Global Connectivity and the Spread of Infectious Diseases

Dirk Brockmann (Berlin)[1]

Abstract

The spread of infectious diseases has become a global concern. In light of the recent outbreaks of Zika in South America and MERS in the Middle East as well as the 2013 Ebola crisis, researchers are developing a range of methods and strategies to mitigate disease spread. One of the key challenges is to understand the key features that shape patterns of global disease spread. The complexity and redundancy of global transport networks suggests that the systematic identification of hidden patterns in spatially incoherent disease dynamics is next to impossible. Here, we will discuss how the concept of effective distance, as a replacement for conventional geographical distance, helps us understand global disease dynamics and how it can be employed as a new technique for developing predictive tools and means for testing effective containment strategies.

1. Tools for Predicting the Dynamics of Human Infectious Disease

'Can we control the world?' is not a humble question. Any person with common sense would immediately answer with a 'no'. Being humble is undoubtedly quite important when talking about infectious diseases and about trying to tackle or understand their spread. On the one hand, infectious diseases are natural phenomena. But especially human infectious diseases are also social phenomena. And the way we move across the globe plays an important role in the spread of these diseases.

In 2009, an outbreak of the influence virus H1N1 started in Mexico and then spread across the world. There was the MERS coronavirus in Saudi Arabia and, of course, the Ebola crisis. And now there is Zika. The gut feeling of most people is that these outbreaks are happening at an increased rate. This is also reflected by the fact that Hollywood is producing more movies about the spread of diseases, blockbusters such as *Outbreak* or *Contagion*. Their plot: there is a huge pandemic and everyone panics, and finally a hero saves the world or fails to do so. The real story is more complex and more interesting.

2. The True Story of Infectious Disease Spread

The human population is increasing and half of mankind now lives in urban areas. At the same time, global mobility is at an all-time high. Together, these developments are bad for humans and good for pathogens. Then again, there is a kind of revolution happening in epidemiology. As Dr. Lengauer showed us, a lot of data is available today, data we can use to

---

1 Humboldt University of Berlin; Robert Koch Institute.

investigate the properties of infectious diseases. Beyond that, a global network of molecular surveillance is now at our disposal; viral or bacterial genomes are routinely sequenced and this micro-evolutionary information is used to model transmission pathways.

3. Networks of Epidemiology

One type of data that is pivotal in reconstructing and predicting pathways of transmission are host mobility networks. Figure 1 depicts a global host mobility network, the worldwide air transportation network. Each node represents a location (airport) in space. Links between these nodes are connections between these airports. In contact pattern networks, nodes are individuals and links are contacts and thus potential transmission pathways between individuals, quantifying, for instance, how much time infected people spend together.
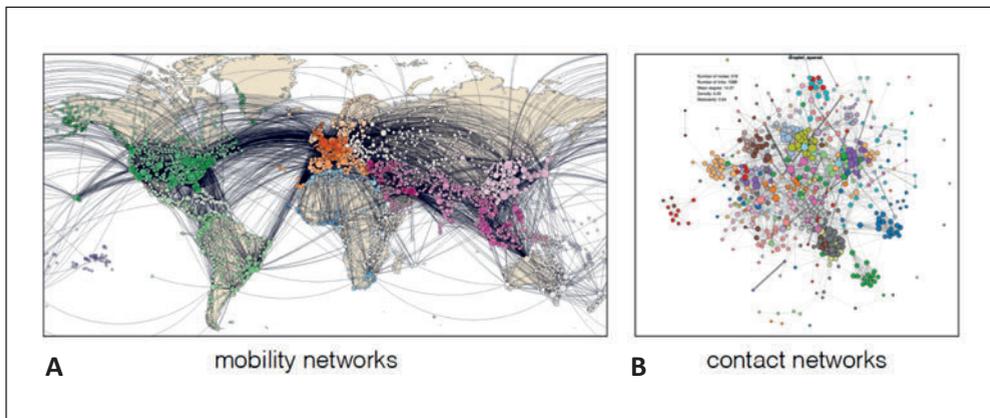

Fig. 1 (*A*): The worldwide air transportation network. This network has approx. 4,000 airports and 25,000 direction connections. More than three billion passengers travel on this network each year. Every day, all passengers travel more than 15 billion km in total; that is three times the radius of our solar system. (*B*): A contact pattern network between 1,000 students. Contacts are reconstructed by measuring individuals' proximity to each other using cell-phone Bluetooth information. Strong links reflect pairs of individuals that spend a lot of time together.

