IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Miguel GUZMÁN
Camille FRANCOIS-MARTIN

May 30th, 2023

SNA: Challenge 2
Fight COVID-19 Outbreak

Problem Statement:

The following statements are entry records describing the evolution of the epidemic emergency due to COVID-19.

December 2019:

“2019 Novel Coronavirus (2019-nCoV, now called COVID-19) is a virus (more specifically, a coronavirus) identified as the cause of an outbreak of respiratory illness first detected in Wuhan, China in December 2019.”

23 January 2020:

“The confinement of the metropolis of Wuhan (center) and its province, Hubei, is decided. It is the beginning of an international crisis.”

31 January 2020:

“The World Health Organization (WHO in English, OMS in French) declares an international emergency on 31 January 2020. While the cost in lives is rising in China, with 213 deaths and 10,000 people infected, and 80 confirmed cases in 18 countries, the WHO is nevertheless urging that travel should not be restricted. The WHO stressed that there is no need to restrict travel and trade with China. ‘WHO (...) even opposes any travel restrictions,’ WHO's director stressed. Restricting the movement of people and goods during a public health emergency can be ‘inefficient’, disrupt the distribution of aid and have "negative effects" on the economies of affected countries, he said.”

3 February 2020:

“Ten days after the start of the crisis, marked by the confinement of the metropolis of Wuhan (center) and its province, Hubei, the Chinese stock exchanges of Shanghai and Shenzhen plunged by around 8% after a ten-day interruption in quotations. This was the biggest drop in Chinese indices since the stock market crash of 2015. And the international repercussions are obviously important from an economic point of view. In spite of WHO's recommendation to keep the travels possible, many countries are concerned and have increased their protective measures. The United States, Australia, New Zealand, Iraq, Israel and the Philippines in particular have banned foreigners who have recently visited China from entering their territory.”

24 Feb 2020 - Latest Entry:

“Panic is growing. You fear the worst health and economic crisis... In your crystal ball, you can see that the whole world is frozen. France will not be spared. People will die, the health care system will be pushed to its limits, or even worse, explode. People will struggle for food... and probably toilet paper... Should people be confined at home? When ? If yes, when? After how many cases/deaths? What would be the impact on the economy? Remote working is not always possible... Shops, potato growers, etc. need customers. A lot of people will lose their jobs, companies will face huge difficulties....”

Challenge Overview:

Business rationale - As members of the Health Ministry, the challenge is to propose strategies to save both people and the economy given the COVID-19 outbreak. While limiting the *diffusion* of the virus, our aim is to preserve business and health. In this report we propose strategies to the management department director, and measure their efficiency with key indicators. The objective is to enable him/her to make decisions, and then he/she will convince, in turn, the Health Minister and the President de la République himself.

1. Formal Objective

Let there be an undirected and unweighted graph defined as $G = (V, E)$ where *V* is the set of nodes describing individuals of a population and *E* their connections. *V is* made up from the union of two disjoint subsets $W \sqcup S$ each of them constructed from tabular datasets containing information from members coming from the working class *W* and student sector *S*. *E* is the set of edges constructed first by creating cliques between nodes $v_i, v_j \in V$ sharing a household, then by creating cliques within the nodes in *W* belonging to the same company and within the nodes of *S* belonging to the same school and finally by creating additional randomly-based connections among members of different affiliations (company / school) for each of the subsets.

Let I be defined as the number of COVID infected nodes given an epidemic model simulation at a time $t \in [0, t_{max}]$, $0 < t_{max} < \infty$. The objective is thus to find an optimal network configuration G^* following a strategy to handle I in such a way that we minimize both the number of deaths of the global population *V* at t_{max} and the economic impact of having a high number of members of the working class *W* in an inactivity state at the peak of infection, defined as the time with the highest number of infected nodes and denoted as t_{peak}, due to contamination.

2. Data Preparation

To begin with the analysis, graph models ought to be created given the provided tabular data.

The creation of *G* is pipelined under the following sequence of steps:

1. *Household Cliques*: Cliques are built among the members of the same household. Intuition is simple and direct: they all live in the same house and interact on a daily basis.

2. *Company & School Cliques*: Even though the members of a company/school may not belong to the same department/grade or have any friendship links, they still share the same *roof* and consequently breathe the same air. Thus, under this intuition we build cliques in the population connecting all the members belonging to the same company/school.

