30/05/23

Challenge #2 Report

Challenge 2 - Fight COVID-19 outbreak - Graph Theory

Team 1

Antoine HORRER, Kadriye Nur BAKIRCI

Contents

Business Understanding

Problem Statement

Background

During an epidemic or pandemic, implementing strict public health measures, such as lockdowns or business closures, can help control the spread of the virus but may have adverse effects on businesses, leading to job losses, economic downturns, and other socio-economic consequences. On the other hand, prioritizing economic activities and reopening businesses without appropriate measures can result in increased infections and strain on healthcare systems. The problem's background revolves around the challenge of balancing public health and the economy in the context of an ongoing epidemic or pandemic, these two notions evolve in opposite directions. The more we save the population by confining them, the more business suffers. It acknowledges the need to limit the spread of the virus to protect people's health while also considering the impact on businesses and the overall economy.

Stakeholders

The stakeholders are the national government authorities, health ministry, health minister, the president de la république France, employers, employees, and household members.

Risks and Assumptions

In our problem the possible risks could be:

- The risk of increased infection rates and potential strain on healthcare systems if the epidemic is not effectively controlled.
- The risk of economic downturn, job losses, and financial instability due to prolonged closures or restrictions on businesses.
- The risk of incomplete or inaccurate data, which can affect the reliability and effectiveness of modeling and decision-making.

to solve this specific problem our assumptions include the followings:

- The assumption is that the provided datasets are accurate and representative of the population and contact networks.
- The chosen epidemic model and its parameters accurately reflect the dynamics of the disease spread.

Objectives

The specific objectives are to:

1. *Minimize the number of Infected Clusters*

Set a threshold for the maximum allowable size of infected clusters within the contact network graph.
Measure the number of clusters exceeding the threshold.
Aim to reduce the number of clusters above the threshold to effectively limit the spread of the virus.

2. *Limit the Disease Transmission between Schools and Workplaces*

Define a threshold for the maximum allowable number of contacts between individuals from different schools or workplaces.

Measure the number of contacts exceeding the threshold between different schools or workplaces.

Focus on minimizing inter-school or inter-workplace contacts to prevent large-scale transmission across these settings.

3. *Maximize Business Continuity while Ensuring Safety*

Strive to keep the number of infected individuals within businesses or sectors below the threshold to enable economic activity while prioritizing safety.

Problem Translation

From a business problem to data analysis

Data Analysis Problem: Analyzing the graphs to minimize the spread of the virus while maximizing business continuity and protecting people's health.

Infected Clusters:

Data: Contact network (household click) graph representing interactions between individuals.

Problem: Identify clusters of infected individuals within the graph.

Objective: Minimize the number of infected clusters above a defined threshold.

Approach: Analyze the graph to detect connected subgraphs (clusters) where infections occur and measure the number of clusters exceeding the threshold. After determining the number of clusters exceeding the threshold, and implement the epidemic model. Develop strategies to reduce the number of clusters above the threshold, effectively limiting the spread of the virus.

Disease Transmission between Schools and Workplaces:

Data: Contact network weight graph with nodes representing schools, workplaces, and their respective connections.

Problem: Identify contacts between individuals from different schools or workplaces.

Objective: Limit the number of inter-school or inter-workplace contacts to prevent large-scale transmission.

Approach: Analyze the graph to detect edges (connections) between different schools or workplaces. Measure the number of contacts exceeding a defined threshold. Implement an epidemic model. Develop strategies to minimize inter-school or inter-workplace contacts and prevent disease transmission across these settings.

"Business Continuity and Safety:

Data: Contact network graph with nodes representing people and connected if their are working in the same company or living in the same house

Problem: Identify the sectors most conducive to virus propagation

Objective: Advise these sectors so that they have a better approach to the pandemic and can act to limit its spread

Approach: Analyze the graph to detect the most influential people sectors. Implement preventive measures within businesses (e.g., testing, mask-wearing, hygiene protocols) to reduce infections. Encourage remote work or flexible scheduling to minimize in-person interactions.

By framing the business problem within the context of graph theory, we can leverage network analysis techniques to analyze the contact network graph, identify patterns, measure the impact of interventions, and develop strategies to balance public health measures and the economy effectively.

Determining the threshold before implementing the epidemic model:

Advantage: Setting the threshold before implementing the epidemic model allows us to define a specific target or goal for the analysis. It provides a clear criterion for evaluating the effectiveness of the implemented measures.