Mobility networks like the worldwide air transportation network are very important in studying global disease dynamics. Contact pattern networks exist on a much finer scale and help us understand transmission dynamics in groups of individuals.

A few decades ago, researchers had to make a lot of assumptions about the rate at which individuals interact. Quantitative experiments had not yet been conducted. Today, however, we have means of reconstructing individual face-to-face interactions. For example, a few years ago, a colleague of mine, Sune LEHMANN of the Technical University of Denmark (DTU), purchased 1,000 cell phones and distributed them among a random group of DTU students. This was the deal: the students got a free cell phone and in return the cell phones monitored everything the students did with high temporal resolution over the course of months or years. The phones provided information about how the students moved around in Copenhagen and how much they interacted physically and on social media through texting,

etc. With this data, the researchers could investigate whether social media interactions were predictive for real world interactions. I started a collaboration with LEHMANN on a particular aspect of their data: the amount of time any two students spent together over the course of a longer period, say three months.

Some students spent a lot of time in a group of people, while others spent a lot of time with only one other person: they were dating, so they spent most of the nights together, which is a substantial fraction of a 24-hour cycle. The data set revealed other interesting phenomena, like strong links breaking or strong links forming. There even were some triplets in the data! We can use this kind of technology to quantify important factors of disease transmission: for instance, how much time any two people spend together on average. For the DTU students, it was about 3.6 minutes on average. But the distribution is very broad: some people spent ten hours a day together and others only seconds.

4. Digital Epidemiology

Experiments like the Danish cell phone study and similar projects kicked off a whole new field called digital epidemiology. Today, more and more of this data is collected on the web. Statistical models use these growing big data sets to extract information about the dynamics of human infectious diseases. In addition to mobility and contact patterns, we have genetic information. Powerful algorithms can exploit this data to derive statistical inferences about the near future, similar to what is happening in meteorology.

This approach is very powerful, but it is also risky, especially because large datasets, powerful computers, and sophisticated statistical methods invite researchers to treat data as a black box. Different, parameter-rich methods applied to the same data can yield contradictory results. Recently, an interesting paper was published in which researchers gave a data set to different teams of scientists, 29 teams in total. They were given a data set on soccer referee decisions and were asked to determine whether the referees were more likely to hand red cards to players with dark skin than to those with light skin. Twenty teams – 69 % – found a statistically significant positive effect and nine teams – 31 % – observed a non-significant effect.

In addition, correlation is not causation. Just because observation A in a data set correlates strongly with observation B does not mean that A causes B, nor that B causes A, nor that both are caused by a third effect. The book *Spurious Correlations*, published by Harvard student Tyler VIGEN, beautifully illustrates this limitation of retrospective, big-data-driven research. For example, there is a very strong correlation between the number of sociology doctorates in the US and worldwide non-commercial space launches. The human brain spots correlations all the time. It is a powerful tool for finding patterns in data, but we often cannot know if these patterns are meaningful if we do not understand the underlying mechanism, as Dr. LENGAUER pointed out earlier.

5. Mobility and its Effect on Disease Spread

So, in trying to understand global disease dynamics, we try to keep our questions simple. For instance, we would like to know where a pathogen came from and when and where it will likely appear next after an outbreak. In an ideal situation, a disease spreads like a wave at a

