Network Tour of Data Science Project: Would Humanity Survive a Spanish-Flu Pandemic Today?

Team 03
Dönz Jonathan, Esguerra Martin, Vojinovic Stefano

January 2020

1 Introduction

Human existence has always been affected by different catastrophic events that have change the course of history. The bubonic plague for example killed more than 25 million. Pandemics have been one of the biggest threats for the survival of the species and even with all the scientific developments we have today, they are still a probable cause for the extinction of humanity. As the world becomes more and more connected, diseases can be spread extremely fast as we have seen with the recent SARS epidemic. Having at our disposal an international flight routes database, we will explore how a new pandemic would spread given different contagion and mortality coefficients and what control strategies could be useful to stop the disease. To do this, we will build a graph of the routes and explore its properties.

2 Network Construction and Analysis

2.1 Data Acquisition and Pre-Processing

The OpenFlights/Airline Route Mapper Database[1] contains 67,663 routes between 3,321 airports on 548 airlines spanning the globe. It is separated into 4 datasets containing information on airlines, airports, planes and routes. We only use the airports and routes datasets.

The nodes of the graph are given by the airports in the corresponding dataset and the edges are given by the routes dataset. In the data there were multiple airports that were not connected to any other airport, hence they were dropped from the graph as they were useless for the analysis (7000 airports at first, 3200 after filtering). There are also some airports present in the routes data that do not appear on the airport data, these were also ignored. To create the graph, we assigned an ID to each airport and built an adjacency matrix whose entries $A_{i,j}$ represent the number of flights from airport i to airport j. This matrix is not symmetric and is weighted. We also built 3 other matrices, a weighted symmetric, and the unweighted symmetric and asymmetric. In our analyses we generally assume that we use the weighted symmetric one and we state when we use a different one. The resulting graph has 3216 nodes and 18857 edges.

[1] The datasets are available at `https://openflights.org/data.html`.

Figure 1: Airports network visualization. Each airport is indicated by a dot at its geographical location. The size of a dot is proportional to its number of connections. The curves between the airports are the edges of the network. The thickness of the curves is proportional to the weight associated with the edge. The colors are an adaptation of the K-means clustering applied to the network.

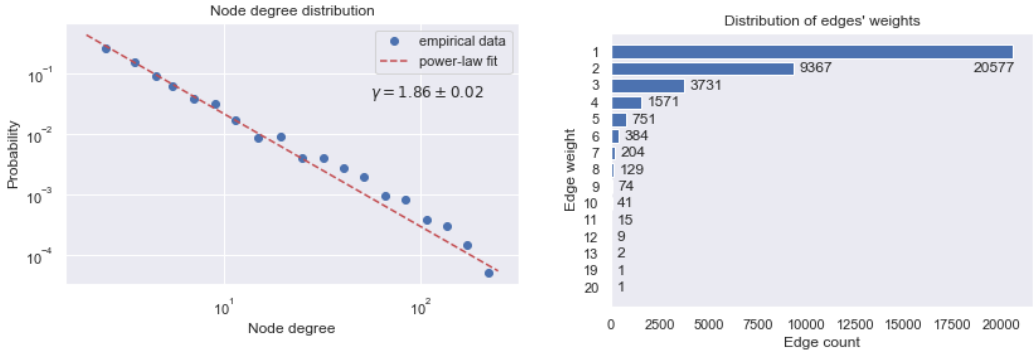
We made some property analysis on this graph and found that there are 11 connected components (CC) and that the largest CC contains almost all the nodes with 3,186 airports. The second largest contains only 10 airports, which are situated on several islands of New Caledonia. The rest of the CCs all contain less than 5 airports. In the remaining of our work, we only keep the largest connected component as the other CCs are of negligible size, and it simplifies the epidemics' simulation. Throughout the rest of this document, we refer to this network as the *airports network*. Its characteristics are given in Table 1.

2.2 Network Characterization

The high value of $\langle k^2 \rangle$ and the low value of the median degree reported in Table 1 suggest that the airports network belongs to the family of scale-free networks. Scale-free networks are characterized

Table 1: Properties of the airports network.

