BRACQUE VENDRELL Maria
CONDETTE Théo

Mars 2024

Graph Analysis Project:
Diseases Spreading Process in the US Air Transportation Network Before and during the COVID-19 pandemic

In this project, you will investigate the spreading of a disease, an opinion, a behavior, information, or another transmissible resource, on a real network. Please leave the instructions in italics. Attach your code as a RmD file.

1. Problem Statement

Describe the network you are studying. What contagion process are you interested in and why is it interesting or relevant to study it in this network? (max 200 words)

COVID-19 is an example of how difficult it is to deal with unknown diseases in an interconnected world. Many restrictions have been taken at national and international levels to slow the diseases and Air transportation has been highly affected.

Thus, we investigate the spreading of different contagious diseases on networks with and without COVID-19 restrictions to evaluate their effectiveness.

The networks under study correspond to the air transportation flow of the United States before and during the COVID-19 pandemic, respectively in 2019 and 2021. The networks have cities as nodes, routes between cities as edges, and the total number of flights per year per route as weight. The idea is to analyze and compare 2019 and 2021 air transportation restrictions for different disease scenarios (low, medium, and high contagious diseases).

2. Expectations

Formulate expectations about the dynamics of contagion. Is it a case of simple or complex contagion? How do you expect it to unfold? Which nodes do you expect to be affected the most? (max 300 words)

Concerning the dynamic of the contagion, we expect :

- Network with COVID-19 restrictions :
    - Low contagious diseases: The diseases do not spread at all. Only a few cities are affected.
    - Medium contagious diseases: First we observe a hub then it propagates to the full network.

    - Highly contagious diseases: The diseases spread in an uncontrolled manner.

- Network without restriction :
    - Low contagious diseases: The diseases might spread in some cities in a hub.
    - Medium contagious diseases: The diseases spread in an uncontrolled manner.
    - Highly contagious diseases: The diseases spread in an uncontrolled manner.

The contagion we are considering is a simple contagion case. The diseases can be transmitted with only one interaction. Nodes connected to most cities are expected to be the most impacted by the diseases since they are well-connected and may act as potential sources or hubs for disease transmission. We expect to see cities such as Atlanta, New York, Dallas, Denver, and Chicago be central as they have the biggest airports.

3. Research design

Propose one centrality measure to identify the nodes that are at higher risk of being affected by the process. This measure should depend on edge weights (if no weights are available, the similarity between node attributes can be used to create a weight). (max 300 words)

In this context of contagion, the Strength (weighted vertex degree) Centrality measure seems to be a good one to use. Indeed, it considers both the number of connections a node has and the strength of those connections. In our case, the weight of an edge represents the volume of air traffic between two cities, thus representing the potential for disease transmission.

Nodes with high Strength values would be likely to be major hubs with extensive connectivity and substantial passenger traffic passing through them. These cities with major airports play a central role in facilitating the spread of contagious diseases through air travel due to the high volume of passengers traveling to and from these locations.

To sum up, by using Strength as a centrality measure in the analysis of air traffic, we can effectively identify key cities that are not only heavily connected but also serve as critical nodes in the air transportation network. This information can be invaluable for implementing targeted measures, such as enhanced screening protocols or travel restrictions, to mitigate the risk of disease transmission via air travel.

Explain how to formulate an ABM representing your process. Clearly explain the rules underlying the ABM (possibly with a diagram or a flow chart), its starting conditions, and the parameters it requires. (max 400 words)

For this project, the objective is to create an Agent-Based Model (ABM) to simulate disease spread in the air transportation networks of the United States in 2019 and 2021. The transmission rate, denoted as β, represents the level of contagion of the disease, and we will compare its effects on disease spreading between the two networks.

1. Agents: represented by the nodes in the network, each representing a city.

1.1. Infection status: Agents are infected by a continuous level of contagion $r_i \in (0, 1)$ where 0 means no one is affected in the concerned city or 1 where everyone is affected by the disease.

1.2. Connectivity to other agents: Agents are connected based on the air transportation network, where edges represent routes between cities and weight represents the daily number of flights.

