country.

Generating synthetic households in a way that reproduces the census data is far from a trivial task. The exact details vary depending on the end goal of the model and the level of details desired but the household size, age, and gender of household members are determined stochastically from the empirically observed distributions and conditional probabilities. One might start by determining the size of the household by extracting from the distribution of household size of the country of interest and selecting the age and gender of the head of the household proportionally to the number of heads of households for that household size that are in each age group. Conditional on this synthetic individual we can then generate the remaining members, if any, of the household. The required conditional probability distributions and correlation tables can be easily generated [54] from high quality census data that can be found for most countries in the world. This process is repeated until enough synthetic households have been generated. Households are then aggregated into neighborhoods by selecting from the households according to the distribution of households in a specific neighborhood. Neighborhoods are similarly aggregated into communities and communities into census tracts.

Each increasing level of aggregation (from household to country) represents a decrease in the level of social contact, with the most intimate contacts occurring at the household level and least intimate ones at the census tract or country level. The next step is to assign to each individual a profession and work place. Workplaces are generated following a procedure similar to the generation of households and each employed individual is assigned a specific household. School age children are assigned a school. Working individuals are assigned to work places in a different community or census tract in a way that reflects empirical commuting patterns.

At this point, we have a fairly accurate description of where the entire population of a city or country lives and works. It is then not entirely surprising that this approach was first used to study in detail the demands imposed on the transportation

system of a large metropolitan city. *TRANSIMS*,[1] the TRansportation ANalysis and SIMulation System [55], used an approach similar to the one described above to generate a synthetic population for the city of Portland, in Oregon (OR) and coupled it with a route planner that would determine the actual route taken by each individual on her way to work or school as a way of modeling the daily toll on Portland's transportation infrastructure and the effect that disruptions or modification might have in the daily lives of its population.

EpiSims [5] was the logical extension of *TRANSIMS* to the epidemic world. *EpiSim*s used the *TRANSIMS* infrastructure to generate the contact network between individuals in Portland, OR. Susceptible individuals are able to acquire the infection whenever they are in a location along with one or more infectious individuals. In this way the researchers are capable of observing as the disease spreads through the population and evaluate the effect that measures such as contact tracing and mass vaccination.

More recent approaches have significantly simplified the mobility aspect of this kind of models and simply divide each 24 h period into day time and nighttime. Individuals are considered to be in contact with other members of their workplace during the day and with other household members during the night. In recent years, modelers have successfully expanded the large scale Agent Based approach to the country [6] and even continent level [56].

As the spatial scale of the models increased further modes of long-range transportation such as flights had to be considered. These are important to determine not only the seeding of the country under consideration through importation of cases from another country but also to connect distant regions in a more realistic way. *FluTE* [6] is currently the most realistic large scale Agent-Based epidemic model of the continental United States. It considers that international seeding occurs at random in the locations that host the 15 largest international airports in the US by, each day, randomly infecting in each location a number of individuals that is proportional to the international traffic of those airports.

FluTE is a refinement of a previous model [57] and it further refines the modeling of the infectious process by varying the infectiousness of an individual over time in the *SIR* model that they consider. At the time of infection each individual is assigned one of six experimentally obtained viral load histories. Each history prescribes the individuals viral load for each day of the infectious period and the infectiousness is considered to be proportional to the viral load. Individuals may remain asymptotic for up to 3 days after infection during which their infectiousness is reduced by 50 % with respect to the symptomatic period. The total infectious period is set to 6 days regardless of the length of the symptomatic period.

Given the complexity of the model the calibration of the disease parameters in order to obtain a given value of the basic reproductive ratio, R_0 requires some finesse. Chao et al. [6] uses the definition of R_0 to determine "experimentally" its value from the input parameters. It numerically simulates 1000 instances of

[1]The source code for *TRANSIMS* can be obtained from https://www.code.google.com/p/transims/.

the epidemic caused by a single individual within a 2000 person fully susceptible community for each possible age group of the seeding individual and use it to calculate the R_0^a of each age group a. The final R_0 is defined to the average of the various R_0^a weighted by age dependent attack rate [57]. The final result of this procedure is that the value of R_0 is given by:

$$R_0 = 5.592\lambda + 0.0068 \tag{4.13}$$

where λ is the infection probability per unit contact and is given as input. *FluTE* was a pioneer in the way it completely released its source code,[2] opening the doors of a new level of verifiability in this area. It has successfully used to study the spread of influenza viruses and analyze the effect of various interventions in the Los Angeles County [58] and United States country level [6].