Feature	Value
number of nodes	3186
number of edges	18,832
average node degree $\langle k \rangle$	11.822
average squared node degree $\langle k^2 \rangle$	763.7
clustering coefficient	0.493
median node degree	3
average shortest path	3.958
diameter	12



(a) Degree distribution

(b) Weights distribution

Figure 2: Degree and edges' weights distribution of airports network.
(a) Degree distribution of the airports and power-law fit. The blue dots represent the normalized degree distribution of the nodes. Note that both axes are logarithmic and the data have been binned logarithmically. It means that the size of the bins grows exponentially with the node degree. The dashed red line shows the theoretical degree distribution of a power law with $\gamma = 1.86$.
(b) Distribution of edges' weights. It also follows a power-law distribution. The majority of the edges have weight 1, meaning that there is a single flight between two airports in general. The maximum number of flights between two airports is 20.

by a single parameter γ, from which their degree distribution depends as $p(k) \propto k^{-\gamma}$. In line with the work of Clauset et al. [1], we computed the value of the γ-parameter by fitting the node degree distribution on a log-log plot, using logarithmic sized bins. The degree distribution and the theoretical power-law distribution using the optimal value γ are represented in Figure 2a.

Visually, the power-law fit matches relatively well the empirical data. The airport network is therefore well approximated by a scale-free network with parameter $\gamma = 1.86$. As $\gamma < 2$, the network belongs to the *anomalous regime* according to the classification of Prof. Barabási [2][2]. This small value can be explained by the presence of many hub airports, contributing to a heavy-tailed nodes' distribution.

The edges also follow a power-law distribution, with 55% of the edges being of weight 1, meaning that 55% of the connected airport pairs have a single flight between them.

To support our use of the *symmetric* weighted adjacency matrix as opposed to the asymmetric one, we calculated that 95.8% of the airport pairs are symmetric, i.e. they have the same number of flights from one to the other. The distribution of the differences in number of flights between pairs of airports is shown in Figure 3.

2.3 Clustering

For our further epidemic simulations, we wanted to explore how the epidemic would evolve within and between the clusters obtained using K-means clustering (KM) and Spectral Clustering (SC). With no a priori incentive for the numbers of clusters k to use, we tried several number of clusters. We show the relative sizes of the clusters with k ranging from 2 to 10 in Figure 4.

[2]The term *anomalous regime* stems from the fact that such a network would require the largest hub to be connected to more nodes than present in the network for a sufficiently large network size.

Figure 3: Distribution of routes' asymmetry. The vertical axis represents the number of routes (aka *pairs of airports*) that have a difference in number of flights from one to the other equal to the value of the horizontal axis. The majority (95.8%) of routes are symmetric, i.e. they have the same number of flights between their airports' pair.

We can see that the clusters' sizes are well balanced for any value of k using KM. On the other hand, there is always one cluster containing most of the airports in the case of SC, no matter the number of clusters. We provide a visual representation of the clusters obtainted by these two methods in Figure 5.

3 Epidemics simulations

3.1 Epidemics model

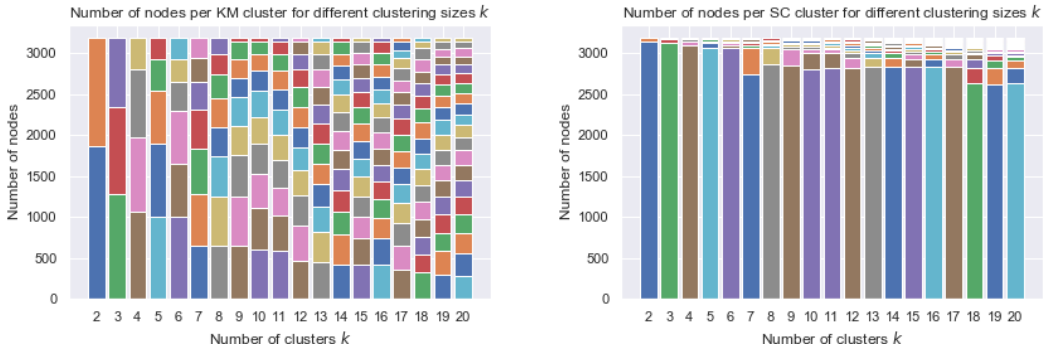
The Susceptible-Infected-Susceptible (SIS) epidemics model was chosen. It is usually used to model an epidemic spreading from person to person, where each person is either in a *susceptible* (S) state, or an *infected* (I) state. A susceptible person can become infected by contact with an infected person with a certain probability, and an infected person can become susceptible after some time. The probability of a person becoming infected through a contact with an infected one, and the expected time required for an infected person to become susceptible are modeled with the parameters β and μ respectively. Figure 6 is a schematic representation of the SIS model.