Consideration: The chosen threshold should be based on expert knowledge(this time imaginary), available data, and public health guidelines. It should align with desired outcomes and strike a balance between public health and economic considerations.

In this approach, we would define the threshold for the maximum allowable size of infected clusters, the maximum allowable number of contacts between different schools or workplaces, or the maximum allowable number of infected individuals within businesses. Once the threshold is defined, we can then implement the epidemic model and analyze the graphs that we create, to measure the number of clusters, contacts, or infected individuals exceeding the threshold at different time points.

Why should we use graph theory?

Graph theory is important for the problem of balancing public health and the economy during an epidemic because it allows us to analyze the contact networks between individuals, households, workplaces, and schools. By representing these networks as graphs, we can identify the patterns of interactions and the potential paths of disease transmission. Also, it helps in identifying critical nodes or hubs within the contact networks that play a significant role in the spread of the virus. By analyzing centrality measures, such as degree centrality, betweenness centrality, or eigenvector centrality, we can pinpoint key individuals, locations, or sectors that have a higher influence on the spread of the virus. Finally, using graph theory, we can simulate the effects of various interventions, such as lockdowns, and business closures. By modeling the interactions and connections within the graph, we can evaluate the potential impact of these interventions on the spread of the virus, as well as on the economy.

Data Understanding

Dataset Description

In our problem, we have 3 main datasets which are “household”, “pro_contacts_adults” and “pro_contacts_children”.

Household dataset:

	household_id	nb_children	nb_adults	type	size
0	0	2	2	two_parent_family	4
1	1	1	2	two_parent_family	3
2	2	0	3	two_parent_family	3
3	3	2	2	two_parent_family	4
4	4	2	2	two_parent_family	4

In this dataset, household_id represents families in the France region. The other column attributes represent information about each family. We have 3802 rows and 5 columns. Also, in this dataset, we don’t have any null values.

Pro. contacts adults:

	household_id	adult_id	job_cat	pro_contacts	company_id
0	0	0	Indus_other	52	6
1	0	1	Hotel_Restaurant	114	21
2	1	2	Shops_other	14	45
3	1	3	Administration_schools	33	10
4	2	4	Services_other	770	34

This dataset reflects the business contacts of adults for each family. From this dataset, we can get the job category, how many contacts they have, and the company’s id information for each individual in each family. We have 6960 rows and 5 columns. Also, in the job category and company id columns we have 2297 null values.

Pro. contacts children:

	household_id	child_id	school_contacts	school_id
0	0	6960	35	2

1	0	6961	58	6
2	1	6962	92	5
3	3	6963	25	9
4	3	6964	85	1

This dataset reflects the school contacts of children for each family. From this dataset, we can get the children's id, how many contacts they have and school id information for each individual in each family. We have 3085 rows and 4 columns. Also, in this dataset, we don’t have any null values.

Data Preparation and Create Graphs:

Objective 1 aims to set a threshold for the maximum allowable size of infected clusters within the contact network graph and measure the number of clusters exceeding this threshold. The goal is to reduce the number of clusters above the threshold to effectively limit the spread of the virus. In order to get this goal, first we want to create a household clique graph. To create a graph we implement some data preparation steps:

1. Choose household_id and adult_id from the pro_contacts_adults dataset, and household_id and child_id from the pro_contacts_children dataset then change the adult_id and child_id name to node_id.
2. Concat these two datasets along the vertical axis.
3. To get the nodes in each household, and build edgelists such as nodes in the same household form a clique, group the concatenated dataset under the household_id column and to get edgelist, iterate over each group, it generates combinations of node indices within each group using the “combinations()” function, and appends the pairs of node IDs to the “edgelist” list.
4. Create a graph by using created nodes and edgelist.

After implementing these data preparation steps, we have the graph with 8980 nodes and 12519 edges.

Objective 2 aims to define a threshold for the maximum allowable number of contacts between individuals from different schools or workplaces and measure the number of contacts exceeding this threshold. The focus is on minimizing inter-school or inter-workplace contacts to prevent large-scale transmission across these settings. The aim is to reduce the potential for the virus to spread between schools and workplaces, thereby containing its transmission. In order to get this goal, first we want to create a contact-weighted graph. To create a graph we implement some data preparation steps:

1. Create an empty graph.
2. Retrieve the unique school IDs from the 'school_id' column of the 'pro_contacts_children' dataset. Then create a loop to add a node to the graph with the label "School {school_id}" and a node attribute 'type' with the value 'school'.
3. Retrieve the unique company IDs from the 'company_id' column of the 'pro_contacts_adults' dataset. Then create a loop to add a node to the graph with the label "Workplace {company_id}" and a node attribute 'type' with the value 'workplace'.
4. To add edges representing contacts between adults and workplaces with weights, create a loop that iterates over each row in the 'pro_contacts_adults' dataset, and for each row, household_id,

company_id, and pro_contacts are extracted from the row. Then add an edge between the workplace node labeled "Workplace {company_id}" and the household node labeled "Household {household_id}". The edge has a weight attribute with the value pro_contacts.

