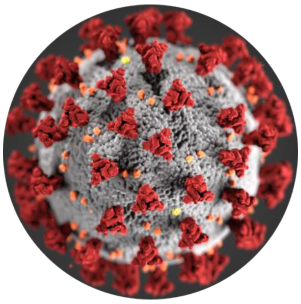
Pandemic Modelling

Abram Schönfeldt

March 2020



[2]

Introduction

This goal of this *mini-project* is to understand how the tools we have been learning in Stochastic Processes and Non-Equilibrium Systems have been applied to pandemics, which are rather topical as of late.

1. Literature Review

1.1. Forecast and control of epidemics in a globalized world [3]

The first paper we consider comes just after the SARS outbreak of the early 2000s and illustrates the importance and implications of considering fluctuations in disease modelling. It also considers the connected nature of our modern world by incorporating disease dispersal via the aviation network into its model.

Hufnagel et al. start by reviewing the deterministic *Systematic Inflammatory Response* (SIR) model. This models the number of *'Susceptible (S), Infected (I), and Recov-*

ered (R)' entities in a population at a time t and takes the form:

$$\frac{ds}{dt} = -\alpha sj$$
$$\frac{dj}{dt} = \alpha sj - \beta j$$

Entities in this case are people, but I imagine the SIR model has applications outside of our kind

where $j = I/N$, $s = S/N$ are the number of infecteds and susceptibles relative to the population size (N), and the relative number of recovered individuals can be obtained from $r(t) = 1 - j(t) - s(t)$.

The important quantities are the *average infectious period* $\tau = \beta^{-1}$ (self-explanatory) and the *basic reproduction number* $\rho_0 = \alpha/\beta$, which tells us the average number of susceptibles an infected individual infects in an otherwise uninfected population.

The paper argues that disease transmission and recovery are inherently stochastic processes, and thus a probabilistic description is more appropriate than the deterministic SIR model. At each time step, the probability $p(S, I; t)$ of finding S susceptibles and I infecteds in a population of size N is governed by the master equation:

$$\begin{aligned}\partial_t p(S, I; t) = \frac{\alpha}{N}(S+1)(I-1)p(S+1, I-1; t) \\ +\beta(I+1)p(S, I+1; t) \\ -(\frac{\alpha}{N}SI - \beta I)p(S, I; t)\end{aligned}$$

with the initial condition $p(S, I; t - t_0) = \delta(I - I_0)\delta(S - (N - I_0))$. In the limit $N \gg 1$, the master equation can be approximated by a Fokker-Planck equation with associated stochastic Langevin equations:

$$\frac{ds}{dt} = -\alpha sj + \frac{1}{\sqrt{N}}\sqrt{\alpha sj}\xi_1(t)$$
$$\frac{dj}{dt} = \alpha sj - \beta j - \frac{1}{\sqrt{N}}\sqrt{\alpha sj}\xi_1(t) + \frac{1}{\sqrt{N}}\sqrt{\beta j}\xi_2(t)$$

where ξ_1, ξ_2 are independent Gaussian white noise forces which reflect the fluctuations in transmission and recovery.

An adapted version of this 'stochastic' SIR model is then used to model the Local dynamics of the disease. The adapted stochastic model also considers people who have been infected but who are not yet infectious. These people are incorporated into the model as *Latent* (L) individuals. The adapted model also does not use a fixed basic reproduction number ρ_0. Instead, $\rho_0(t)$ is a function of time.

Dispersal via the aviation network is modelled using the *probability rate matrix* $\gamma_{ij} = w_{ij}/\tau_j$, which divides weights w_{ij} describing the relative probability of travelling from an airport in region i to one in region j, by the typical amount of time an individual remains in region j, denoted τ_j. The weights are computed as $w_{ij} = M_{ij}/\sum_i M_{ij}$, where M_{ij} is the number of passengers per unit time departing from an airport in region i, arriving at an aiport in region j.