3.2 Choice of parameters

Several research papers specifically focused on estimating the Spanish flu's reproducible number and incubation period reached similar results (see for instance [3, 4, 5]). In accordance with the work of Vynnycky et al. [4], we set the reproducible number R_0 to 2.5. The relationship between R_0 and the parameters of the model is

$$R_0 = \frac{\beta}{\mu}. \tag{1}$$

In our case, as the simulation method proceeds in continuous time and has no straightforward interpretation in terms of concrete time, we arbitrarily chose a value for μ and derived the corre-



(a) K-Means clustering

(b) Spectral clustering

Figure 4: Relative sizes of clusters for different number of clusters k for both methods K-means (a) and Spectral Clustering (b). The clusters sizes are balanced for any value of k in K-means clustering. A single cluster always contains the majority of the nodes in the case of Spectral Clustering.

sponding value for β from equation 1. In all the following simulations, we use $\beta = 0.025$, $\mu = 0.01$, and we run the simulation for 100 time units.

The weight of an edge between two airports multiplies the probability of infection by the weight's value.

3.3 Limitation of the model

The application of the SIS model to an epidemic taking place on an *airports* graph requires some different interpretations than usual.

The most important limitation is that an airport is either in a susceptible or an infected state. This assumption is adequate only if the flow of infected people are large, so that it can be assumed that an airport transitions from susceptible to infected after the income of passengers from a single other airport. The model is therefore more suited to study an epidemic at a stage where it is already widespread.

3.4 Experiment 1: Impact of the degree of the initially infected airport

We first investigated whether the number of connections of the only initially infected airport led to a different spread dynamics. To this end, we selected one airport with a high degree, one with a medium, and two with low degrees and ran the epidemics simulation on the airports graph. The results are shown in Figure 7.

We can see that the widespread is almost immediate for the most connected airport of the dataset, Amsterdam Airport Shiphol with 248 flights connections. In the case of Sibu Airport, which has 6 connections, a short period with unnoticeable infections precedes a widespread similar to the one obtained with the Amsterdam airport. We interpret this initial period as the time needed by the epidemic to reach a hub, enabling it to widely spread through the network then. This initial period is longer in the case of Nanaimo airport, which has only two connections, but the epidemic spreads wide eventually. In the case of Peawanuck airport, which has a single connection and is