4.6 Digital Epidemiology

The unprecedented amount of data on human dynamics made available by recent advances technology has allowed the development of realistic epidemic models able to capture and predict the unfolding of infectious disease at different geographical scales [59]. In the previous sections, we described briefly some successful examples that have been made possible thanks to high resolution data on where we live, how we live, and how we move. Data availability has started a second golden age in epidemic modeling [60].

All models are judged against surveillance data collected by health departments. Unfortunately, due to excessive costs, and other constraints their quality is far from ideal. For example, the influenza surveillance network in the USA, one of the most efficient systems in the world, is constituted of just 2900 providers that operate voluntarily. Surveillance data is imprecise, incomplete, characterized by large backlogs, delays in reporting times, and the result of very small sample sizes. Furthermore, the geographical coverage is not homogeneous across different regions, even within the same country. For these reasons the calibration and test of epidemic models with surveillance data induce strong limitations in the predictive capabilities of such tools. One of the most limiting issues is the geographical granularity of the data. In general, information are aggregated at the country or regional level. The lack of ground truth data at smaller scales does not allow a more precise selection and training of realistic epidemic models.

How can we lift such limitations? Data, data and more data is again the answer. At the end of 2013 almost 3 billion of people had access to the Internet while almost 7 billion are phone subscribers, around 20 % of which are actively using smartphones. The explosion of mobile usage boosted also the activity of social

[2] http://www.cs.unm.edu/~dlchao/flute/.

media platforms such as Facebook, Twitter, Google+ etc. that now count several hundred million active users that are happy to share not just their thoughts, but also their *GPS* coordinates. The incredible amount of information we create and access contain important epidemiologically relevant indicators. Users complaining about catching a cold before the weekend on Facebook or Twitter, searching for symptoms of particular diseases on search engines, or Wikipedia, canceling their dinner reservations on online platforms like OpenTable are just few examples. An intense research activity, across different disciplines, is clearly showing the potential, as well as the challenges and risks, of such digital traces for epidemiology [61]. We are at the dawn of the digital revolution in epidemiology [7, 8]. The new approach allows for the early detection of disease outbreaks [62], the real time monitoring of the evolution of a disease with an incredible geographical granularity [63–65], the access to health related behaviors, practices and sentiments at large scales [66, 67], inform data-driven epidemic models [68, 69], and development of statistical based models with prediction power [67, 70–78].

The search for epidemiological indicators in digital traces follows two methodologies: active and passive. In active data collection users are asked to share their health status using apps and web-based platforms [79]. Examples are *influenzanet* that is available in different European countries [64], and *Flu near you* in the USA [65] that engage tens of thousands of users that together provide the information necessary for the creation of interactive maps of ILI in almost real time. In passive data collection, instead, information about individuals health status is mined from other available sources that do not require the active participation of users. News articles [63], queries on search engines [74], posts on online social networks [67, 70–73], page view counts on Wikipedia [75, 76] or other online/offline behaviors [77, 78] are typical examples. In the following, we focus on the prototypical, and most famous, method of digital epidemiology, Google Flu Trends (GFT) [80], while considering also other approaches based on Twitter and Wikipedia data.

