Would Humanity Survive a Spanish-Flu Pandemic Today?

A Network Tour of Data Science Project

TEAM 03
Doenz Jonathan, Esguerra Martin, Vojinovic Stefano

I. INTRODUCTION

Human existence has always been affected by different catastrophic events that have changed the course of history. Among them, pandemics have been a major threat for the survival of humanity [1]. The Spanish flu that ravaged the world in 1918 killed between 50 and 100 million people for example [2]. Even with all the scientific developments we have today, they remain one of the most probable cause of sudden death of millions of people [3].

As the world becomes more and more connected, especially thanks to the aerial network, diseases can spread through the world extremely fast as we have seen with the SARS epidemic in 2003 [4]. In this work, we tackle the question: *Would Humanity Survive a Spanish-Flu Pandemic Today?* To this end, we first construct the network of the airports and flight routes of the world. Then we explore how a pandemic with similar contagion properties as the Spanish flu would spread on this network and compare control strategies to limit the spread.

II. 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 entry (i, j) represents the number of flights from airport i to airport j. This matrix is asymmetric and weighted. We also built the corresponding weighted symmetric, unweighted symmetric, and unweighted asymmetric matrices. The further analyses are performed on the weighted symmetric matrix except if we explicitly state otherwise. The choice of using the *symmetric* adjacency matrix as opposed to the asymmetric one is that 95.8% of the edges are symmetric.

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

TABLE I: 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

The resulting network has 3216 nodes and 18857 edges and is composed of 11 connected components (CC). 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. We only keep the largest connected component as the other CCs are of negligible size, and it simplifies the epidemics' simulation. This single component network is our final network and we henceforth refer to it as the *airports network*. A visual representation is displayed in Figure 1.

III. NETWORK CHARACTERISTICS

The characteristics of the airports network are given in Table I. The airports' degree distribution follows a power-law with a parameter γ equal to 1.86 ± 0.02 as can be seen from Figure 2.

As $\gamma < 2$, the network belongs to the *anomalous regime* according to the classification of Prof. Barabási [5][2]. This small value can be explained by the presence of many hub airports, contributing to a heavy-tailed nodes' distribution.

A. Clustering

For our further epidemic simulations, we wanted to explore how an epidemic evolves within and between clusters obtained using K-means clustering (KM) and Spectral Clustering (SC). With no a priori incentives for the numbers of clusters k to use, we tried several number of clusters. We noticed that the clusters' sizes are well

[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.

Fig. 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 edges is proportional to the weight associated with the edge and its color is the same as the source node. The colors are an adaptation of the K-means clustering applied to the network.

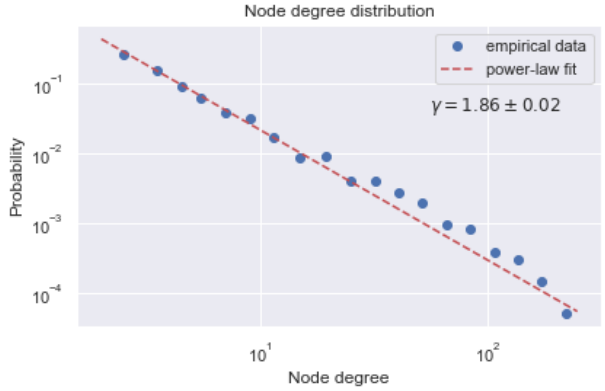


Fig. 2: Degree distribution of the airports and power-law fit. The blue dots represent the normalized degree distribution of the airports. Note that both axes are logarithmic and the data have been binned logarithmically. It means that the size of the bins grows exponentially with the airports' degree. The dashed red line shows the theoretical degree distribution of a power law with $\gamma = 1.86$.

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. These results can be seen in Figure 3. We provide a visual representation of the clusters obtained using the two clustering methods in the *extended report*[3].

[3]The *extended report* is a more rigorous and comprehensive document available on our github repository.