(a) K-means

(b) Spectral Clustering

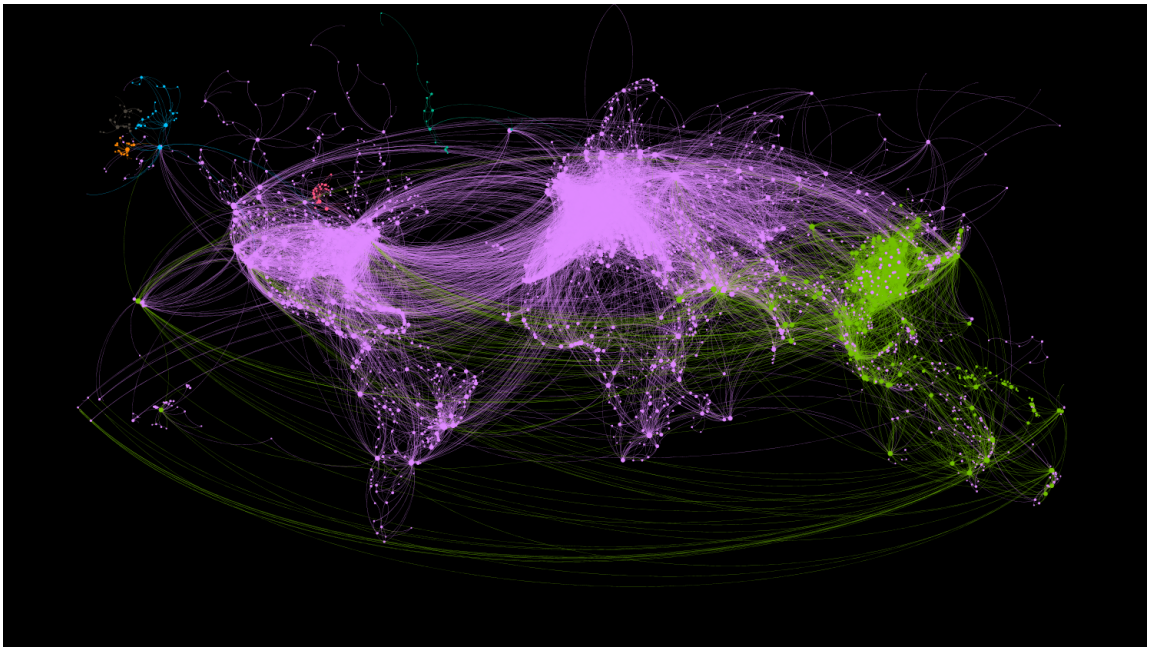


Figure 5: Clusters from K-means and Spectral Clustering using $k = 8$.

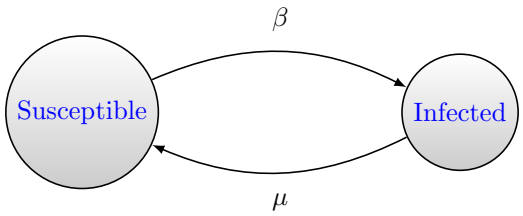


Figure 6: SIS model. An individual can only be in two states, *Susceptible* and *Infected*. β is the transmission rate, the rate at which an infected person infects a susceptible one. μ is the recovery rate, the rate at which an infected person becomes susceptible.

one of the remotest airport from the dataset[3], the epidemic never spreads over the network.

In all the cases where the epidemic spreads through the whole network, a similar endemic state is reached with 2781 ± 15 airports infected. This is about 87% of all the airports.

From this experiment, we conclude that the node degree influences the time at which the spread explodes in the network. The only way to avoid a widespread epidemic characterized by the present parameters is if the initially infected airport is very remote.

In all the next experiments, we use Hong Kong International Airport as initially infected airport. We made this choice because we believe that the likelihood of it being the initially infected airport in a real-world scenario is very high. Here are a few reasons supporting this claim.

- It has 133 connections and is therefore a hub.
- The city is highly and densely populated with a population of 7.48 million and a density of 6,777 people per km2 [6].
- Southern China was the cradle of the SARS [7] epidemic in 2002 and the Black Death [8] in the 14th century.

3.5 Experiment 2: Epidemics spread from the clusters perspective

We then investigated whether some insight could be derived by viewing the epidemic's spread from the clusters obtained by K-means (KM) and spectral clustering (SC). To this end, we ran 4 simulations and we isolated the time series of the infection from every cluster. The time series of one of the simulations accompanied by the display of the susceptible and infected airports on a world map is shown in Figure 8. The time-series of the clusters is shown in Figure 9.

It can be observed that, irrespective of the clustering method, the larger a cluster is, the earlier and the faster the spread occurs within the cluster. The small clusters from spectral clustering are noisy because they contain very few airports and therefore are subject to a high variance over the different simulations.

4 Control strategies

Next we wanted to design control strategies and compare their effectiveness on the extent of epidemics' spread. We implemented the 5 following methods.

[3]Six flights are necessary to reach the nearest international airport from it.