constant speed, making a calculation of subsequently affected areas easy. During the bubonic plague, this assumption of constant speed actually corresponded to reality because most people did not travel very far. But today, that is no longer the case, as illustrated in Figure 1 on the left. Each day, humans are travelling about 14 billion kilometres on planes alone, three times the radius of our solar system. In just a day, you can go from anywhere on the globe to anywhere else. What does this long range mobility mean for spreading phenomena? Computer simulations show that today, because of long-range traffic, an initial outbreak can yield new secondary outbreaks far away. Geographic distance to the original outbreak location correlates no longer well with arrival time, and the spatial pattern of disease spread is spatially incoherent. Regular wave patterns no longer exist. Educated guesses about when the disease is going to hit a particular location become next to impossible. We can actually see this in real data: Figure 2 shows a simulation of outbreaks based on today's mobility network and two real recent events, H1N1 in 2009 and SARS in 2003.
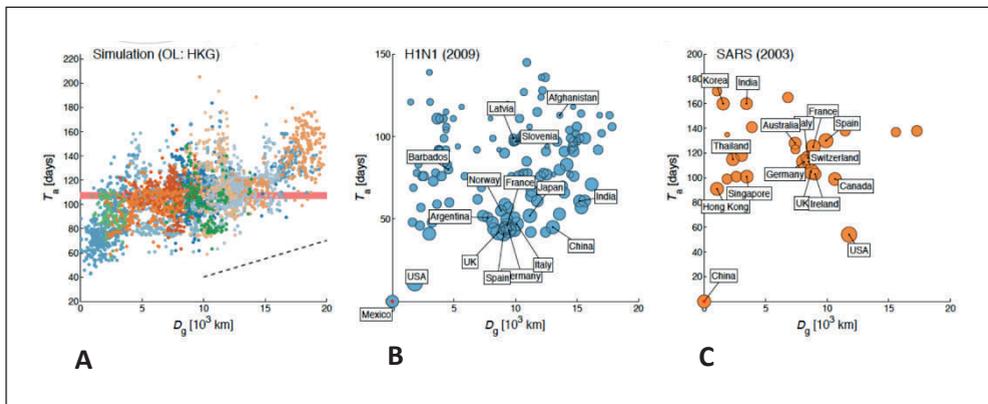

Fig. 2 (*A*): A computer simulation for a virus like H1N1 with a hypothetical outbreak in Hong Kong. (*B*): Actual data from the 2009 H1N1 outbreak. (*C*): Actual data from the 2003 SARS outbreak. D_g: distance of a particular location to the origin of the outbreak, T_a: time of arrival at that particular location.

The simulation still shows a correlation of distance and speed of disease spread of about 250 to 400 kilometres per day. This is about 100 times faster than the Black Death in the 14th century. However, both in the simulation and during the actual events, mere geographic distance is not a predictor of arrival time. That is why we have to develop mechanistic mathematical models and algorithms that take into account global connectivity and have to rely on computer simulations to make predictions on the spatio-temporal pattern of modern, global disease dynamics.

6. Mobility-Based Models of Disease Spread

Computational, predictive models for global disease dynamics are similar to weather forecast systems. Researchers are putting a lot of effort into making them more precise and reliable. There is even a tool for making global disease dynamics predictions, called "GLEaM", the