IV. Epidemics simulations

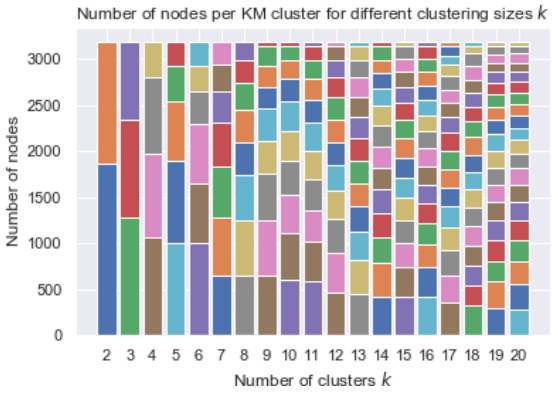
A. Epidemics model

We chose to use the Susceptible-Infected-Susceptible (SIS) epidemics model. The SIS model applied to our airports network leads to the following interpretation. An airport j is either susceptible (S) or infected (I). If it is connected to an infected airport i, it can become infected with probability $\beta \cdot w_{i,j}$, where β is the transmission rate of the disease, and $w_{i,j}$ is the weight of the edge between the two airports. An infected airport can become susceptible at a rate μ. The values of the disease's parameters β and μ are chosen to be similar to the parameters of the Spanish flu, whose reproducible number R_0 is documented to be between 2.0 and 3.0 [6]. We therefore chose to set R_0 to 2.5. It is related to the disease's parameters as $R_0 = \beta/\mu$, so we arbitrarily set μ to 0.01 and computed the value of β to be 0.025 accordingly.

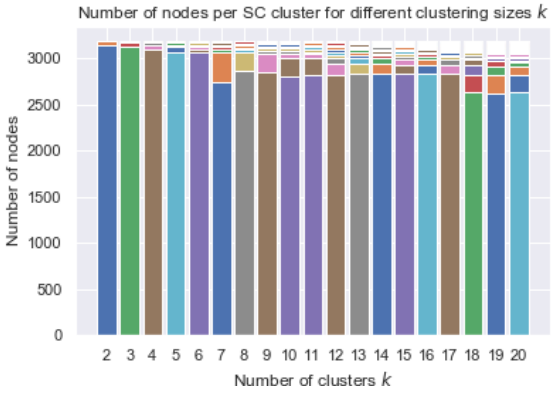
The assumption that an airport is either in a susceptible or an infected state is adequate only if the flow of infected people are large, so that it can be assumed that an airport transitions from susceptible to infected after the income of passengers from a single infected airport. The model is therefore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

B. The disease becomes widespread in the majority of the cases

We first investigated the conditions for the disease to become *widespread* (i.e. infect a significant fraction of the airports worldwide). We found out that it depends



(a) K-Means clustering



(b) Spectral clustering

Fig. 3: 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.

on the number of connections of the initially infected airport. The disease eventually becomes widespread when the initially infected airport has a sufficiently large number of connections. The only case where we managed to reliably not observe a widespread epidemics is in the case of Peawanuck airport as initially infected. This is one of the remotest airport of the dataset[4].

In all the cases where the epidemic becomes widespread, a similar endemic state is reached with 2781 ± 15 airports infected. This is about 87% of all the airports. For more details on this set of simulations and their results, see the extended report.

In all the next experiments, we use Hong Kong International Airport as initially infected airport as we believe it is likely to be the initially infected airport in a real-world scenario due to its highly and densely population and its history of being situated in the cradle of the SARS [4] and the Black Death [7] epidemics. This airport has 133 connections in the network, so it is considered a hub.

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

C. Spread from the clusters' perspective

We investigated whether some insight could be derived by observing 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 (visualizations of the results can be seen in the extended report). We observed that, irrespective of the clustering method, the larger a cluster is, the earlier and the faster the spread occurs within the cluster.

D. Control strategies

Next we wanted to apply different control strategies and compare their effectiveness on ther ability to limit the extent of the epidemic's spread. We implemented the 4 following methods.

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_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.

E. Control strategies' impact on extent of epidemic spread

We investigated the relative impact of the control strategies on the extent of the epidemic spread as a function of the time of treatment (time at which the control strategy was applied). To this end, we ran simulations for different treatment-times ranging from $t = 10$ to $t = 95$ with every increment of 5 in between. Each simulation lasts a total of 100 time units. For each treatment-time, we ran 4 simulations and computed the average number of airports being infected at the end of the simulation. This is our proxy for how well the control strategy worked. The results are shown in Figure 4.