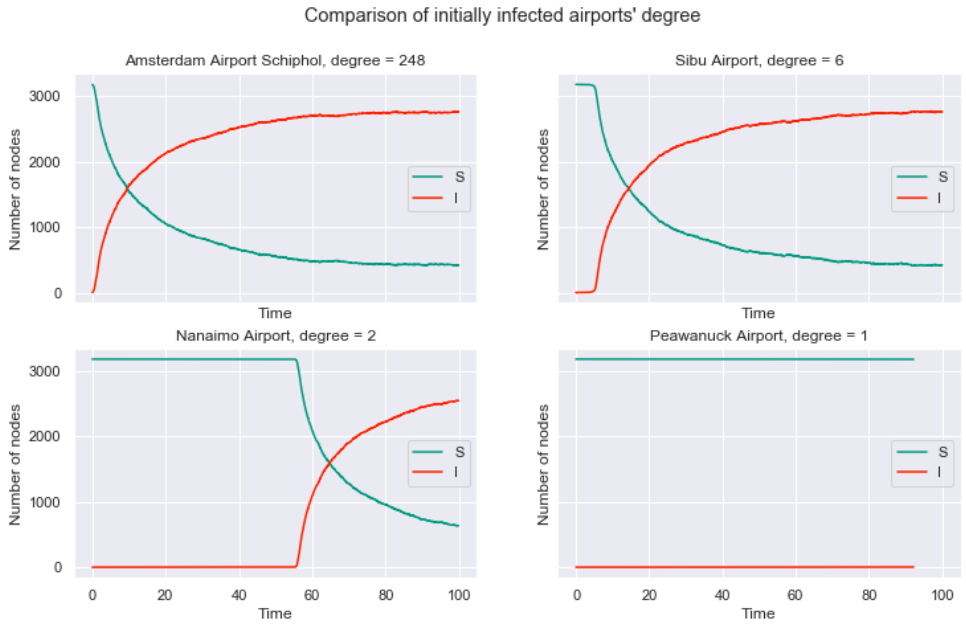


Figure 7: Impact of the degree of the initially infected airport on spread dynamics. Each plot represents a simulation with a different initially infected airport indicated in the title. Each plot shows the time evolution of the number of susceptible and infected airports.

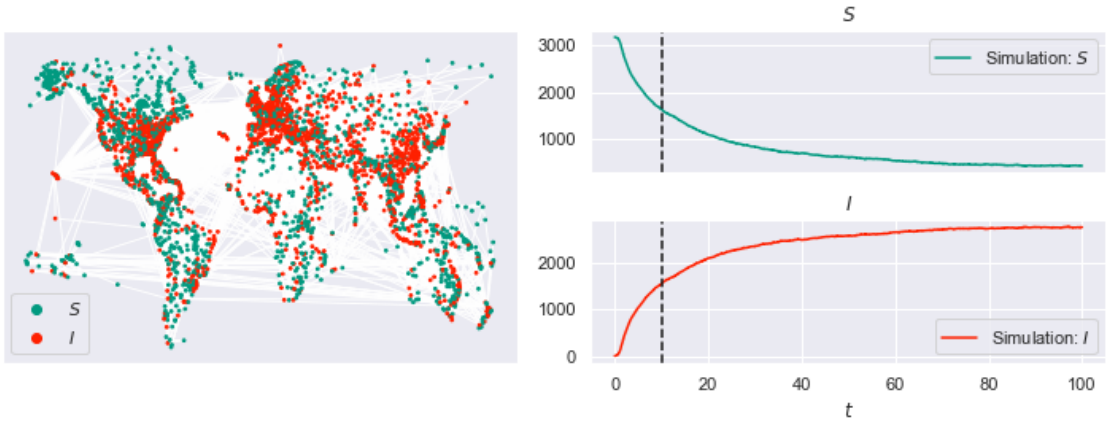
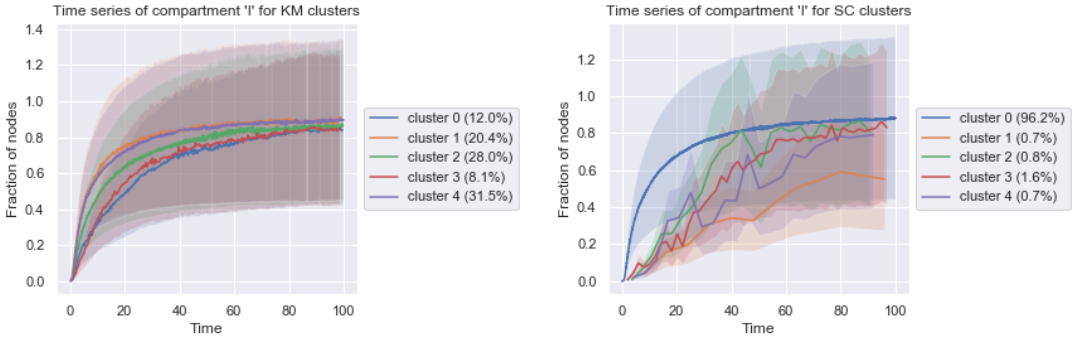


Figure 8: Snapshot of epidemic simulation starting with Hong Kong international airport as initially infected. The snapshot is taken after 10 time units. The left panel shows the susceptible (S) and infected (I) nodes in respectively green and red. The two panels on the right show the time series of the number of nodes in the susceptible (top) and infected (bottom) state.



(a) K-Means clustering

(b) Spectral clustering

Figure 9: Time series of infected nodes from the clusters' perspective. Each cluster is associated with a given color. The solid line represents the average proportion of nodes from the cluster being infected at a given time. The light colored overlay is the standard error of the mean of the solid curve calculated from the different simulations. The respective size of the clusters is indicated in parentheses. (a) and (b) represent the time series using the clusters from respectively K-means and spectral clustering.