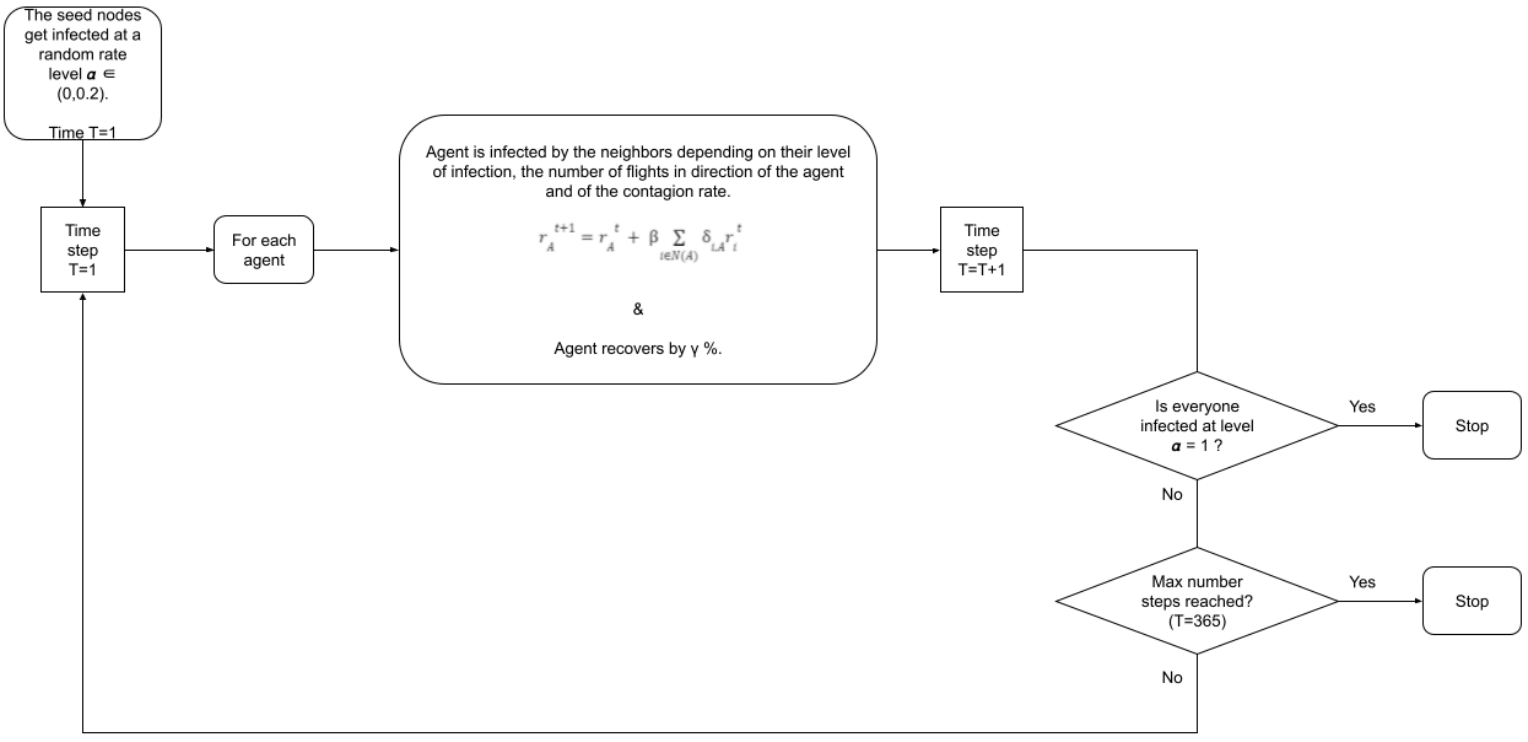
2. Initialization: The ABM begins with a set of initial conditions:
   2.1. Some agents are randomly selected to be infected. (15% of the nodes). Their level of infection is randomly selected from a uniform distribution $U(0, 0.2)$.
   2.2. Mobility patterns and connectivity between cities are based on the air transportation network data.

3. Contagion Rules: The contagion process follows a set of rules governing disease transmission and recovery:
   3.1. Transmission: At each step, the infection level of the agents is affected by a weighted average of their neighbors. The iterative step is described with the following formula:

$$r_A^{t+1} = r_A^t + \beta \sum_{i \in N(A)} \delta_{i,A} r_i^t \ , \ where \ \delta_{i,A}: \frac{number\ of\ daily\ flights\ from\ i\ to\ A}{maximum\ number\ of\ daily\ coming\ flights\ over\ all\ the\ cities\ airports}$$

   3.2. Recovery: Infected agents have a probability of recovering from the disease based on the recovery rate γ = 0.01.
   3.3. Termination Condition: The simulation continues until all nodes are fully infected or we reach the 365 steps (we assume that the contagious state is updated once a day, and we will look at how it spreads for a year)

This is the flow-chart of our ABM:



Propose one parameter to vary in your simulations and explain why you chose this particular one and how you will vary it. (max 200 words)

To initiate the simulations, we will start with a low transmission rate, β, such as β = 0.05, to represent a low contagious disease. We will observe and compare how this disease spreads in the air transportation networks of 2019 and 2021.

Next, we will repeat the simulations with a medium transmission rate, β, such as β = 0.2, to represent a moderately contagious disease. Again, we will compare the spread of the disease between the two networks.

Finally, we will choose a high transmission rate, β, such as β = 1, to simulate a highly contagious disease. We will repeat the simulations and analyze the differences in disease spread patterns between the networks in 2019 and 2021.