3. *Inter company / school connections :* Every person has a predetermined number of work or school connections (degree) defined by the tabular datasets. This is expressed by the *pro_contacts* feature found in the adults and children datasets. Some people have more work/school connections than people in their school or company. To fill that gap we have decided to add random connections between the person and persons from other companies / schools. This can simulate extra-work activities such as business meetings with other company, sports, etc.

The generated graph *G* representing the whole population contains 9699 nodes and 1244291 edges.

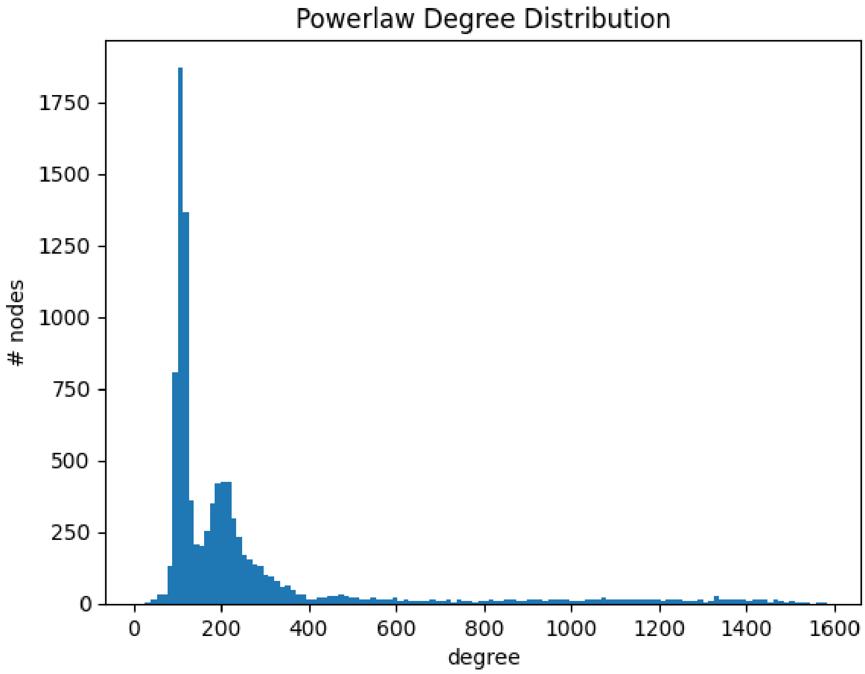

Figure 1. Node degree *k* distribution for *G*. The average node degree $\langle k \rangle$ is ≈256.58.

In addition to creating the graph *G*, a feature engineering task is required to compute the economic loss indicator expressed in Section 3. It consists of fitting all the job categories found in the business contacts dataset into the activity branches considered by INSEE [3]. The engineered 'Industry Sector' categorical feature notion is established in Table 1.

Activity Branch (New Feature)	Job Categories (Associated Categorical Values)
Agriculture and food industries	{'Indus_food', 'Agriculture_fishing', 'Shops_market_food'}
Industry excluding agri-food	{'Indus_other'}
Construction	{'Construction'}
Merchant services	{'Transportation', 'Shops_other', 'Hotel_Restaurant'}
Non-market services	{'Administration_schools', 'Services_other', 'Health'}

Table 1. Activity branch feature associated job categories.

3. Objective Function

3.1 Minimization Problem

As stated in Section 1, our objective is to minimize the number of deaths in the global population and the economic impact given the inactivity (or death) of members of the workforce. Consequently, our object of study is a bi-objective minimization problem defined as:

$$G^{*} = minimize\, f(G) \;=\; [f_1(G),\, f_2(G)]$$

Where f_1 represents the function that runs an epidemic model given an initial network configuration G and returns the total number of deaths in the population at t_{max}.

And where f_2 evaluates the same epidemic model as $f_1$but returns an economic loss indicator δ resulting from the weighted sum of the elements of a vector $\vec{y} = (y_1, y_2, y_3, y_4, y_5)$ and its respective fixed weights denoted as $\vec{w} = (w_1, w_2, w_3, w_4, w_5)$ at the peak of infections. Table 2 describes the features represented by every element of the aforementioned vectors.

Activity Branch	$\vec{y}$: number of inactive/infected people from *W* for given activity branch	$\vec{w}$: share in GDP (in %) for given activity branch in normal conditions
Agriculture and food industries	y_1	w_1 = 4% = 0.04
Industry excluding agri-food	y_2	w_2 = 12% = 0.12
Construction	y_3	w_3 = 6 = 0.06
Merchant services	y_4	w_4 = 56 = 0.56
Non-market services	y_5	w_5 = 22 = 0.22

Table 2. Economic loss indicator (δ) components description. The share in GDP estimates are taken from pre-pandemic estimations by INSEE [3].