1. `random_airports_removal`
   Shutdown 20% of airports of the network. Airports are randomly chosen.
2. `random_neighbors_removal`
   Select 20% of airports at random, then shutdown a random neighbor for each of these.
3. `largest_airports_removal`
   Shutdown the top 20% connected airports.
4. `largest_infected_airports_removal`
   Shutdown the top 20% connected airports that are infected.
5. `largest_routes_removal`
   Remove the top 45% connected routes.

The choice of the percentages of removals is chosen to have a significant and observable effect on the epidemics spread.

4.1 Experiment 3: Control strategies' impact on extent of epidemic spread

We ran simulations where we applied each control strategy at the same arbitrary time of $t = 40$. The resulting time-series are shown in Figure 10. We make the following observations on these results:

- Random airports removal seems quite ineffective as the epidemic's growth (slope of the infected curve) is weakly attenuated after time 40, if at all.

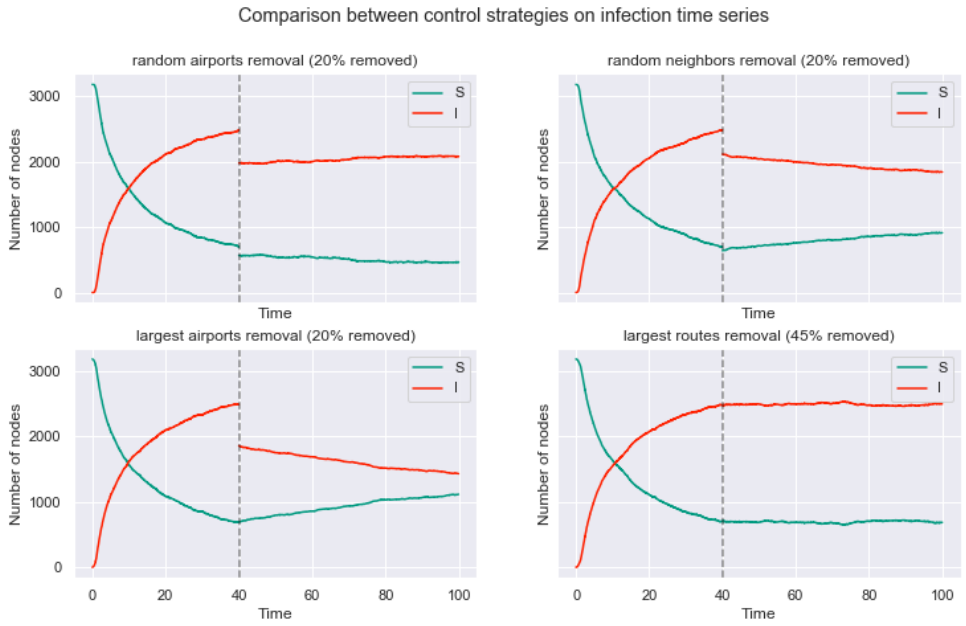


Figure 10: Control strategies' time series. Each plot is associated with the control strategy indicated in its title. It shows the time evolution of the number of airports in the susceptible (S) and infected (I) states in a simulation where the control strategy is applied at time $t = 40$.

- Random neighbors removal is more efficient than random airports removal. This is expected from the friendship paradox. We notice that most of the random neighbors airports that have been selected turn out to be infected airports.
- Largest airports removal is the most efficient control strategy but it is also the most disruptive of the aerial network: shutting down the largest hubs rapidly breaks the airports network into several isolated networks.
- Largest routes removal requires to remove about 45% of the routes for the epidemic to stay constant and not grow after the treatment.

Note that we did not display the results of the method `largest_infected_airports_removal` because they were essentially the same as the ones from `largest_airports_removal`. This means that at time $t = 40$, most of the airports with the most connections are infected.

4.2 Experiment 4: Control strategies' sensitivity to time of treatment

We finally investigated the impact of the time at which we apply a control strategy. We wanted to be able to respond to the questions: *Is acting early important to stabilize the spread for all control strategies? Is it more beneficial for certain strategies?*

We ran simulations for different treatment-times ranging from $t = 10$ to $t = 95$ with every increment of 5 in between. For each treatment-time, we ran 4 simulations and computed the

average number of airports being infected at the end of the simulation (time = 100). This is our proxy for how well the control strategy worked. The results are shown in Figure 11.