By varying the transmission rate and observing its effects on disease spread, we can gain insights into the impact of air transportation networks on the spread of contagious diseases and assess the effectiveness of control measures implemented during the COVID-19 pandemic.

4. Data collection

Describe your data. Which real network did you select, and how did you collect and store the data? Provide a graph visualization and three descriptive measures of the network that are useful for your analyses. (max 300 words)

The data used for this project is from the Bureau of Transportation Statistics from the United States Department of Transformation. We downloaded the data month-by-month and we then merged the different .csv files per year for 2019 and 2021.

The datasets contain the following variables: FlightDate, Reporting_Airline, OriginAirportID, OriginCityMarketID, Origin, OriginCityName, OriginSate, OriginStateName, DestAirportID, DestCityMarketID, Dest, DestCityName, DestState, DestStateName, DepTime, ArrTime. We decided only to keep the variables OriginCityName and DestCityName. We created an additional variable for the weights. To do so, we counted the number of times a route was made during the year.

Then the data is structured as two directed graphs, one for 2019 and another one for 2021. The graph without restriction has 352 nodes (one node corresponds to a city of departure, there can be multiple airports for one city) with 6 039 routes and the one with restriction has 347 nodes with 6 266 routes. This means that there are fewer active airports with restrictions than without. Also, there are more possible routes with restrictions, which might not look intuitive. The diameter of the graph without restriction is 4153 compared to 3158 for the one with restriction. Also, the graph is very transitive according to the CUG test we performed.

Let’s visualize a subpart of it:

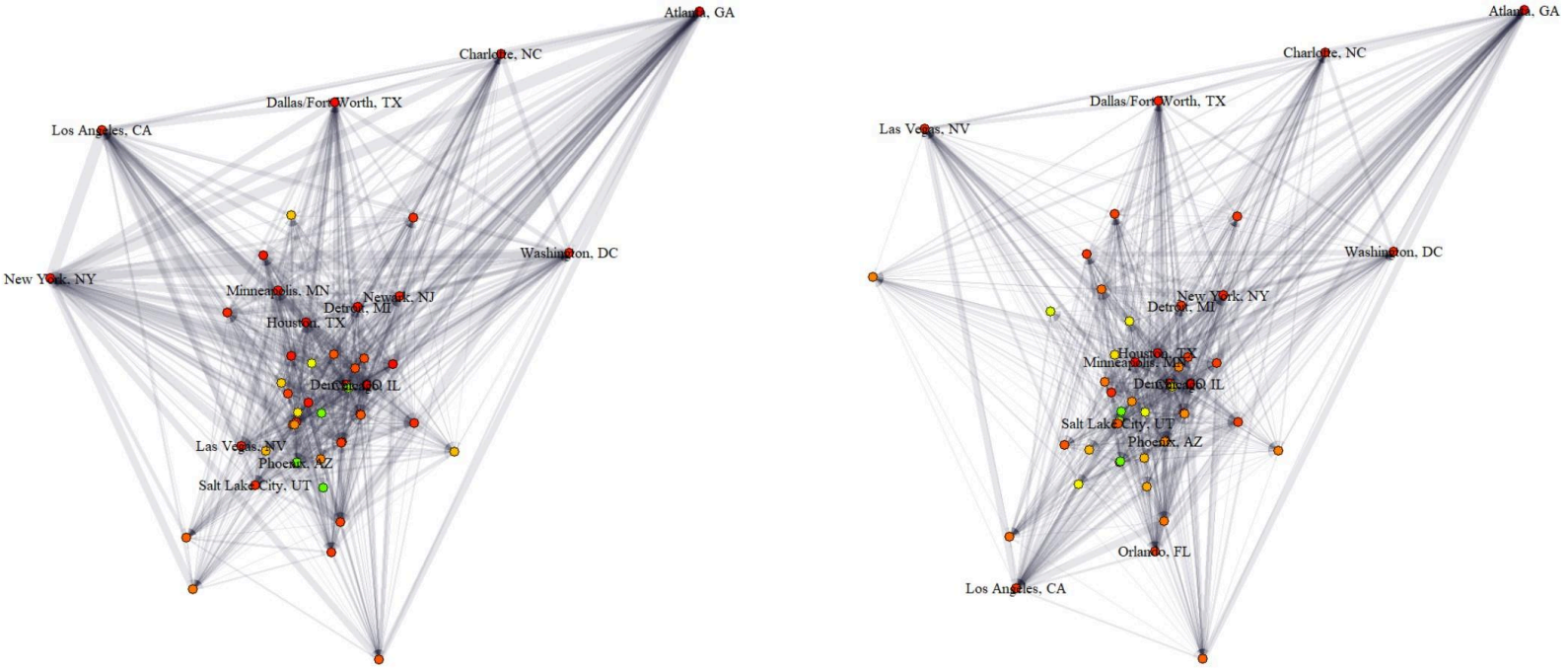


Figure 1: US air transportation subgraph in 2019 (left) and 2021 (right).