Consequently, we can express:

$$f_2(G) = \delta = \sum_{i=1}^{5} w_i y_i, \; when\; t = t_{peak}$$

3.2 Epidemic Model Dynamics

The epidemic model that defines both objective functions follows the SIR model. This model divides the population into compartments. Each compartment is expected to have the same characteristics. SIR represents the three compartments segmented by the model.

- Susceptible
- Infectious
- Recovered

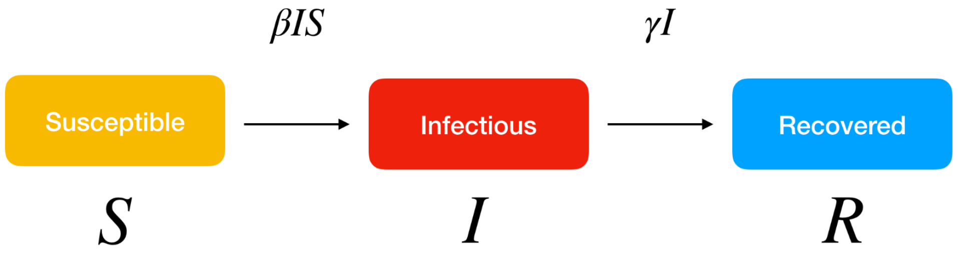

Figure 2. Compartments and transitions in the SIR model [8]

Susceptible (*S*) is a group of people who are vulnerable to exposure with infectious people. The group of ***infectious*** (*I*) represents the infected people. They can pass the disease to susceptible people and can be recovered in a specific period. ***Recovered*** (*R*) people get immunity (or die) so that they are not susceptible to the same illness anymore. SIR model is a framework describing how the number of people in each group can change over time [8].

Each susceptible individual gets the disease with probability β after each contact with an infected individual. Each infected individual has a probability γ of recovering (or dying) from the disease at each time step [5].

Defining the appropriate SIR model **parameters** for the epidemic simulation is not an easy task, yet it is **critical** in order to estimate an accurate diffusion dynamic. Sasaki proposes a model to estimate SIR parameters given COVID time series datasets per country [8]. However, even though his parameter estimations fit the COVID-19 dataset decently, the estimated parameters, R_0 in particular, seem extremely unrealistic. Manivel [4] estimated the characteristics of a classic SIR model which fitted best the actual deaths by COVID per 24 hours in hospital published by *Santé Publique France* between February 19th and July 1st of 2020. In his studies, R_0 is estimated to be 5.5 and β = 1.53E-06. However he mentions that the R_0 estimations should not be highly reliable and may assimilate more to the most cited value of French publications $R_0 \approx 3.0$. Other study estimates R_0 = 4.8 for the French population [7]. For our study we select the value of R_0 proposed by the challenge requirements since it is more coherent to the French publications apparent consensus.

Infection recovery rate should not have a strict correlation to geographic location. The infection recovery rate of our study corresponds to the median viral shedding period of 20 days described by Zhou [11]. Several studies concerned on predicting the diffusion dynamics of COVID in France agree with this commonized proposal [4, 7].

Our proposed SIR model parameters are then set as:

$$R_0 = 2.5$$
$$\gamma = 0.05 \text{ (20 days)}$$

From Social Network Analysis theory [5], we know that the basic reproduction number R_0 can be estimated by the following formula:

$$R_0 = \frac{\beta}{\gamma}\langle k \rangle, \ where\ \langle k \rangle\ is\ the\ average\ node\ degree\ in\ the\ graph$$

Solving for β, we obtain:

$$\beta = \frac{R_0\gamma}{\langle k \rangle}$$

We can note that the estimation of β relies heavily on⟨k⟩. Given this, we consider that in order to estimate a reliable β for a SIR simulation on *G*, we should compute β directly considering the real $\langle k \rangle$ of *G,* instead of using estimations from other models or studies whose node degree distribution may not resemble the one from our simulated demography. Thus, for graph *G*, β is estimated as:

$$\beta = \frac{R_0\gamma}{\langle k \rangle} = \frac{(2.5)(0.05)}{256.58} \approx 0.00048$$