The plots give insight on the sensitivity of each method to the time at which the control strategy is applied. As the relationship between the number of infected airports and the time of treatment (time at which the control strategy is applied) is linear in all cases, we fitted a linear regression for each control strategy.

The *y-intercept* of the linear regression can be interpreted as the efficiency of the control strategy. The smaller it is, the more effective a control strategy is. The reason is that it represents the number of infected airports if the control strategy had been applied at the very beginning of the epidemic's spread.

The *slope* of the linear regression can be interpreted as the sensitivity of the control strategy to the time of treatment. The larger it is, the more sensitive the control strategy is, and the more beneficial it is to apply it early.

Equipped with these interpretations, we can comment the results from Figure 11.

- Random airports removal is quite ineffective as the y-intercept of its regression is large. It is not sensitive to the time of treatment either as its linear regression has a small slope.

- Choosing a random neighbor of a random airport is more effective than just choosing a random airport as can be seen from the improved performance of *random neighbors removal* strategy. We can see that it is effective (y-intercept of 839 infected airports versus 1484 with the former strategy) and it is sensitive to the time of treatment (slope = 9.8 [infected airports / time of treatment])

- The best strategy is - with no surprise - the removal of the largest airports. The y-intercept of about 50 infected airports is impressively low, meaning this control strategy almost interrupts the spread if it is applied very early. It can be seen from the two first data points that a very early application of the strategy almost completely disrupts the epidemic. The sensitivity is also the largest, meaning this strategy benefits the most of being implemented early.

- Removing 45% of the routes is less effective than the other strategies but it differs fundamentally from the other strategies that are all about removing airports. Therefore we don't hold strong comparison claims. We note that it is nevertheless more sensitive to the time of treatment than the random airports removal strategy as it has a larger slope.

5 Discussion

The simulations of epidemic spread (experiments 1 and 2) showed that if an epidemic was starting in a relatively connected airport, there would be very little chance that it would stay local and not spread through the whole network at some time. The endemic state reached by a Spanish flu-like disease spread onto the airports network would result in about 87% of the airports being infected.Even if we downweight this result by considering the limitations our epidemic model, this result indicates that a relatively contagious disease has a high potential to spread over the whole world quickly through the aerial network.

The experiments on the control strategies (experiments 3 and 4) demonstrate that the efficiency of a control strategy depends on the knowledge of the graph onto which the epidemic spreads.

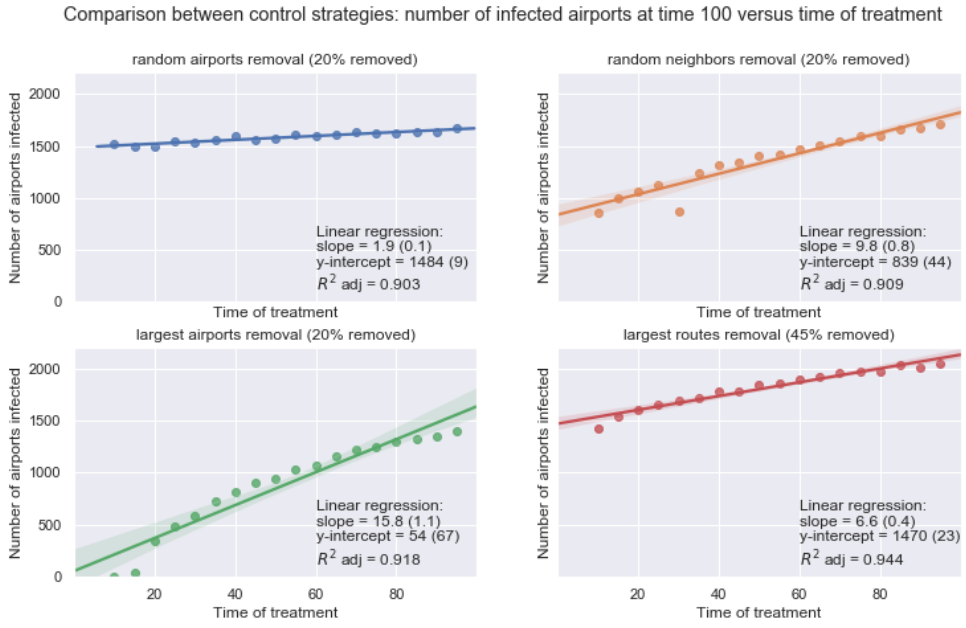


Figure 11: Time of treatment impact with the different control strategies. Each plot is associated with a control strategy. For each control strategy, the dots represent the number of infected airports at the end of the simulation as a function of the time at which the control strategy was applied. Each dot is the average obtained over 4 simulations. The line is a fitted linear regression of the dots. Its statistical parameters are reported on the bottom right of the plots.