We are not able to visualize the full graph since it looks like a hairball, thus we decided to represent only the subgraph that corresponds to the top 50 cities in terms of degree. We use a kamada.kawai layout which plots nicely this graph. The width of the edges corresponds to the number of yearly flights. The gradient color of the nodes corresponds to their degree. Green corresponds to a low degree of centrality and red to a high degree. We see that the number of flights has clearly decreased between 2019 and 2021. This is also what we conclude by looking at the degree distribution :

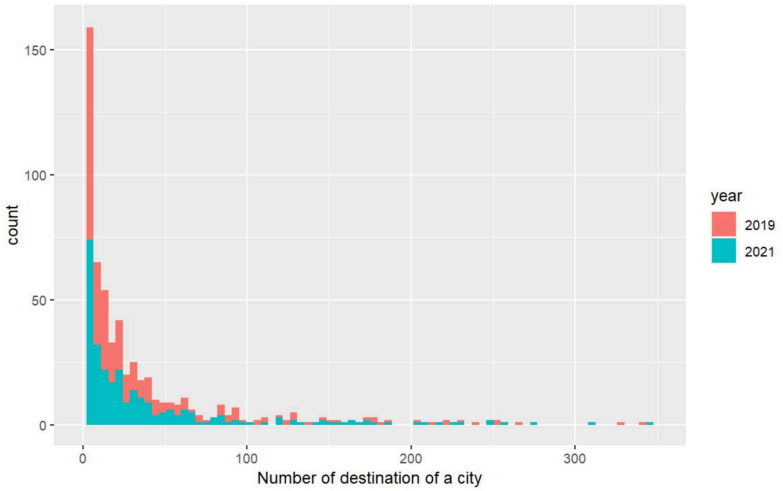


Figure 2: Degree distribution for 2019 - 2021 graphs

We see that the degree distribution is heavy-tailed in both graphs. There is no clear difference between both graphs.

5. Exploration and Analysis

Compute your centrality measure. Describe its distribution and provide a graph visualization of the distribution of these measures in the network. (max 200 words)

As previously mentioned, we decided to use the Strength (weighted vertex degree) centrality measure. Its distribution is described in Figure 3:

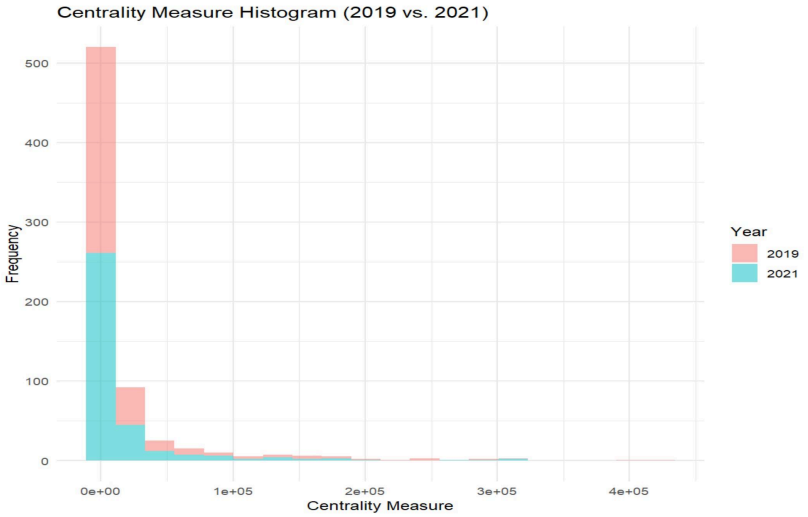


Figure 3: Strength centrality distribution for 2019 - 2021 graphs.

As we can see, the centrality measure is heavy-tailed for both graphs. There is no clear difference between both distributions. Let's look at the graph visualization of the distribution :