The plots give insight on the sensitivity of each method to the time at which the control strategy is applied. The relationship between the number of infected airports and the time of treatment is linear in all cases, so 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. As it represents the number of infected airports if the control strategy had been applied at the very beginning of the epidemic's spread, the smaller it

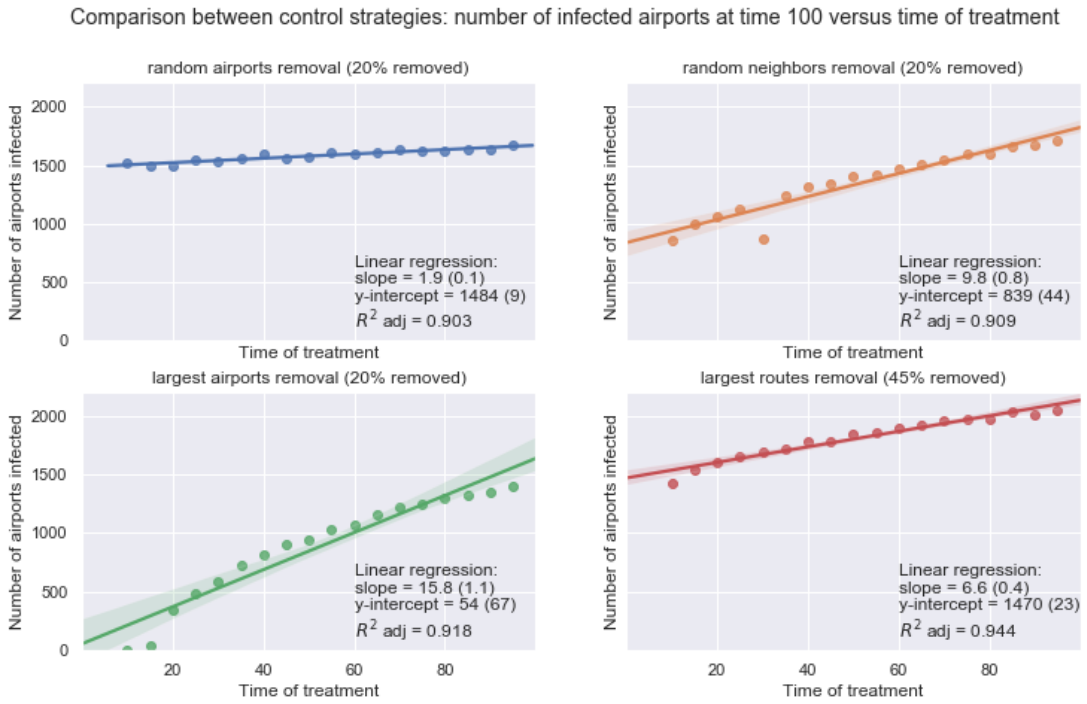


Fig. 4: Time of treatment impact with the different control strategies. Each plot is associated with a control strategy. For each control strategy, the dots represent the number of infected airports at the end of the simulation as a function of the time at which the control strategy was applied. Each 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. The numbers in parentheses are the standard errors of the paremeters' estimates.

is, the more effective is the control strategy. 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 compare the strategies based on the results from Figure 4.

- Random airports removal is quite ineffective (large y-intercept) and is not sensitive to the time of treatment either (small slope).
- Choosing a random neighbor of a random airport is more effective than just choosing a random airport (random neighbors removal strategy's y-intercept is 839 [infected airports] compared to 1484 with random airports removal). It is also relatively 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 that 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.

V. Discussion

The simulations of epidemics' spread showed that if a Spanish-flu like 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 the disease 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 simulations with the control strategies demonstrate that the efficiency of a control strategy depends on the

knowledge of the graph onto which the epidemic spreads. 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 [8]. 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.

VI. Conclusion

Based on the simple SIS model, 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 could be completely disrupted.

A real-world scenario could differ from our model in many ways, but it seems likely for a highly contagious disease to infect a significant part of the world's population. Whether the societal infrastructures keep functioning and humanity as a whole survives to such an event can be argued in both ways.

References

[1] Marcia C. Inhorn and Peter J. Brown. The anthropology of infectious disease. *Annual Review of Anthropology*, 19(1):89–117, 1990.

[2] Niall Johnson and Juergen Mueller. Updating the accounts: Global mortality of the 1918-1920 "spanish" influenza pandemic. *Bulletin of the history of medicine*, 76:105–15, 02 2002.

[3] Ezra Klein. The most predictable disaster in the history of the human race. https://www.vox.com/2015/5/27/8660249/gates-flu-pandemic. Last accessed: January 10, 2020.

[4] Wikipedia. Severe acute respiratory syndrome. https://en.wikipedia.org/wiki/Severe_acute_respiratory_syndrome. Last accessed: January 8, 2020.

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

[6] 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.

[7] Wikipedia. Black death. https://en.wikipedia.org/wiki/Black_Death. Last accessed: January 8, 2020.

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

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

[10] 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 [9] to make the visualizations of our network, and the Python library *Epidemics on Networks* [10] to perform the epidemics simulations.

Appendix B
Airport network's fun facts

See the extended report to discover these fun facts!