5. To add edges representing contacts between children and schools with weights, create a loop that iterates over each row in the 'pro_contacts_children' dataset, and for each row, household_id, school_id, and school_contacts are extracted from the row. Then add an edge between the school node labeled "School {school_id}" and the household node labeled "Household {household_id}". The edge has a weight attribute with the value school_contacts.

In the end, the graph is populated with nodes representing schools, workplaces, and households, and edges representing contacts between adults and workplaces, as well as contacts between children and schools. Also, the graph has 4064 nodes and 9399 edges with an average degree 4.625492125984252.

Strategies and Epidemic Model:

For the first objective, first, we set a threshold (threshold = 3) for the minimum number of nodes in a cluster and then find the connected components in a graph. Later filter out clusters that exceed the threshold and count the number of such clusters which is 913. Here our aim is to reduce the number of clusters above the threshold to effectively limit the spread of the virus like implementing lockdowns, putting distance between people, extending the vaccine operation, etc. In order to reduce the spread of disease and the number of clusters exceeding the threshold, we develop an edge removal strategy. Our strategy is to identify highly influential or central nodes in the graph, such as those with high degrees or betweenness centrality. Targeted removal of these nodes can disrupt the spread of the disease by breaking connections between different clusters or reducing the overall connectivity of the graph. After implementing this strategy the number of clusters that exceeds the same threshold reduces to 379. Before and after implementing the strategy we also implement an epidemic model to see how effective our strategy is. We will discuss model results in the evaluation section. We run an epidemic model for 10 simulations and store the results in a dataframe. For each simulation:

- Initialize the infection status for all nodes in the graph.
- Select a random subset of clusters that exceed the threshold.
- Randomly choose an initial infected node from each selected cluster.
- Run the epidemic model using the EoN library, specifying the graph, transmission rate (tau, 0.05), recovery rate (gamma, 0.02), initial infected nodes, and maximum simulation time (tmax = 100).
- Calculate the number of deaths, infectious rate, and recovery rate based on the model outputs.
- Store the simulation results in a dictionary and append it to the results list.

For the second objective, first, we set a threshold value (threshold = 10) for maximum allowable contacts and then measure the number of contacts in the graph that exceed this threshold which is 6600. Here our aim is to focus on minimizing inter-school or inter-workplace contacts to prevent large-scale transmission across these settings like in the first objective but the difference between first and second, in the first objective we want to reduce the spread of disease in the household while in the second objective we want to reduce the spread of disease inter-school or inter-workplace that's why we decided to generate two different graphs and implement two strategies for different scenarios. In order to reduce the number of contacts in the graph while considering the spread of disease is by targeting nodes with higher degrees of centrality. The degree centrality of a node measures the number of connections it has with other nodes in the graph. We set a threshold for the degree centrality (0.05) and then identify nodes with degree centrality above the threshold. Later remove the identified nodes and their associated edges from the graph finally recalculate the number of contacts exceeding the threshold after removing nodes which is 3707. Before and after implementing strategy we also implement an epidemic model to see how effective our strategy is. We will discuss model results in the evaluation section. We run an epidemic model for 10 simulations and store the results in a dataframe. For each simulation:

- Randomly selects 100 nodes from the graph as the initially infected nodes
- Run the epidemic model using the EoN library, specifying the graph, transmission rate (tau, 0.5), recovery rate (gamma, 0.02), initial infected nodes, and maximum simulation time (tmax = 100).
- Calculate the number of deaths, infectious rate, and recovery rate based on the model outputs.
- Store the simulation results in a dictionary and append it to the results list.

Business Continuity and Safety:

Here's the strategy and our method

Loading the data: We began by loading the data from a CSV file containing information on individuals, their sector of activity, their place of work, and their place of residence. This provided us with the basic information we needed to create our graph.