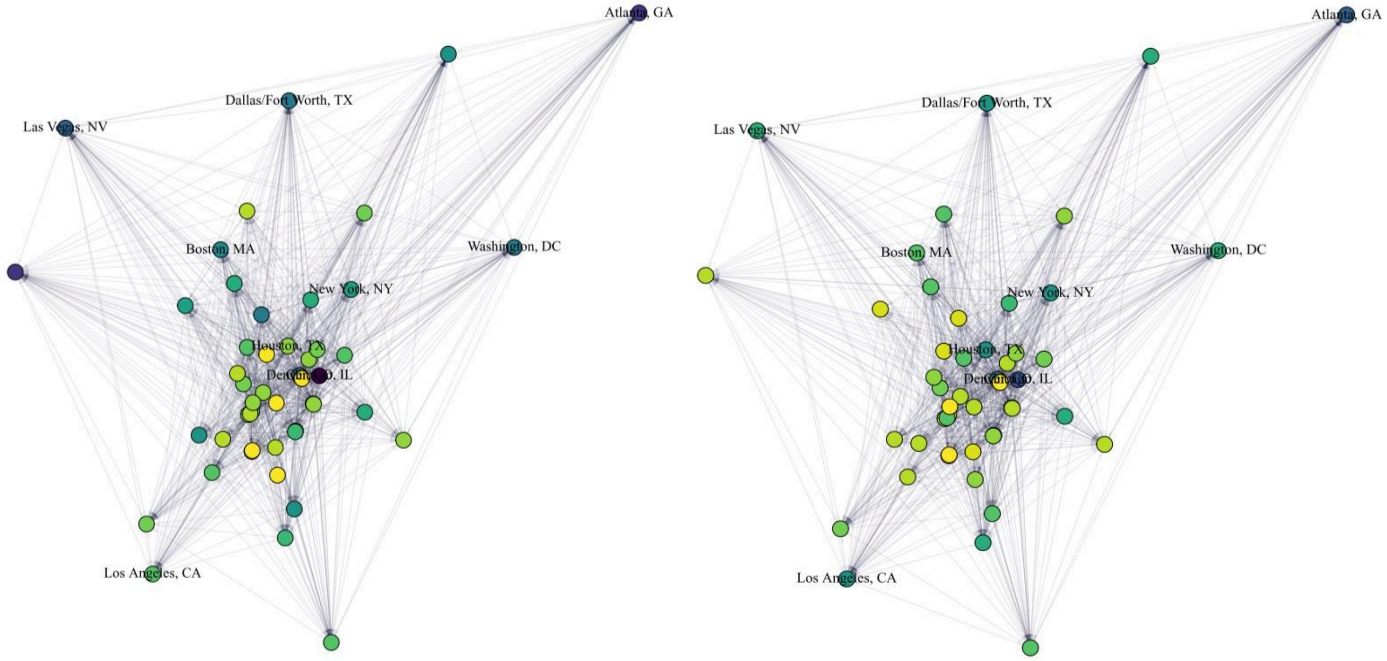


Figure 4: US air transportation subgraph in 2019 (left) and 2021 (right).

Figure 4 shows the 2019 and 2021 subgraphs of the US air transportation which corresponds to the top 50 cities in terms of degree. The color corresponds to the strength centrality. The gradient of color goes from yellow (low centrality) to purple (high centrality). We can see that the air transportation restrictions had an effect. Most of the cities are less central in 2021 than in 2019 to control COVID-19 spreading.

Describe the outcome of your simulations for different values of the varying parameter, with descriptive statistics and plots. (max 300 words)

In this section, we want to model the spread of a virus among the two networks and see if the virus spreads quicker in 2019 than in 2021. This will allow us to see if the flight restrictions during the pandemic of COVID-19 were effective (we are not evaluating other restrictions such as shielding measures or vaccination restrictions). We are going to vary the transmission rate and see if it has an impact on the transmissions, here are our results. The following plots display the evolution of diseases (more or less contagious depending on beta) in the graphs with and without the COVID-19 air transport restrictions.

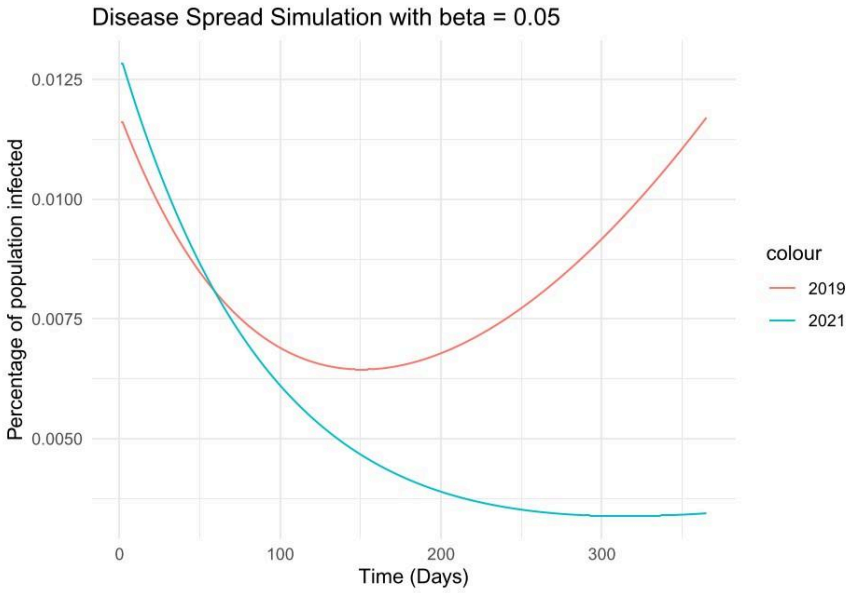


Figure 5: Evolution of a low contagious disease with and without air transportation restrictions

Figure 5 shows that the restrictions have a huge effect on the spreading of the disease. We see that for low contagious diseases, the infection is stopped in one year with the covid 19 restrictions. However this is not the case without the restrictions, the spread slows down and then comes back, maybe due to the infection of a hub. However, the level of people infected is less than 1.5% since it is not a contagious disease. Thus we can claim that the air transport restriction is very effective in stopping low contagious diseases. Let's look at the spreading of medium contagious diseases where beta equals 0.2.