4.6.1 Social Media Based Epidemic Models

GFT is by far the most famous model in digital epidemiology. Launched in November 2008 together with a Nature paper [80] describing its methodology, it has continuously made predictions on the course of seasonal influenza in 29 countries around the world.[3] The method used by *GFT* is extremely simple. The percentage of *ILI* visits, a typical indicator used by surveillance systems to monitor the unfolding of the seasonal flu, is estimated with a linear model based on search engine queries. This approach is general, and used in many different fields of Science. A quantity of interest, in this case the percentage of ILI visits P, is estimated using a correlated

[3]Data available at http://www.google.org/flutrends.

signal, in this case the *ILI* related queries fraction Q, that acts as surrogate. The fit allows the estimate of P as a function of the value of Q:

$$\text{logit}\,(P) = \beta_0 + \beta_1 \text{logit}\,(Q) + \epsilon, \tag{4.14}$$

where $\text{logit}\,(x) = \ln\left(\frac{x}{1-x}\right)$, β_0 and β_1 are fitting parameters, and ϵ is an error term. As clear from the expression, the *GFT* is a simple linear fit, where the unknown parameters are determined considering historical data. The innovation of the system lies on the definition of Q that is evaluated using hundreds of billions of searches on Google. Indeed, *GFT* scans all the queries we submit to Google, without using information about users' identity, in search of those that *ILI* related. This is the paradigm of passive data collection in digital epidemiology. In the original model the authors measured the correlation of 50 millions search queries with historic CDC data, finding that 45 of them were enough to ensure the best correlation between the number of searches and the number of *ILI* cases. The identity of such terms has been kept secret in order to avoid changes in users' behavior. However, the authors provided a list of topics associated with each one of them: 11 were associated with influenza complications, 8 to cold/flu remedies, 5 to general terms for influenza, etc. Although the search for the terms has been performed without prior information, none of the most representative terms were unrelated to the disease. In these settings *GFT* showed a mean correlation of 0.97 with real data and was able to predict the surveillance value with 1–2 weeks ahead.

GFT is based on proprietary data that for many different constraints cannot be shared with the research community. Other data sources, different in nature, are instead easily accessible. Twitter and Wikipedia are the two examples. Indeed, both systems are available for download, with some limitations, through their respective *API*s.

The models based on Twitter are built within the same paradigm of *GFT* [67, 71–73, 81]. Tweets are mined in search of *ILI*-related tweets, or other health conditions such as insomnia, obesity, and other chronic diseases [67, 82], that are used to inform regression models. Such tweets are determined either as done in *GFT*, or through more involved methods based on support vector machine (*SVM*) or other machine learning methods that, provided an annotated corpus, find disease related tweets beyond simple keywords matches [67, 71–73, 81]. The presence of *GPS* information or other self-reported geographical data allows the models to probe different granularities ranging from countries [67, 71, 73, 81] to cities [72].

While models based on Twitter analyze users' posts, those based on Wikipedia focus on pages views [75, 76]. The basic intuition is that Wikipedia is used to learn more about a diseases or a medication. Plus, the website is so popular that is most likely one of the first results of search queries on most search engines. The methods proposed so far monitor a set of pages related to the disease under study. Examples are *Influenza*, *Cold*, *Fever*, *Dengue*, etc. Page views at the daily or weekly basis are then used a surrogates in linear fitting models. Interestingly, the correlation with surveillance data ranges from 0.02 in the case of Ebola to 0.99 in for *ILI*s [75, 76], and allows accurate predictions up to 2 weeks ahead. One important limitation of

Wikipedia based methods is the lack of geographical granularity. Indeed, the view counts are reported irrespective of readers' location but the language of the page can be used as a rough proxy for location. Such approximation might be extremely good for localized languages like Italian but it poses strong limitations in the case of global languages like English. Indeed, it is reported that 51 % of pages views for English pages are done in the USA, 11 % in the UK, and the rest in Australia, Canada and other countries [76]. Besides, without making further approximation such methods cannot provide indications at scales smaller than the country level.

Despite these impressive correlations, especially in the case of ILIs, much still remains to be done. *GFT* offers a particular clear example of the possible limitations of such tools. Indeed, despite the initial success, it completely failed to forecast the 2009 H1N1 pandemic [61, 83]. The model was updated in September 2009 to increase the number of terms to 160, including the 40 terms present in the original version. Nevertheless, *GFT* missed high 100 out of 108 weeks in the season 2011–2012. In 2013 *GFT* predicted a peak height more than double the actual value causing the underlying model to be modified again later that year.