Creating the graph: We used the NetworkX library to create a graph. Each node in the graph represents a person, identified by their ID, and the edges represent the links between people. Edges can represent either work links between people sharing the same workplace (same workplace ID), or residence links between people sharing the same home (same home ID)

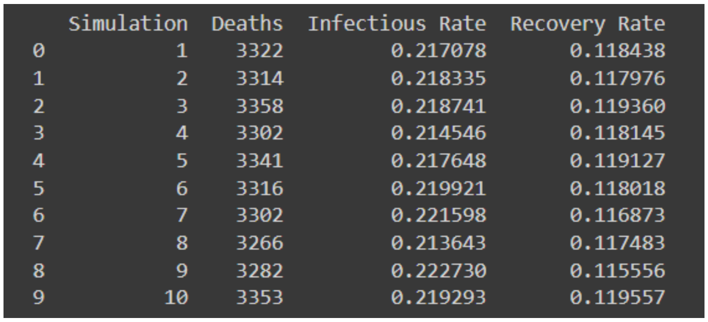
Community detection: We used the community detection algorithm to identify groups of people who are strongly connected to each other. This enabled us to identify workplaces or households where there is a high concentration of contacts, which can facilitate the spread of disease.

Identifying the most influential nodes: We calculated the degree centrality for each node in the graph, enabling us to measure the influence of nodes. Persons with a high degree of centrality are those with the greatest number of connections with other people, meaning they affect a large number of individuals.

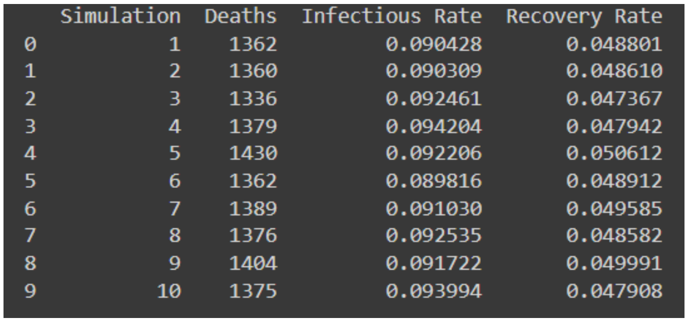
Industry analysis of influential persons: Using the IDs of the most influential companies, we extracted the corresponding business sectors from the data. This enabled us to identify the types of companies that have the greatest impact on individuals, and to make appropriate health decisions based on these industries.

Evaluation

Minimize the number of Infected Clusters:



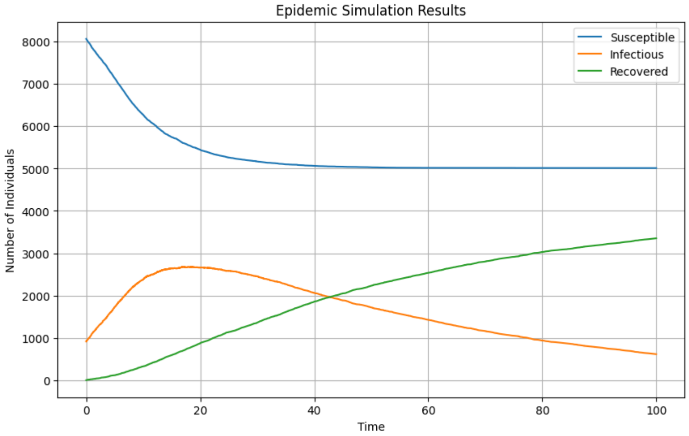
```
   Simulation  Deaths  Infectious Rate  Recovery Rate
0           1    3322         0.217078       0.118438
1           2    3314         0.218335       0.117976
2           3    3358         0.218741       0.119360
3           4    3302         0.214546       0.118145
4           5    3341         0.217648       0.119127
5           6    3316         0.219921       0.118018
6           7    3302         0.221598       0.116873
7           8    3266         0.213643       0.117483
8           9    3282         0.222730       0.115556
9          10    3353         0.219293       0.119557
```