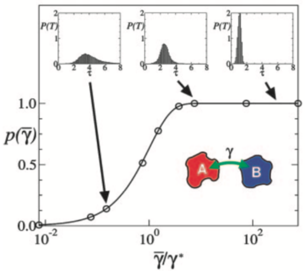
In the paper, the time periods $\tau_j = N_j / \sum_i M_{ij}$ are assumed to be a constant global rate γ. This assumption is made on the basis of the capacity of each airport j, $\sum_i M_{ij}$, reflecting the need of the population surrounding each airport, N_j, i.e. $N_j \propto \sum_i M_{ij}$.

Is this a justified assumption to make?

The resulting simulations of this model rather accurately model what happened in reality with the SARS outbreak. The order of magnitude of the predicted number of infecteds in each country matched those of the actual number of infecteds in all countries except for Japan and Canada. The model also predicted a small number of cases (≤ 10) in the Netherlands and in Bangladesh, whereas the WHO did not record any confirmed cases in these countries. However, other than these excpetions, the countries predicted to have infecteds via the simulations matched the actual countries that had infecteds, and the actual number of infecteds fell within the maximum and minimum numbers predicted by the simulations.

The paper goes on to discuss interesting, counter-intuitive results on the relationship between the variability of the rate-matrix γ_{ij} governing exchanges of individuals in a network (in this case the aviation network) and the predictability of the overall stochastic model. The aviation network has a relatively variable rate matrix γ_{ij}, yet this variability did not have a big impact on the variability of the global dynamics shown in the simulations, which is a somewhat counter-intuitive result. To investigate this behaviour, two simplified networks were considered. The first simplified network has only two populations, A and B, of equal size N, confined from the rest of the world, exchanging individuals at a rate γ. In this network, given a sufficient number of infecteds in A, an epidemic will occur, and for a rate of exchange $\gamma > 0$ of individuals from A to B, a subsequent outbreak may also occur in B, after time lag T.

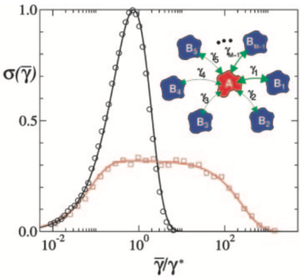
The probability $p(\gamma)$ of an outbreak occurring in population B as a function of the exchange rate γ is shown in the graph on the right. The insets in the top of the graph show histograms of the number of realizations where outbreaks occurred and the time lag after which they occurred. Three histograms are shown, each based on a different exchange rate γ. As the exchange rate increases, the variability in the time lag of the outbreak decreases. For high exchange rates between A and B, there is shorter, more predictable time lag before the outbreak in B occurs. However, the paper notes that, even for exchange rates where $p(\gamma) \approx 1$, the time lag still has a high degree of variance. This first network highlights that a stochastic exchange rate between two confined populations means there is uncertainty in how long it will take for an outbreak in the first population to cause an outbreak in the second.



Now, in order to illustrate how predictability is 'regained' in the more complex case of the aviation network, a second simplified network in the shape of a star is considered. A central population A, which we assume has had an outbreak, is connected to $M-1$ populations $B_1, ..., B_{M-1}$. Two different types of exchanges between A and its surrounding populations are considered - one in which the exchange rates γ_i between A and each of its surrounding populations are identical ($\gamma_i = \bar{\gamma}$) and one in which the exchange rates are drawn from a distribution $q(\gamma)$ with a high degree of variability between γ_{min} and γ_{max}.

$$q(\gamma) = \frac{1}{log(\gamma_{min}/\gamma_{max})} \cdot \frac{1}{\gamma},$$
$$\gamma_{min} \leq \gamma \leq \gamma_{max}$$

The relationship between the *cumulative variance per population* $\sigma(\bar{\gamma})$ and the mean exchange rate $\bar{\gamma}$ of the two different types of exchange rates are compared in the graph on the right. The black line represents the network where all exchange rates are the same, and the brown line represents the network with distributed exchange rates. As we can see, in the left half of the graph ($\bar{\gamma} \leq \gamma^*$), the network with the same exchange rates has a higher cumulative variance and thus a lower predictability than the network with distributed exchange rates. For higher exchange rates, however, the network with equal exchange rates rapidly becomes less variable, whereas the network with distributed exchange rates remains variable for longer. The purpose of this star network is to give us insights into how stochastic exchange rates in the aviation network can counter-intuitively lead to an overall higher degree of predictability.