What are the reasons underlying the limitations of *GFT* and other similar tools? By construction, *GFT* relies just on simple correlations causing it to detect not only the flu but also things that correlate strongly with the flu such as winter patterns. This is likely one of the reasons why the model was not able to capture the unfolding of an off-season pandemic such as the 2009 H1N1 pandemic. Also, changes in the Google search engine, that can inadvertently modify users' behavior, were not taken into account in *GFT*. This factor alone possibly explains the large overestimation of the peak height in 2013. Plus, simple auto-regressive models using just CDC data can perform as well or better than *GFT* [84]. The parable of *GFT* clearly shows both the potential and the risks of digital tools for epidemic predictions. The limitations of *GFT* can possibly affect all similar approaches based on digital passive data collection. In particular, the use of simple correlations measures does not guarantee the ability of capturing the phenomena across different scales in space and time with respect to those used in the training. Not to mention that correlations might be completely spurious. In a recent study for example, a linear model based on Twitter simply informed with the timeline of the term *zombie* was shown to be a good predictor of the seasonal flu [71].

Despite such observations the potential of these models is invaluable to probe data that cannot be predicted by simple auto-regressive models. For example, flu activity at high geographical granularities, although very important, is measured with great difficulties by the surveillance systems. *GFT* and other spatially resolved tools can effectively access to these local indicators, and provide precious estimates that can be used a complement for the surveillance and as input for generating epidemic models [49, 68].

4.7 Discussion

The field of epidemiology is currently undergoing a digital revolution due to the seemingly endless availability of data and computational power. Data on human behavior is allowing for the development of new tools and models while the commoditization of computer resources once available only for world leading research institutions is making highly detailed large scale numerical approaches feasible at last.

In this chapter, we present a brief review not only of the fundamental mathematical tools and concepts of epidemiology but also of some of the state-of-the-art and computational approaches aimed at describing, modeling, and forecasting the diffusion of viruses. Our focus was on the developments occurring over the past decade that are sure to form the foundation for developments in decades to come.

Acknowledgements BG was partially supported by the French ANR project HarMS-flu (ANR-12-MONU-0018).

References

1. Bernulli, D. (1760). Essai dune nouvelle analyse de la mortalité causée par la petite vérole et des advantages de l'inocoulation pur la prévenir. *Mémoires de Mathematique Physique de l'Academie Royale des Sciences, 8*, 1–45.
2. Anderson, R. M., & May, R. M. (1992). *Infectious diseases in humans*. Oxford: Oxford University Press.
3. Colizza, V., Barrat, A., Barthélemy, M., & Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences, 103*(7), 2015–2020.
4. Balcan, D., Gonçalves, B., Hu, H., Ramasco, J. J., Colizza, V., & Vespignani, A. (2010). Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of Computational Science, 1*(3), 132–145.
5. Eubank, S., Guclu, H., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., et al. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature, 429*(6988), 180–184.
6. Chao, D. L., Halloran, M. E., Obenchain, V. J., & Longini, I. M. (2010). Flute, a publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology, 6*(1), e1000656.
7. Brownstein, J. S., Freifeld, C. C., & Madoff, L. C. (2009). Digital disease detection: Harnessing the web for public health surveillance. *New England Journal of Medicine, 360*(21), 2153–2157. PMID:19423867.
8. Salathé, M., Bengtsson, L., Bodnar, J. T., Brewer, D. D., Brownstein, J. S., Buckee, C., et al. (2012). Digital epidemiology. *PLoS Computational Biology, 8*, 7.
9. Bailey, N. T. (1975). *The mathematical theory of infectious diseases*. London: Griffin.
10. Keeling, M. J. & Rohani, P. (2008). *Modeling infectious diseases in humans and animals*. Princeton: Princeton Univeristy Press.
11. Kermack, W., & McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Series A, Containing Papers of a Mathematical and Physical Character* (Vol. 115, pp. 700–721)