The mortality rate of COVID in France, which we will denote θ, is obtained from the ratio within the number of confirmed cases and number of deaths caused by COVID in the country since the beginning of the outbreak. The data is the one reported by the World Health Organization as of May 28, 2023 [9].

$$\theta = \frac{\#\, of\, deaths}{\#\, of\, confirmed\, cases} = \frac{163{,}437}{39{,}010{,}097} \approx 0.00418 \approx 0.04\%$$

For this study, we work with a simple intuition of multiplying the rate of death factor by the total number of nodes that were infected after a full SIR simulation as an estimator of the number of deaths. The estimated number of deaths $\widehat{x}$ is expressed as follows:

$$\widehat{x} = \theta I,\ when\, t = t_{max}$$

4. Strategies Description

The strategies to mitigate the consequences of the COVID epidemic outbreak consist of a complete isolation policy of a subset of individuals of *G* at an early stage, meaning when $t = 0$. A complete isolation of an individual in the population consists of the removal of its corresponding node in the graph *G* representation. In our proposed business case scenario we will propose the government to only isolate 5% of the population in an early stage to avoid extremely complex logistics and a potential severe impact on the economy.

In order to choose which nodes to remove, we benchmark the impact of 4 potential isolation policies based on different graph centrality measures. The measures' intuition description follows the one proposed by Bhasin [1].

- *Degree Centrality*: In a non-directed graph, the degree *k* of a node is defined as the number of direct connections a node has with other nodes. Degree Centrality metric defines importance of a node in a graph as being measured based on its degree i.e the higher the degree of a node, the more important it is in a graph.

- *Closeness Centrality*: Closeness centrality metric defines the importance of a node in a graph as being measured by how close it is to all other nodes in the graph. For a node, it is defined as the sum of the geodesic distance between that node to all other nodes in the network. The geodesic distance between two nodes *a* and *b* is defined as the number of edges between these two nodes on the shortest path between them.

- *Betweenness Centrality*: This metric defines and measures the importance of a node in a network based upon how many times it occurs in the shortest path between all pairs of nodes in a graph. A sample application of betweenness centrality is to find bridge nodes in graphs.

- *Eigenvector Centrality*: This metric measures the importance of a node in a graph as a function of the importance of its neighbors. If a node is connected to highly important nodes, it will have a higher Eigenvector Centrality score as compared to a node which is connected to lesser important nodes.

It is important to mention that particularly long computation times shall be expected for betweenness centrality since it is a very slow calculation and *G* is relatively big. An alternative was used as an approximate measure. The standard betweenness measure considers every single pair of nodes and the paths between them. NetworkX offers an alternative which uses a random sample of just *n* nodes and then finds shortest paths between those *n* nodes and all other nodes in the network. The chosen *n* value represents 5% of *V* which yields a reasonable computation time of ≈9 minutes.

Rank	Degree Centrality	Closeness Centrality	Betweenness Centrality	Eigenvector Centrality
1	6	6	1013	6
2	2636	2636	1296	2925
3	1412	1412	1342	1412
4	2170	2170	1412	63
5	2925	2925	2271	984

Table 3. 5 most influential nodes per centrality measure.

By analyzing the top 5 most influential nodes per centrality measure (Table 3), we can see that betweenness centrality provides the most discordant results, although it agrees on tagging node 1412 as highly influential like the other metrics. This could be explained given the approximation method used for this measure and the fact that unlike other centrality measures it is totally independent of node degree. Also, it is interesting to see that for the top 5 most influential nodes, the rankings of degree and closeness centrality are fully concordant. Node 6 seems to be the unanimous most influential node in *G*.