The paper ends in demonstrating how this model can be used to come up with control strategies for future infectious diseases. The first way that it can be of use is in identifying other countries at risk given an initial outbreak in a certain location. This is visualized for two hypothetical outbreaks - one originating in New York and one in London . As a control mechanism, the paper focuses on vaccinating a portion of the population, reducing the number of susceptibles. In the current COVID-19 outbreak, vaccinations have not yet been fully developed, but if we abstract the 'vaccinations' mentioned in this paper to 'control mechanisms that reduce the number of susceptibles' (for instance, making individuals self-isolate), we could still take lessons from it.

Spoiler alert, South Africa is more at risk in the case of an initial outbreak in London than in New York

The second source of advice for devising control strategies comes from randomly placing an infected individual at a region covered by the model and examining the proportion of the population that needs to be vaccinated in order to prevent an epidemic from spreading. This proportion increases rather dramatically if the infected individual is allowed to travel 2 or 3 times. In the case that an infected individual is allowed to travel twice, an expected 74.58% of the population needs to be vaccinated to prevent the pandemic spreading ($\langle v \rangle = 74.58\%$), and in the case that an infected individual is allowed to travel 3 times, global vaccination is required ($\langle v \rangle = 100\%$).

These seem like rather dramatic implications - is a single infected individual sufficient to start a pandemic?

In order to prevent these rather dramatic implications, the effectiveness of two types of travel restrictions were considered (assuming an infected individual would otherwise be allowed to travel twice):

1. *The isolation of cities* - based on their simulations, isolating 2% of the largest cities reduces $\langle v \rangle$ from 74.58% to 37.50%.
2. *Shutting down individual connections* - to obtain a similar reduction in $\langle v \rangle$ as we get in isolating major cities, the top 27.5% of connections would need to be taken off the network

What the paper does not mention is the number connections accounted for by 2% of the largest cities

It would seem that the most effective measure in preventing an outbreak is preventing travel through major cities. Ultimately, the danger of allowing epidemics to become pandemics in our connected world cannot be underestimated and reacting quickly to effectively stop the spread of infectious diseases is vital. To what degree does effectively stopping the spread translate into effectively disconnecting our world, and what are the implications of these disconnections?

1.2. *The role of the airline transportation network in the prediction and predictability of global epidemics [1]*

This second paper comes a couple of years after the article discussed above, and, as opposed to primarily developing tools and modelling the SARS outbreak 'a posteriori', it develops more general quantitative tools that can be used to asses the reliability of epidemic forecasts.

The focus of the paper is on investigating the relationship between the structure of the network that allows an infectious disease to spread and the overall pattern, both in time and space, of the infectious disease. Specifically, it investigates the link between the seemingly erratic evolution of infectious diseases and our modern, large-scale aviation network with its associated *heterogeneities*. The complexity of the modern aviation network makes it unwise to use standard homogeneous approaches to model the spread of disease. Instead, this article uses a model similar to the one proposed in the above article - a global, stochastic, epidemic model which couples local disease dynamics (using a stochastic formulation of the SIR model) with a model of the complete worldwide airport network (*WAN*).

In this paper the aviation network is referred to as the *air-transportation network* or the *worldwide airport network* (*WAN*)

The air-transportation network is represented by a weighted graph. Vertices represent airports and the edges between them represent flight paths or connections between airports. The weights associated with each edge w_{il} represent the passenger flow between airports j and l. In addition, there is information on the size N_j of the large metropolitan area served by each airport j. The data-set, obtained from International Air Transport Association database, includes information on 3 100 airports and 17 182 connections. Their model of the aviation network allegedly accounts for 99% of the worldwide traffic.