Global Epidemic and Mobility Simulation Tool. It is very sophisticated and integrates a lot of information and data. You can choose from thousands of origins, specify the number of available hospital beds in that location, the global mobility, the commuter traffic, and much more. The hope with models like these is that the more data they are fed, the more precise their predictions will get. This is a major direction of research in this area. And even though they have become quite good at making short term predictions, there are several issues with these mechanistic models, and I remain sceptical for many reasons that an emphasis on computational detail is by itself a promising path to take. One of them is that dynamical systems at the core of computer simulations require the correct set of parameters (including their values) and initial conditions. But in situations like the Ebola, Chikungunya, MERS or other emer-
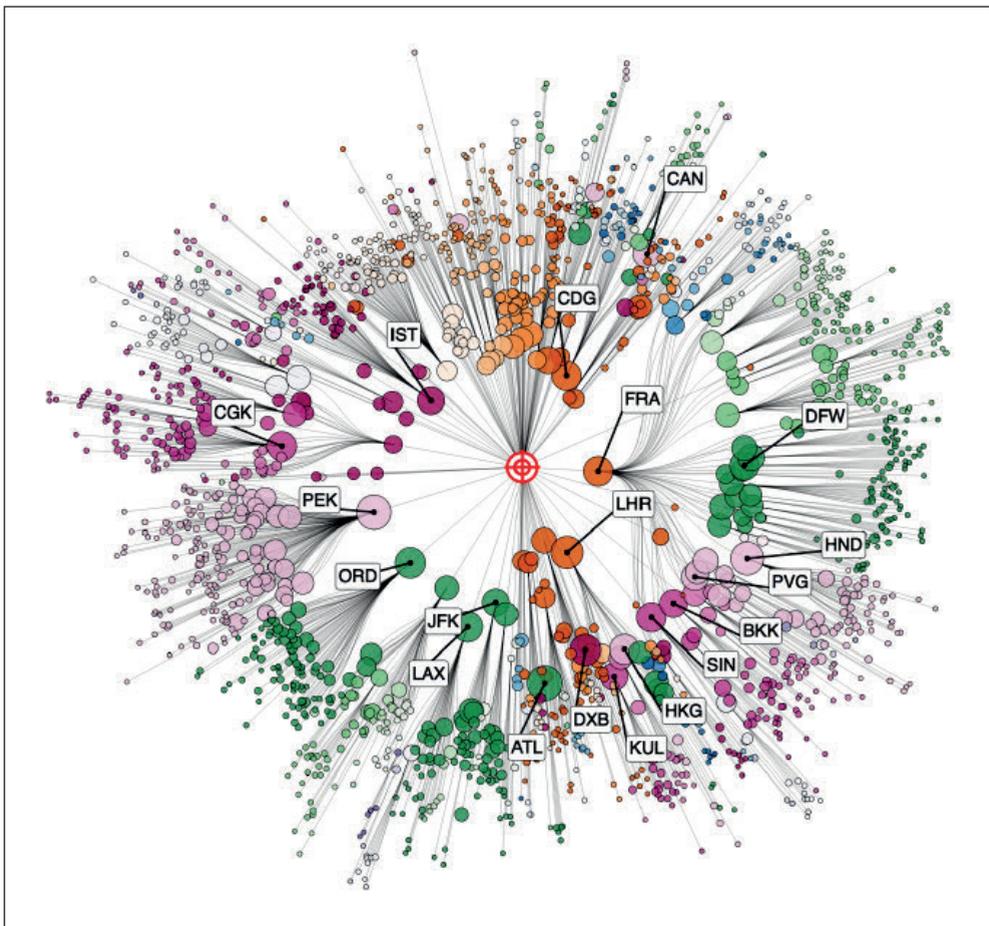

Fig. 3 The network depicts the perspective of the Berlin airport Tegel (TXL, central red node) using the concept of effective distance. All airports of the worldwide air transportation network are arranged in a circle, the distance to the centre is the effective distance from TXL to the respective airport. The tree represents the most probable spreading route from a hypothetical outbreak in Berlin. A number of airports that may be geographically distant, for example Beijing (PEK), are effectively close to TXL and a gateway to many other airports.

gent epidemics, the initial conditions and the disease parameters are exactly the unknown factors. So, in situations where we actually need these models, we cannot use them reliably.

7. Redefining Distance

Recently, we started playing with one of the parameters of our statistical models: distance. Looking at the worldwide air transportation network, we realised that from the point of view of a pathogen, cities like London or Frankfurt are much closer to New York than many small American towns, simply because there is more traffic connecting both places. You can redefine distance in this manner for any two locations on the planet to restructure the air traffic map. We developed a mathematical theory that accounts for this fact and introduced a new distance measure that is small between places that are strongly connected by traffic and large for places that are connected by small passenger flux. When this theory is applied and visualised, radial distances become what we call effective distances (see Fig. 3). Now you can look at places that are relevant for disease dynamics, like Freetown in Sierra Leone, where the Ebola outbreak happened. London Heathrow, for instance, is very close to Freetown effectively. Beijing is not far from Sierra Leone either, as connections go through Heathrow. This way, we can find the most effective spreading routes and identify the nodes in the network that are most effective for spreading diseases.