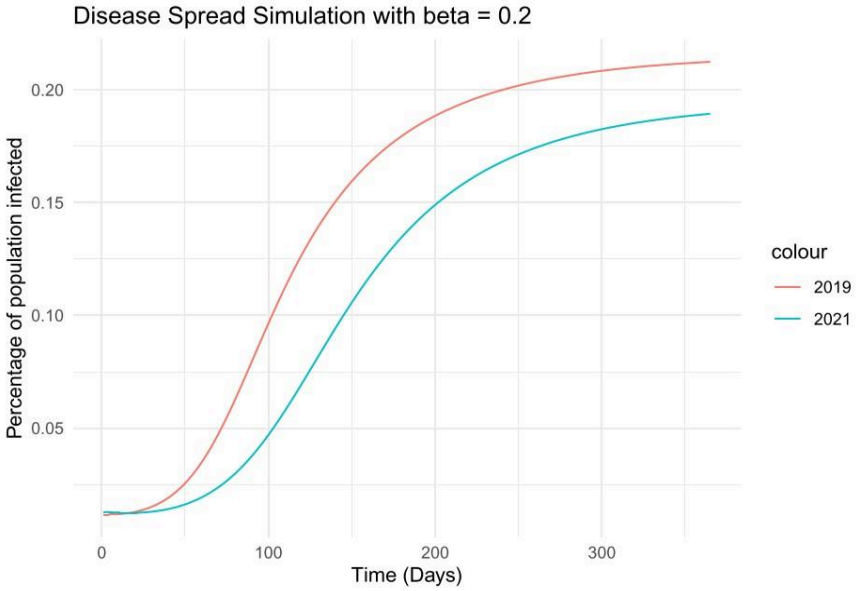


Figure 6: Evolution of a low contagious disease with and without air transportation restrictions

Figure 6 shows that the restrictions affect the spreading of the disease. We see that for medium contagious diseases (beta =0.2) the infection is stabilized in one year with and without the covid 19 restrictions. As we can see, the restrictions help to slow down the spread of the disease and to reduce the amount of people infected. For example, without the restrictions, 10% of the population is infected after 100 days, with the restrictions, 10% of the population is infected after 150 days. Thus we can say that the air transport restriction is effective in slowing down a medium contagious disease. Let's look at the spread of very contagious diseases where beta equals 1.

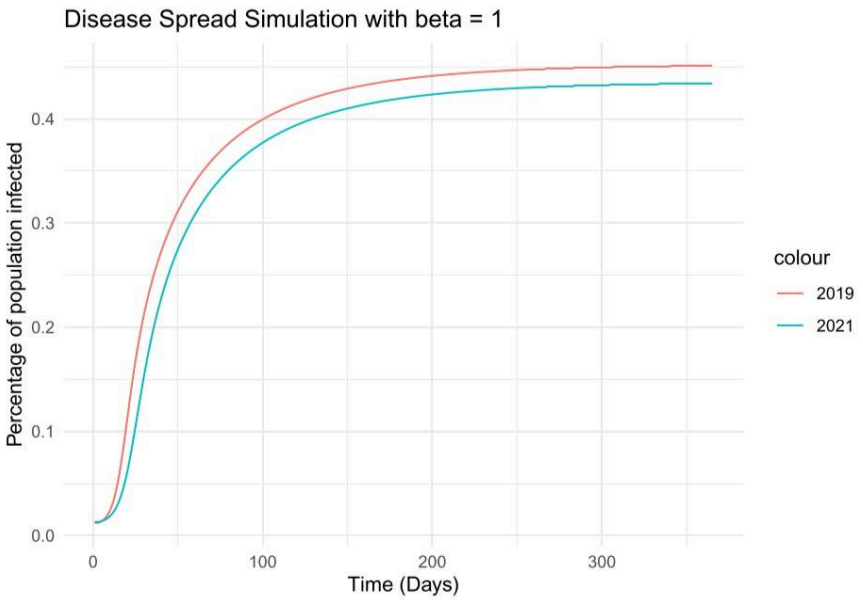


Figure 7: Evolution of a very contagious disease with and without air transportation restrictions

Figure 7 shows that the restrictions have a very small effect on the spreading of the disease. The spread of the disease is a bit slowed down but this is not very significant, other restrictions are needed in case of very contagious diseases. However, the structure of the graph is such that even if we have a very contagious disease, it stabilizes at 40 % of the population. This might be due to the disease spreading mainly on hubs and the cities with low air traffic are less impacted. Thus, for very contagious diseases, the COVID-19 restrictions are not enough.