The aviation network is highly heterogeneous in its connectivity pattern and in the traffic capacities. In contrast to the linear relationship assumed by the previous article between the capacity of each airport and the size of the surrounding population, this article finds a number of nonlinear relationships. The relationship between the local population N_j and the traffic $T_j = \sum_l w_{lj}$ handled by the corresponding airport is found to follow the nonlinear relationship $N \approx T^{\alpha}, \alpha \cong 0.5$. Additionally, the relationship between the traffic of an airport, and its number of connections k is nonlinear with the form $T \approx k^{\beta}, \beta \cong 1.5$. Along with these nonlinear relationships, there are a number of heavy-tailed distributions associated with each airport. The distribution of the probability that an airport j has k_j connections and handles T_j passengers has heavy tails and exhibits large statistical fluctuations. The probability distribution of the weights w of the connections is skewed and has heavy tails, and the distribution of the surrounding populations N_j is heavy-tailed.

I imagine that "heterogeneities in connectivity patterns and traffic capacities" translates to a high variation in the degrees of each vertex, and in the weights of each edge

This is apparently in agreement with 'Zipf's Law' - self-prompt for further reading

The article goes on to develop quantitative tools in order to discern whether the network structure influences the pattern of spread of an infectious disease. The heterogeneity in the pattern of an epidemic may purely reflect the stochastic nature of the transmission of infections. In order to monitor the spread of the disease, snapshots in the form of the *prevalence* of the disease are recorded. The prevalence of an infectious disease in city j at time t is the proportion of the local population that is infected $i_j(t) = I_j(t)/N_j$. A global snapshot of the relative prevalences in each city can be captured by the normalized vector $\vec{\rho}$ with components $\rho_j = i_j / \sum_l i_l$. The heterogeneity in the disease prevalence in these global snapshots can be measured by the disorder, or *entropy* in $\vec{\rho}$. The tool which was used to measure entropy is the 'normalized entropy function $H(t)$', defined as

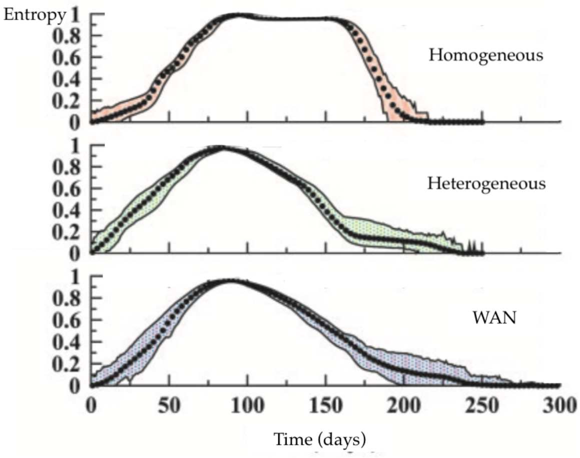
$$H(t) = -\frac{1}{\log V} \sum_j \rho_j(t) \log \rho_j(t)$$

$H = 1$ when all prevalences are equal (i.e. homogeneity), $H = 0$ when there is only one initial infected city - the most localized and heterogeneous situation

The entropy of an infection over time in the WAN is compared to the entropy of infections that spread over 2 other constructed networks:

1. A homogeneous random graph
2. A heterogeneous graph that retains the topology of the WAN

The homogeneous graph is constructed to have a Poissonian degree distribution, peaked around the average degree of the vertices of the WAN. There are no other details of the construction of the heterogeneous graph - how is it different from the WAN?



This visual comparison above provides evidence that the heterogeneity in the epidemic pattern is a result of the heterogeneity in the network structure. The shape of the entropy-time curve of the simulations with a heterogeneous network matches the shape of the curve of the simulations with the worldwide airport network. The pattern showcased in these last two curves is of an initially isolated infection which spreads. Just before 100 days, there is briefly an almost equal, global prevalence of the disease. From there on there is an increase in the heterogeneity of the epidemic pattern, and wider confidence interval (less certainty) for the average entropy value $H(t)$. In contrast, the simulations of the homogeneous network have a longer window of almost even global prevalence of the disease, followed by a rapid decline in the average entropy value. From this quantitative examination, the article concludes that the broad nature of the degree distribution largely determines the overall properties of the epidemic pattern.