Effective distance maps are more than just an illustration of how the world looks from a different angle. Figure 2 shows the result of a computer simulation for the two pandemics already (SARS and H1N1) discussed above.

In Figure 4*B*, the spread creates a concentric pattern on the world map of effective distance. Now we can measure the speed of the disease wave and predict when the epidemic is going to hit other locations. This way, we get a much higher predictive power.

8. Understanding Disease Transmission

The speed at which diseases travel through populations depends not only on effective distance between locations, but also on the way the disease is transmitted between individuals in those locations. Traditional epidemiology usually tries to extract some limited information about transmission dynamics from data recorded during recurrent epidemics or endemic diseases. In contrast, it would be helpful if one could inject a pathogen into the population, record how it spreads, and then repeat that experiment 1,000 times in order to develop a correct theory – but of course we cannot do that. Nevertheless, with good contact pattern data we can approximate this experimental scenario.

Vaccination is a very powerful weapon against some infectious diseases. Today, a number of eradication programmes are in place, for instance against polio and measles. A very simple question in this context is: what fraction of the population must be vaccinated in order to eradicate a disease? It has to be a number between zero and 100%. What that number is depends on the properties of the disease itself. Another number is crucial in finding the critical vaccination threshold for eradication. It is called the basic reproduction number or R_0. It is the average number of secondary infections caused by an infected individual and describes at which exponential rate a disease can spread through a susceptible population. For influenza
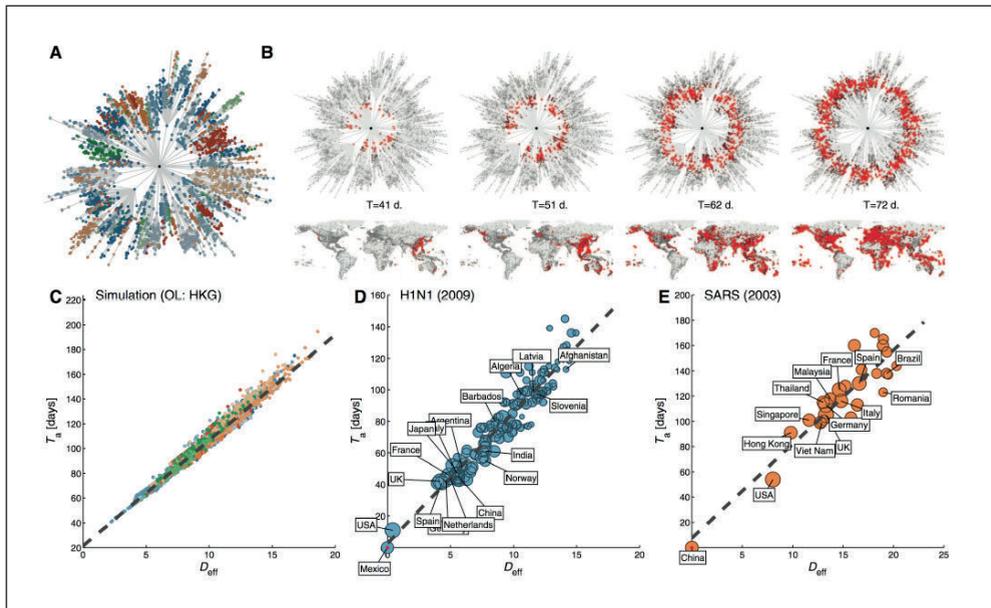
Fig. 4 An illustration of the usefulness of effective distance in a hypothetical outbreak scenario, the same as in Figure 2 with an initial outbreak location in Hong Kong. (*A*): Depicts the effective distance perspective of Hong Kong onto the rest of the world. (*B*): A computer-simulated pandemic in the effective distance representation compared to the traditional visualisation. With the effective distance approach, complex spatio-temporal patterns are mapped onto generic concentric wave fronts that are much easier to quantify, understand, and employ in pattern based predictions. Panels (*C*), (*D*) and (*E*) depict the same information as Fig. 2, the only difference is that geographic distance is replaced by effective distance, which is a much better predictor for epidemic arrival times.
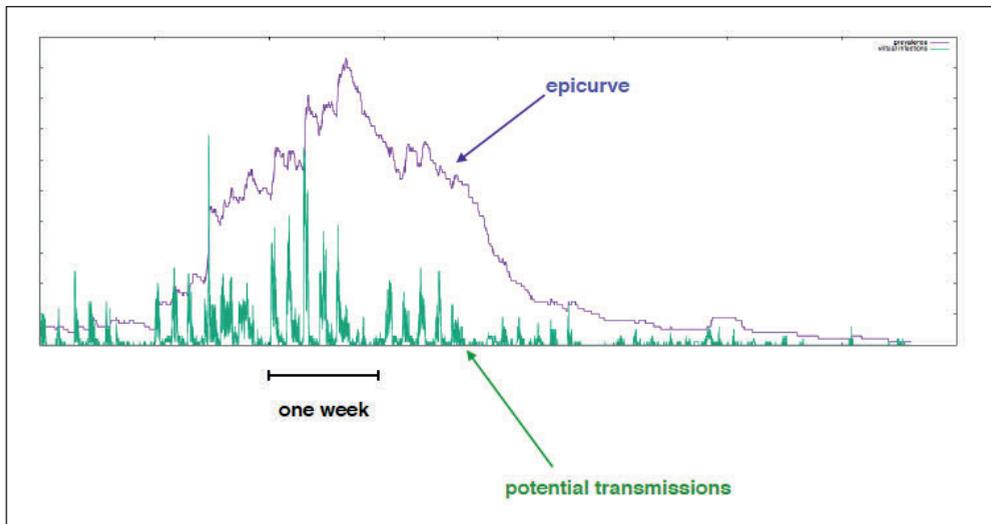