How well does the centrality measure you defined predict how individuals are affected by the contagion process? You can choose the statistic and/or visualization that best shows this. (300 words)

The strength centrality measure describes perfectly how individuals are affected by the contagion process as we can see in the following visualization. The following figures display the evolution of a disease (more or less contagious) through the different cities of our network (except the seeds). The gradient of color represents the strength centrality where light colors represent low strength and dark colors represent high strength centrality.

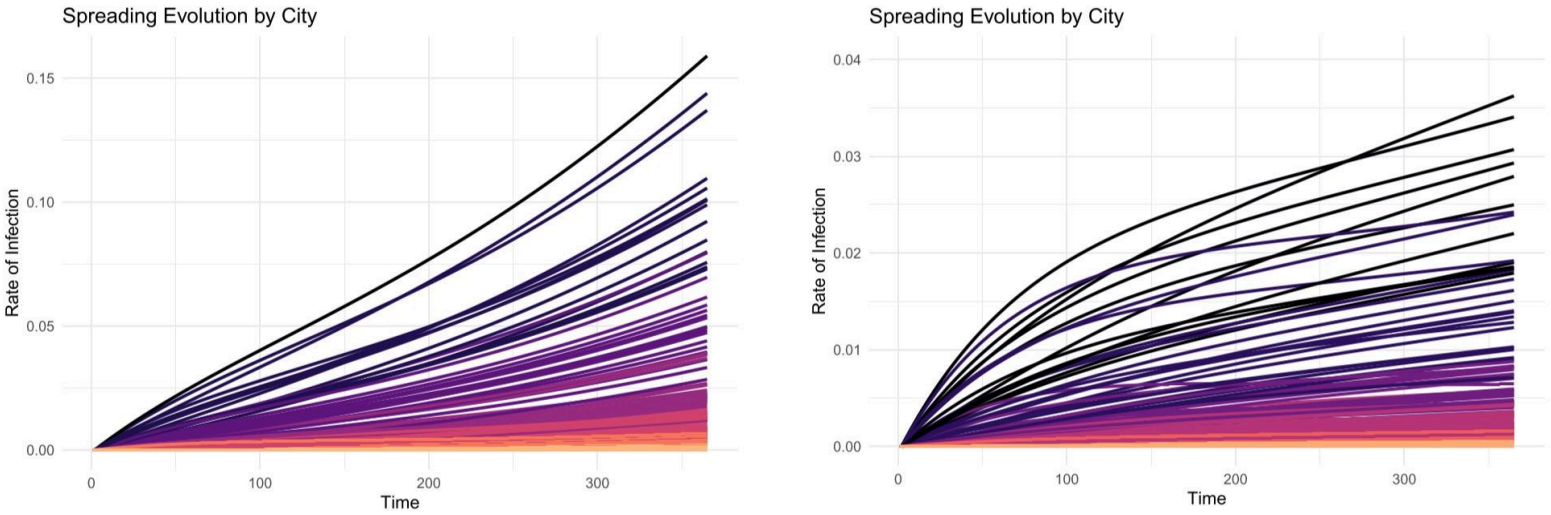


Figure 8: Evolution of a low contagious disease with (right) and without (left) air transportation restrictions

As we can see, the hubs, represented in dark colors, are the nodes that propagate the virus and the ones that lose control of the disease. The cities with low centrality measures (the one in orange) are not affected by the spread of the low contagious diseases. Thus, the strength centrality has a clear predictive power in the spread of diseases.

Also, this visualization is interesting, without restrictions, more than 15% of the major hub population is affected compared to less than 4% with restrictions. Also, the restrictions stabilize the spread of the disease in the hub compared to the case without restrictions where the hubs lose control of even low contagious disease.

We can display the same visualization for medium contagious diseases where we observe the same phenomenon.