The next tool developed is concerned with quantifying the reliability of the forecasts produced by stochastic epidemic models. A forecast is reliable if different realizations of the stochastic epidemic model are reasonably similar. This time, to obtain snapshots of the realizations, we consider the normalized probability that, at time

t, an infected individual is in city j, captured by the vector $\pi(\vec{t})$ with components $\pi_j(t) = I_j(t)/\sum_l I_l(t)$. Let $\vec{\pi}^I$ and $\vec{\pi}^{II}$ be two realizations of the epidemic. The statistical similarity between these two realizations can be measure by the 'Hellinger affinity'

$$sim(\vec{\pi}^I, \vec{\pi}^{II}) = \sum_j \sqrt{\vec{\pi}_j^I \vec{\pi}_j^{II}}$$

and, because normalized similarity measures "do not account for the difference in total epidemic prevalences", we also have to consider the similarity between the global prevalences of two realizations, $sim(\vec{i}^I(t), \vec{i}^{II}(t))$ where, as before $i(t) = \sum_j I_j(t)/N$ is the proportion of the global population that is infected and $\vec{i} = (i, 1-i)$.

This is the first time I have come across the Hellinger affinity. It seems that Hellinger distance $H(P,Q)$ between two discrete distributions P & Q is a measure of the similarity of two distributions, and the Hellinger 'affinity' in this article is related to the Hellinger distance by $sim(P,Q) = 1 - H(P,Q)^2 = \sum_j \sqrt{p_j q_j}$

The overlap function,

$$\Theta(t) = sim[\vec{i}^I(t), \vec{i}^{II}(t)] \times sim[\vec{\pi}^I(t), \vec{\pi}^{II}(t)]$$

is used to quantify the overlaps between two realizations. When $\Theta(t) = 1$, a maximal overlap occurs indicating the very same cities have the same numbers of infectious individuals in both realizations. When $\Theta(t) = 0$, there are no infected cities in common between the two realizations.

In the article, the last term of the overlap function is written as $sim[$

The overlap function was applied to the models based on the three different networks (homogeneous, heterogeneous and worldwide). The forecast produced by the homogeneous network was found to be the most reliable with $\Theta(t) > 80\%$. The heterogeneous network and the WAN were less reliable, particularly at the initial stage of the outbreak. The article goes on to relate the level of predictability to the outcome of the conflict between:

1. Heterogeneity of the connectivity pattern, which provides multiple equivalent channels for the travel of infected individuals, lowering predictability
2. Heterogeneity of traffic flows, which introduces dominant connections and preferred pathways for the travel of infected individuals, increasing predictability

Ultimately, the heterogeneity of the connectivity pattern influences the predictability more than the heterogeneity of traffic flows. To illustrate this last point, two hypothetical outbreaks were compared. The first was an outbreak that started in a city with a large, highly-connected airport, and the second was one that started in a city with an airport with few available connections. The epidemic that started in the city with a highly-connected airport had many equivalent paths for the disease to spread, and the resulting forecast was less reliable, with an overlap that decreased to $50 - 60\%$. The outbreak in the city with few connections developed more predictability because of the few available paths for the disease to spread.

Although this article admits that additional details with regard to the disease dynamics would have to be included in order to make the forecasts more realistic, it highlights that it is possible to use large-scale mathematical models that incorporate the complexities of the full aviation network to model infectious diseases. It concludes that the general air transport network properties are relevant and responsible for global patterns of infectious diseases and it offers useful quantitative tools that can measure the predictability of future epidemic models.

References

[1] Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. 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, Mar 2006.

[2] Alissa Eckert and Dan Higgins. An illustration revealing the ultrastructural morphology exhibited by coronaviruses., 2020.

[3] L. Hufnagel, D. Brockmann, and T. Geisel. Forecast and control of epidemics in a globalized world. *Proceedings of the National Academy of Sciences*, 101(42):15124–15129, Nov 2004.