Fig. 5 A simulated epidemic on student interaction data. *Violet*: Prevalence data from classical infection models based on basic reproduction values. *Green*: Actual contact data.

it is maybe 1.4 to four. For measles, the basic reproduction number is very high: 12 to 18. In a susceptible population, measles will explode. Epidemiologists have been using an equation to estimate from R_0 the critical vaccination threshold: $1 - 1/R_0$. For instance, if $R_0=2$, 50 % would have to be vaccinated. For measles, you get a number between 91 and 96 % because its R_0 is so large.

The equation for the critical vaccination threshold comes from a very simple mechanistic model that rests on a couple of very crude assumptions: all individuals behave the same way and are identical, any pair of individuals is as likely to interact as any other pair, fluctuations do not matter, and the system is in equilibrium. However real contact patterns between people are more diverse: they show strong temporal modulations, such a circadian or weekly rhythms. Groups form and individuals may leave a group and join another. In a nutshell, everyone does not interact with everyone else in the same way.

So instead of using a crude theory to derive critical vaccination thresholds, we ran virtual Bluetooth epidemics on the Copenhagen student interaction data. Figure 5 shows an epicurve that epidemiologists usually look at and the actual transmission rates in the Copenhagen mobile phone epidemic.

The actual contact patterns are very jagged. And it is these contacts during which infected individuals and susceptible ones meet. Usually, this kind of data is not accessible in traditional epidemiology. Instead, prevalence models are relatively smooth curves. However, critical vaccination thresholds depend on the particular patterns of interaction.

9. Virtual Vaccination Programmes Find Real Critical Vaccination Thresholds

We ran thousands of these virtual epidemics and virtually vaccinated individuals, measured the critical vaccination threshold in the population as a function of disease parameters, and found that not only R_0 matters, but also the infectious period. The simulations show that hypothetical critical vaccination thresholds overestimate those actually required for disease eradication – at least in a population of Danish students. As a next step, we aim to find more representative populations to apply this novel method to more realistic scenarios.

In summary, the spread of infectious diseases around the globe depends on both global mobility and the transmission dynamics within populations. We could show that using effective distance as a model parameter instead of geographic distance increases predictive power considerably. Novel experimental methods, such as mobile phone interactions, now provide a type of data previously inaccessible, allowing a more accurate estimation of host-to-host interaction. Together, these novel approaches can increase the ability of data-based statistical modelling to predict the most likely paths of disease spread in our highly interconnected world.