12. Newman, M. E. J. (2010). *Networks: An introduction*. Oxford: Oxford University Press.
13. Caldarelli, G. (2007). *Scale-free networks*. Oxford: Oxford University Press.
14. Cohen, R., & Havlin, S. (2010). *Complex networks: Structure, robustness and function*. Cambridge: Cambridge University Press.
15. Barrat, A., Barthélemy, M., & Vespignani, A. (2008). *Dynamical processes on complex networks*. Cambridge: Cambridge University Press.
16. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A. -L., Brewer, D., et al. (2009). Computational social science. *Science, 323*, 721.
17. Barabasi, A.-L. (2002). *Linked: How everything is connected to everything else and what it means.* Plume Editors.
18. Holme, P., & Saramäki, J. (2012). Temporal networks. *Physics Reports, 519*, 97.
19. Morris, M. (1993). Telling tails explain the discrepancy in sexual partner reports. *Nature, 365*, 437.
20. Rocha, L. E. C., Liljeros, F., & Holme, P. (2011). Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Computational Biology, 7*(3), e1001109.
21. Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J. J., & Van den Broeck, W. (2011). What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of Theoretical Biology, 271*, 166.
22. Perra, N., Gonçalves, B., Pastor-Satorras, R., & Vespignani, A. (2012). Activity driven modeling of dynamic networks. *Scientific Reports, 2*, 469.
23. Karsai, M., Perra, N., & Vespignani, A. (2014). Time varying networks and the weakness of strong ties. *Scientific Reports, 4*, 4001.
24. Sun, K., Baronchelli, A., & Perra, N. (2014). Epidemic spreading in non-markovian time-varying networks. arxiv:1404.1006.
25. Castellano, C., & Pastor-Satorras, R. (2010). Thresholds for epidemic spreading in networks. *Physical Review Letters, 105*, 218701.
26. Vespignani, A. (2012). Modeling dynamical processes in complex socio-technical systems. *Nature Physics, 8*, 32–30.
27. Boguna, M., & Pastor-Santorras, R. (2002). Epidemic spreading in correlated complex networks. *Physical Review E, 66*, 047104.
28. Newman, M. E. J. (2002). Spread of epidemic disease on networks. *Physical Review E, 66*, 016128.
29. Serrano, M. A., Boguna, M., & Pastor-Satorras, R. (2006). Correlations in weighted networks. *Physical Review E, 74*, 055101(R).
30. Cohen, R., Havlin, S., & ben Avraham, D. (2003). Efficient immunization strategies for computer networks and populations. *Physical Review Letters, 91*, 247901.
31. Garcia-Herranz, M., Egido, E. M., Cebrian, M., Christakis, N. A., & Fowler, J. H. (2014). Using friends as sensors to detect global-scale contagious outbreaks. *PLoS One, 9*, 4.
32. Liu, S., Perra, M., Karsai, N., & Vespignani, A. (2014). Controlling contagion processes in activity driven networks. *Physical Review Letters, 112*, 118702.
33. Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bulletin of the ESA, 15*(3), 237–240.
34. Sattenspiel, L. (1987). Population structure and the spread of disease. *Human Biology, 59*, 411–438.
35. Sattenspiel, L., & Dietz, K. (1995). A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences, 128*(1), 71–91.
36. Britton, N. F. (1986). *Reaction-diffusion equations and their applications to biology*. New York: Academic.
37. Colizza, V., & Vespignani, A. (2008). Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: theory and simulations. *Journal of Theoretical Biology, 251*(3), 450–467.
38. Meloni, S., Perra, N., Arenas, A., Gómez, S., Moreno, Y., & Vespignani, A. (2011). Modeling human mobility responses to the large-scale spreading of infectious diseases. *Scientific Reports, 1*, 62.