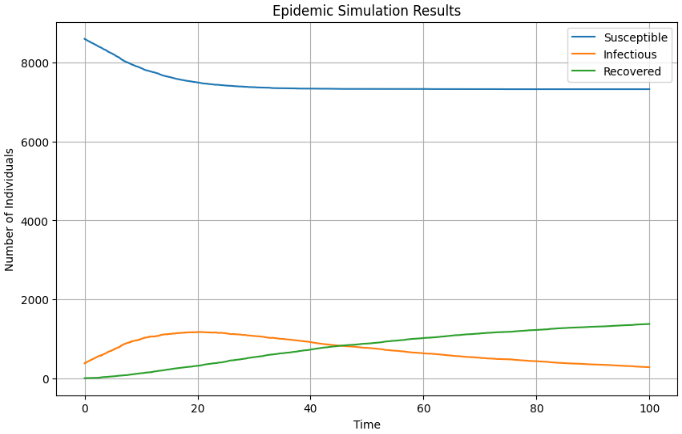


```
   Simulation  Deaths  Infectious Rate  Recovery Rate
0           1    1362         0.090428       0.048801
1           2    1360         0.090309       0.048610
2           3    1336         0.092461       0.047367
3           4    1379         0.094204       0.047942
4           5    1430         0.092206       0.050612
5           6    1362         0.089816       0.048912
6           7    1389         0.091030       0.049585
7           8    1376         0.092535       0.048582
8           9    1404         0.091722       0.049991
9          10    1375         0.093994       0.047908
```

Before implementing strategy

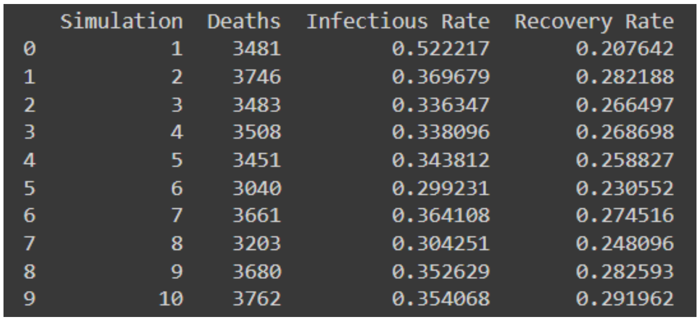
After implementing strategy



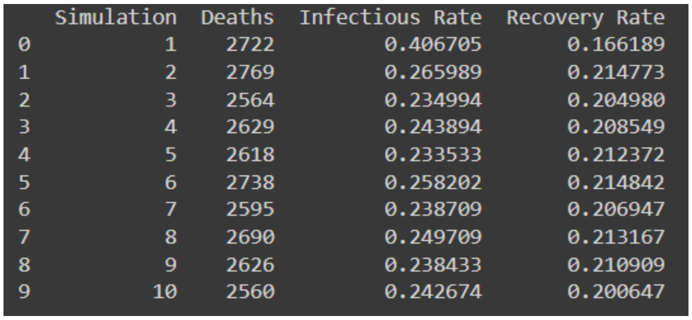


The aim of this objective is to reduce the number of clusters above the threshold to effectively limit the spread of the virus. To do this, an edge removal strategy is developed to identify highly influential or central nodes in the graph, such as those with high degrees or betweenness centrality. Targeted removal of these nodes can disrupt the spread of the disease by breaking connections between different clusters or reducing the overall connectivity of the graph. We can see the results of before and after implementing the strategy from above tables. As results show us that our strategy works pretty well because it reduces the number of deaths(also include recovered) and infectious rate. This strategy can be improved by changing thresholds or these results can be improved by changing hyperparameters or selecting initial infected clusters.

Disease Transmission between Schools and Workplaces:



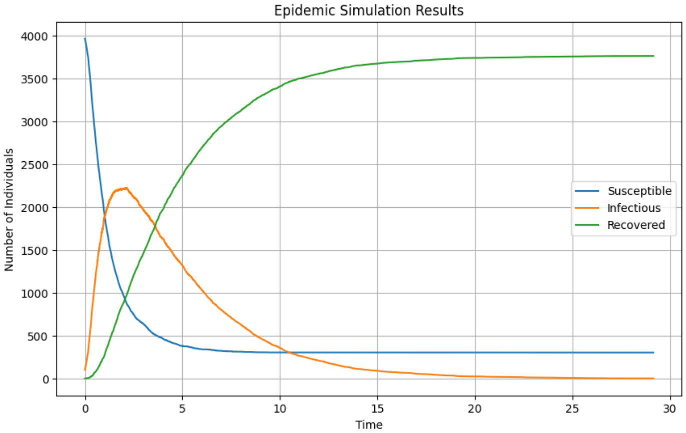
```
   Simulation  Deaths  Infectious Rate  Recovery Rate
0           1    3481         0.522217       0.207642
1           2    3746         0.369679       0.282188
2           3    3483         0.336347       0.266497
3           4    3508         0.338096       0.268698
4           5    3451         0.343812       0.258827
5           6    3040         0.299231       0.230552
6           7    3661         0.364108       0.274516
7           8    3203         0.304251       0.248096
8           9    3680         0.352629       0.282593
9          10    3762         0.354068       0.291962
```