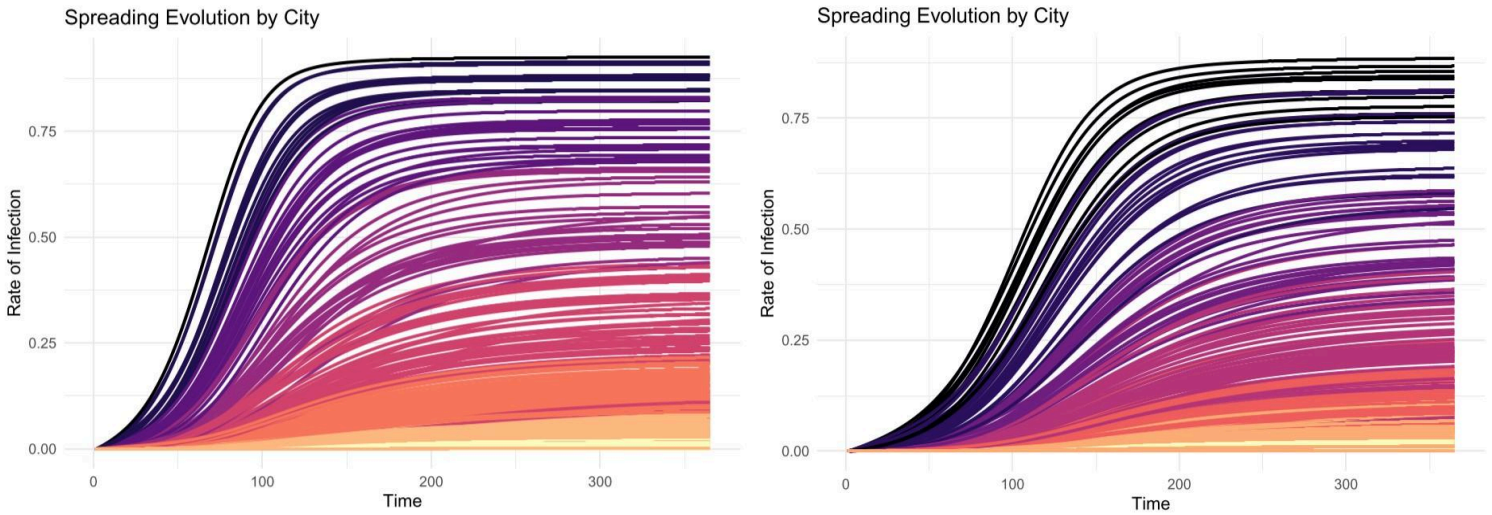


Figure 9: Evolution of a medium contagious disease with (right) and without (left) air transportation restrictions

As we can see, the hubs, represented in dark colors due to their high strength centrality, are the nodes that propagate the virus and the ones that lose control of the disease. The cities with low centrality measures (the one in light yellow) are not affected by the spread of a medium contagious disease. Again, the strength centrality has a clear predictive power in the spread of the disease. We do not display the equivalent visualization for very contagious diseases since it does not bring more information, however, you can find them in the Rmarkdown.

6. Interpretation and conclusions

Summarize what we learn from your simulations. Are they all in line with what you expected? (max 300 words)

Regarding the dynamic of the contagion, our results are mostly in line with what we expected.

In the context of air traffic restrictions, low contagious diseases are quickly stabilized and the most infected cities are the central ones. We can already observe the importance of hubs even for low contagious diseases. However, the disease spreads through the hubs without restrictions and does not stabilize. Slowly, the noncentral cities are affected and the disease spreads over the network.

Medium contagious diseases spread faster than low contagious ones. It spreads first through the hubs and then propagates to less central nodes as expected. However, the contagion is stable at some points, we do not observe a fully infected network. The comparison between the restrictive and nonrestrictive context leads us to conclude that the restriction slows down the speed at which the diseases propagate to the network. Also, the effect is very important for the cities that are between the noncentral and the most central ones.

Highly contagious diseases, like the other ones, spread first through the hubs and then propagate to noncentral cities. However, the diseases spread very quickly. In this context, the air traffic restrictions do not have a big impact. Only a few days are gained to limit the spread of the diseases. Thus, as expected, highly contagious diseases spread in an uncontrolled manner in both cases. Also, here

again, the contagion is stable at some points, we do not observe a fully infected network. However, the hubs are fully infected and they are the most at risk.

Finally, this analysis shows the predictive power of the strength centrality. The diseases spread by the most central cities and the strength effectively describe this centrality considering both the number of routes and the number of flights the route has.

Discuss the limitations of this study and identify possible ways to improve or enrich your study. Could this type of analysis be used to provide insight into or answer real-life problems? (max 200 words)

This study has limitations and possible ways of improvement. First, the agent-based model is not representative of reality because of the hypothesis we used to simplify the problem :

- The agent can only be affected by the situation of its neighbors.
- The same demography for all the cities.
- The spreading process follows a linear relationship.
- The cities communicate only via air transportation.
- There are only air transport restrictions to stop the disease.
- Assumption value for the contagion level, the recovery level.

To improve the study, we should read the literature about this topic to see other approaches and ways to overcome this kind of hypothesis we made.

An analysis of this type could be used to answer a real-life problem. It would help to mitigate the risk under unseen situations. Also, it would help to prepare concrete measures beforehand to anticipate and react as quickly as possible to events depending on their gravity.