39. Schläpfer, M., Bettencourt, L. M., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., et al. (2014). The scaling of human interactions with city size. *Journal of the Royal Society Interface, 11*, 20130789
40. Tizzoni, M., Sun, K., Benusiglio, D., Karsai, M., & Perra, N. (2014). The scaling of human contacts in reaction-diffusion processes on heterogeneous metapopulation networks. arxiv:1411.7310.
41. Center for International Earth Science Information Network (ciesin), Columbia University, & Centro Internacional de Agricultura Tropical (ciat). (2004). *The gridded population of the world version 3 (gpwv3): Population grids*. Palisades, NY: Socioeconomic Data and Applications Center (sedac), Columbia University (2004).
42. Center for International Earth Science Information Network (ciesin), Columbia University; International Food Policy Research Institute (ifpri); The World Bank; & Centro Internacional de Agricultura Tropical (ciat). (2004). *Global rural-urban mapping project (grump), alpha version: Population grids*. Palisades, NY: Socioeconomic Data and Applications Center (sedac), Columbia University (2004).
43. International Air Transport Association. http://www.iata.org
44. Official Airline Guide. www.oag.com/
45. Colizza, V., Barrat, A., Barthélemy, M., & Vespignani, A. (2007). Predictability and epidemic pathways in global outbreaks of infectious diseases: The sars case study. *BMC Medicine, 5*(1), 34.
46. Gonçalves, B., Balcan, D., & Vespignani, A. (2013). Human mobility and the worldwide impact of intentional localized highly pathogenic virus release. *Scientific Reports, 3*, 810.
47. Gomes, M. F. C., Pastore y Piontti, A., Rossi, L., Chao, D., Longini, I., Halloran, M. E., et al. (2014, September 2). Assessing the international spreading risk associated with the 2014 west African Ebola Outbreak. *PLOS Currents Outbreaks* (1st ed.). doi: 10.1371/currents.outbreaks.cd818f63d40e24aef769dda7df9e0da5.
48. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Science, 106*(51), 21484–21489.
49. Balcan, D., Hu, H., Gonçalves, B., Bajardi, P., Poletto, C., Ramasco, J. J., et al. (2009). Seasonal transmission potential and activity peaks of the new influenza a (h1n1): A monte carlo likelihood analysis based on human mobility. *BMC Medicine, 7*(1), 45.
50. Yang, Y., Sugimoto, J. D., Halloran, M. E., Basta, N. E., Chao, D. L., Matrajt, L., et al. (2009). The transmissibility and control of pandemic influenza a (h1n1) virus. *Science, 326*(5953), 729–733.
51. Fraser, C., Donnelly, C. A., Cauchemez, S., Hanage, W. P., Van Kerkhove, M. D., Hollingsworth, T. D., et al. (2009). Pandemic potential of a strain of influenza a (h1n1): Early findings. *Science, 324*(5934), 1557–1561.
52. Tizzoni, M., Bajardi, P., Poletto, C., Ramasco, J. J., Balcan, D., Gonçalves, B., et al. (2012). Real-time numerical forecast of global epidemic spreading: Case study of 2009 a/h1n1pdm. *BMC Medicine, 10*(1), 165.
53. Ajelli, M., Gonçalves, B., Balcan, D., Colizza, V., Hu, H., Ramasco, J. J., et al. (2010). Comparing large-scale computational approaches to epidemic modeling: agent-based versus structured metapopulation models. *BMC Infectious Diseases, 10*(1), 190.
54. Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice, 30*(6), 415–429.
55. Barrett, C. L., Beckman, R. J., Berkbigler, K. P., Bisset, K. R., Bush, B. W., Eubank, S., Hurford, J. M., Konjevod, G., Kubicek, D. A., Marathe, M. V., et al. (1999). Transims (transportation analysis simulation system). In *Volume 0: Overview. Report LA-UR-99-1658*. Los Alamos, NM: Los Alamos National Laboratory.
56. Merler, S., Ajelli, M., Pugliese, A., & Ferguson, N. M. (2011). Determinants of the spatiotemporal dynamics of the 2009 h1n1 pandemic in europe: Implications for real-time modelling. *PLOS Computational Biology, 7*(9), e1002205.