We saw that removing airports completely at random, even in such high proportion as 20%, is quite ineffective, no matter at what time the control strategy is applied. Having the additional information of who the neighbors of randomly selected airports are led to a very significant increase in efficiency. It also increased the sensitivity to the time of treatment. This is a good illustration of the friendship paradox [9]. The best efficiency and the largest sensitivity were obtained when the knowledge of the degree of the airports of the network was used to target them.

6 Conclusion

The airports network is very well designed to reach any part of the world in very few flights thanks to its particular network properties. The very same properties however allow the spread of contagious epidemics through the whole world in very little time. We estimate that a Spanish flu-like disease would reach 87% of all airports if no control strategy is applied.

The strategy of shutting down the most connected airports is the most effective. If it is applied early enough, setting aside the issue of disrupting the aerial network, the pandemics can be completely disrupted.

References

[1] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, Nov 2009.

[2] Albert-László Barabási. *Network Science*. Cambridge University Press, 2016.

[3] Christina E Mills, James M Robins, and Marc Lipsitch. Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904, 2004.

[4] Emilia Vynnycky, Amy Trindall, and Punam Mangtani. Estimates of the reproduction numbers of Spanish influenza using morbidity data. *International Journal of Epidemiology*, 36(4):881–889, 05 2007.

[5] Gerardo Chowell, Hiroshi Nishiura, and Luis MA Bettencourt. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society Interface*, 4(12):155–166, 2006.

[6] Wikipedia. Hong kong. `https://en.wikipedia.org/wiki/Hong_Kong`. Last accessed: January 8, 2019.

[7] Wikipedia. Severe acute respiratory syndrome. `https://en.wikipedia.org/wiki/Severe_acute_respiratory_syndrome`. Last accessed: January 8, 2019.

[8] Wikipedia. Black death. `https://en.wikipedia.org/wiki/Black_Death`. Last accessed: January 8, 2019.

[9] Scott L. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477, 1991.

[10] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks, 2009.

[11] Joel C Miller and Tony Ting. Eon (epidemics on networks): a fast, flexible python package for simulation, analytic approximation, and analysis of epidemics on networks.

Appendix A Methods

We used the software Gephi[10] to make the visualizations of our network. We used the Python library *Epidemics on Networks* [11] to perform the epidemics simulations.

Appendix B Airport network's fun facts

In this section we report the airports displaying notable traits. Note that the `airports` and `routes` datasets that we used have not been updated since respectively January 2017 and June 2014. The information we show here is therefore of historical relevance only.

Table 2: Extreme locations airports

Trait	Name	Country	Details
Northernmost	Svalbard Airport	Norway	latitude = 78.25 degrees
Southernmost	Malvinas Argentinas Airport	Argentina	latitude = -54.84 degrees
Highest altitude	Daocheng Yading Airport	China	altitude = 4,411 meters
Lowest altitude	Atyrau Airport	Kazakhstan	altitude = -21.9 meters

Table 3: Top 5 airports with largest number of flights

Rank	Name	City	Country	Number of flights
1	Amsterdam Airport Schiphol	Amsterdam	Netherlands	248
2	Frankfurt am Main Airport	Frankfurt	Germany	244
3	Charles de Gaulle International Airport	Paris	France	240
4	Atatürk International Airport	Istanbul	Turkey	233
5	Hartsfield Jackson Atlanta International Airport	Atlanta	USA	217

Table 4: Top 5 longest flights

Rank	Route			Distance [km]
1	Los Angeles (USA)	—	Lusaka (Zambia)	16,090
2	Los Angeles (USA)	—	Ndola (Zambia)	15,945
3	Dallas (USA)	—	Sidney (AUS)	13,804
4	Atlanta (USA)	—	Johannesburg (South Africa)	13,581
5	Dubai (UAE)	—	Los Angeles (USA)	13,420