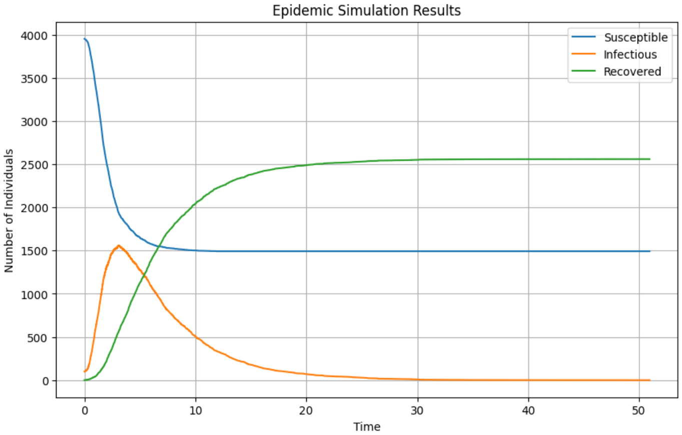


```
   Simulation  Deaths  Infectious Rate  Recovery Rate
0           1    2722         0.406705       0.166189
1           2    2769         0.265989       0.214773
2           3    2564         0.234994       0.204980
3           4    2629         0.243894       0.208549
4           5    2618         0.233533       0.212372
5           6    2738         0.258202       0.214842
6           7    2595         0.238709       0.206947
7           8    2690         0.249709       0.213167
8           9    2626         0.238433       0.210909
9          10    2560         0.242674       0.200647
```

Before implementing strategy

After implementing strategy

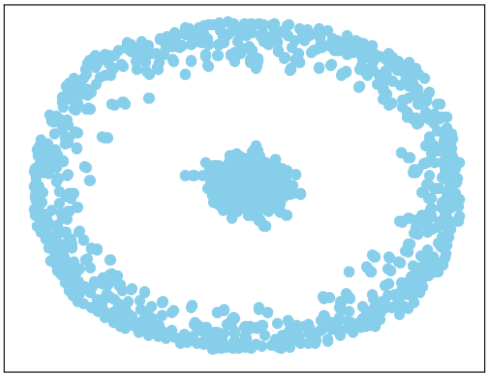




The aim of this objective is to focus on minimizing inter-school or inter-workplace contacts to prevent large-scale transmission across these settings like implementing lockdowns, putting distance between people, extending the vaccine operation etc. In order to achieve our target, we develop a degree centrality strategy. Targeting nodes with higher degrees of centrality can help reduce the number of contacts in the graph while considering the spread of disease. This measures the number of connections it has with other nodes. We can see the results of before and after implementing strategy from above tables. As results show us that our strategy again works pretty well because it reduces the number of deaths(also include recovered) and infectious rate. This strategy can be improved by changing thresholds or these results can be improved by changing hyperparameters or selecting initial infected clusters.

Business Continuity and Safety:

Here are our results and our recommendations



30/05/23

This is the graph of the relations between the nodes linked by their relation at work and at home.

The most influential nodes are:

```
  nodes_id
0      713
1     2436
2     3788
3      864
4     1019
5     1834
6     2059
7     2354
8     5295
9     5349
```

Those nodes are respectively linked to those activities:

```
 adult_id        job_cat
  713          Health
  864  Transportation
 1019          Health
 1834     Indus_other
 2059  Transportation
 2354           Health
 2436      Indus_food
 3788           Health
```

5295	Indus_food
5349	Indus_food

We can therefore conclude that the sectors most prone to virus propagation are the healthcare, transport and food sectors.

As the health sector is already highly regulated, and much better informed than we are about what to do and what not to do to transmit a virus, we have no specific recommendations. Nevertheless, the transport sector is key to the economy, and we must avoid confining employees in this area as much as possible. The use of a mask and hydroalcoholic gel therefore seems obvious. We can also suggest to carriers' customers that they limit their orders as much as possible, by ordering in large quantities for example, to limit the number of orders. For the food industry, we can make the same recommendation, and also perhaps try to eat more locally to limit the amount of food that has been transported by several people.