57. Germann, T. C., Kadau, K., Longini, I. M., & Macken, C. A. (2006). Mitigation strategies for pandemic influenza in the united states. *Proceedings of the National Academy of Science, 103*(15), 5935–5940.
58. Chao, D. L., Matrajt, L., Basta, N. E., Sugimoto, J. D., Dean, B., Bagwell, D. A., et al. (2011). Planning for the control of pandemic influenza a (h1n1) in los angeles county and the united states. *American Journal of Epidemiology, 173*(10), 1121–1130.
59. Vespignani, A. (2009). Predicting the behavior of techno-social systems. *Science, 325*(5939), 425.
60. Pastor-Satorras, R., Castellano, C., Van Mieghem, P., & Vespignani, A. (2014). Epidemic processes in complex networks. arXiv preprint. arXiv:1408.2701.
61. Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of google flu: Traps in big data analysis. *Science, 343*(6176), 1203–1205.
62. Chan, E. H., Brewer, T. F., Madoff, L. C., Pollack, M. P., Sonricker, A. L., Keller, M., et al. (2010). Global capacity for emerging infectious disease detection. *Proceedings of the National Academy of Sciences, 107*(50), 21701–21706.
63. Health Map. http://www.healthmap.org/.
64. Paolotti, D., Carnahan, A., Colizza, V., Eames, K., Edmunds, J., Gomes, G., et al. (2014). Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience. *Clinical Microbiology and Infection*, 17–21.
65. Flu Near You. http://www.flunearyou.org.
66. Salathé, M., & Khandelwal, S. (2011). Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Computational Biology, 7*(10), e1002199.
67. Paul, M. J. & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. In *ICWSM* (pp. 265–272).
68. Shaman, J., & Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences, 109*(50), 20425–20430.
69. Zhang, Q., Perra, N., & Vespignani, A. (in preparation). Forecasting seasonal influenza with stochastic microsimulations models assimilating digital surveillance data.
70. Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS One, 6*(5), e19467.
71. Bodnar, T., & Salathé, M. (2013). Validating models for disease detection using twitter. In *Proceedings of the 22nd International Conference on World Wide Web Companion* (pp. 699–702). International World Wide Web Conferences Steering Committee, 2013.
72. Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through twitter: an analysis of the 2012–2013 influenza epidemic. *PloS One, 8*(12), e83672.
73. Culotta, A. (2010). Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics* (pp. 115–122), ACM.
74. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014
75. Generous, N., Fairchild, G., Deshpande, A., Del Valle, S. Y., & Priedhorsky, R. (2014). Detecting epidemics using wikipedia article views: A demonstration of feasibility with language as location proxy. arXiv preprint. arXiv:1405.3612.
76. McIver, D. J., & Brownstein, J. S. (2014). Wikipedia usage estimates prevalence of influenza-like illness in the united states in near real-time. *PLoS Computational Biology, 10*(4), e1003581.
77. Nsoesie, E. O., Buckeridge, D. L., & Brownstein, J. S. (2014). Guess who is not coming to dinner? Evaluating online restaurant reservations for disease surveillance. *Journal of Medical Internet Research, 16*(1), e22.
78. Butler, P., Ramakrishnan, N., Nsoesie, E. O., & Brownstein, J. S. (2014). Satellite imagery analysis: What can hospital parking lots tell us about a disease outbreak? *Computer, 47*(4), 94–97.

79. Wójcik, O. P., Brownstein, J. S., Chunara, R., & Johansson, M. A. (2014). Public health for the people: participatory infectious disease surveillance in the digital age. *Emerging Themes in Epidemiology, 11*(1), 7.
80. Google Flu Trends. http://www.google.org/flutrends/.
81. Signorini, A., Polgreen, P. M., & Segre, A. M. (2010). Using twitter to estimate h1n1 influenza activity. In *9th Annual Conference of the International Society for Disease Surveillance*.
82. De la Torre-Díez, I., Díaz-Pernas, F. J., & Antón-Rodríguez, M. (2012). A content analysis of chronic diseases social groups on facebook and twitter. *Telemedicine and e-Health, 18*(6), 404–408.
83. Olson, D. R., Konty, K. J., Paladini, M., Viboud, C., & Simonsen, L. (2013). Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Computational Biology, 9*(10), e1003256
84. Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences, 107*(41), 17486–17490.