5. Experimental Results

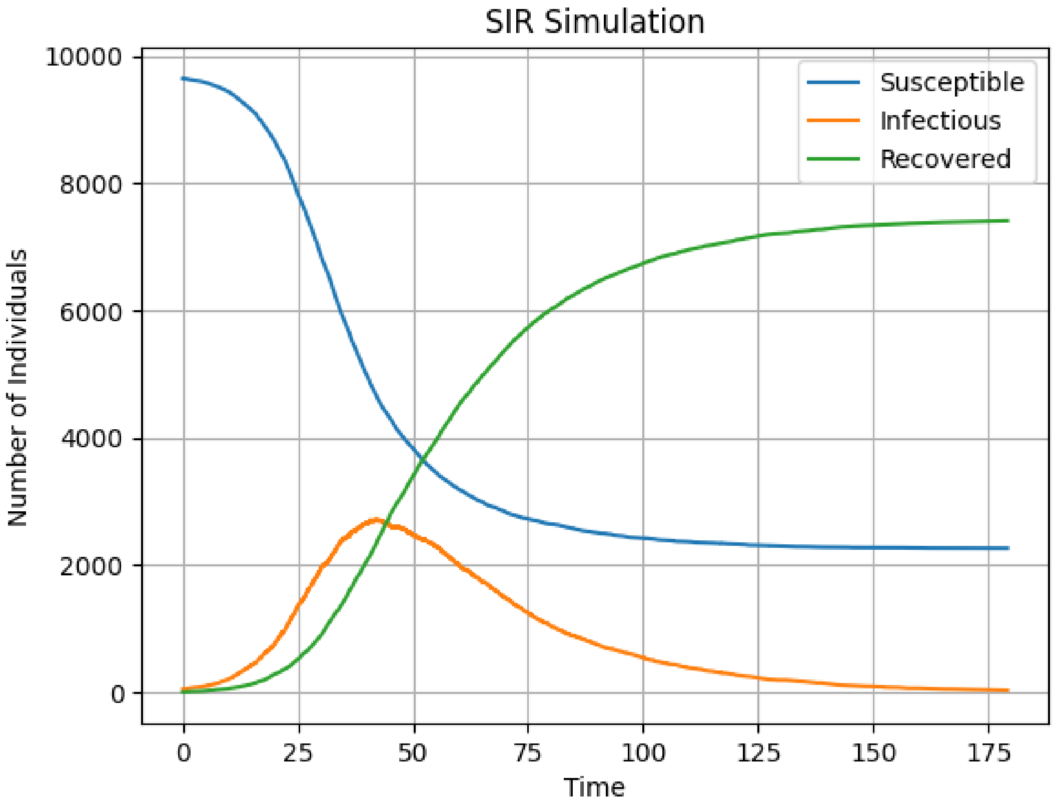

Figure 3. Sample SIR model simulation on graph *G* without any isolation policy.

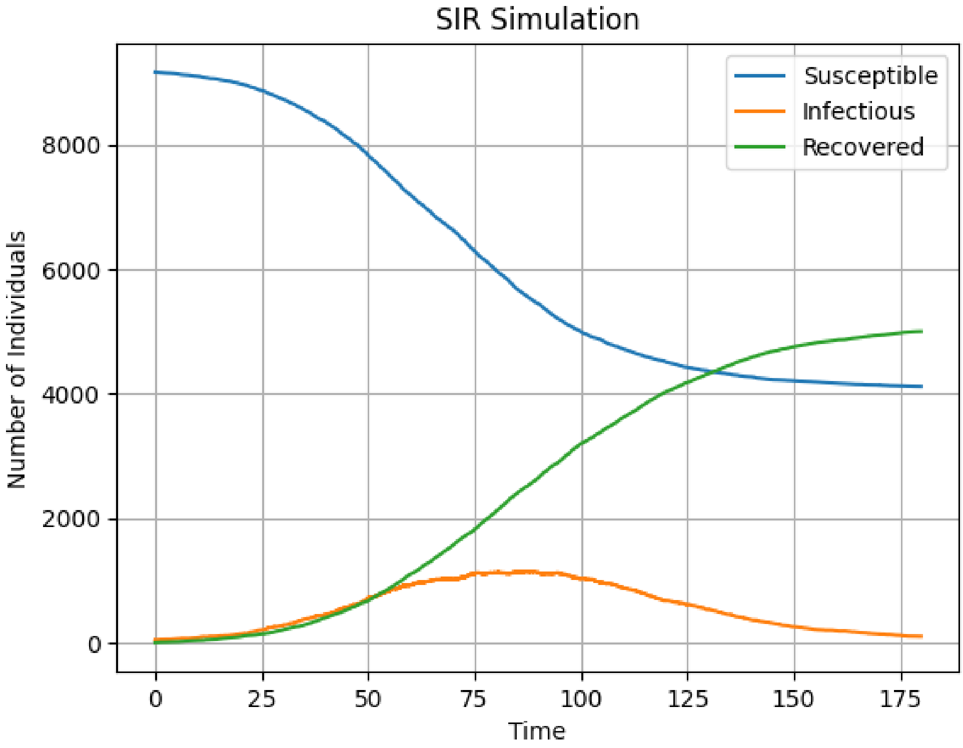

Figure 4. Sample SIR model simulation on graph *G* with Degree Centrality isolation policy.

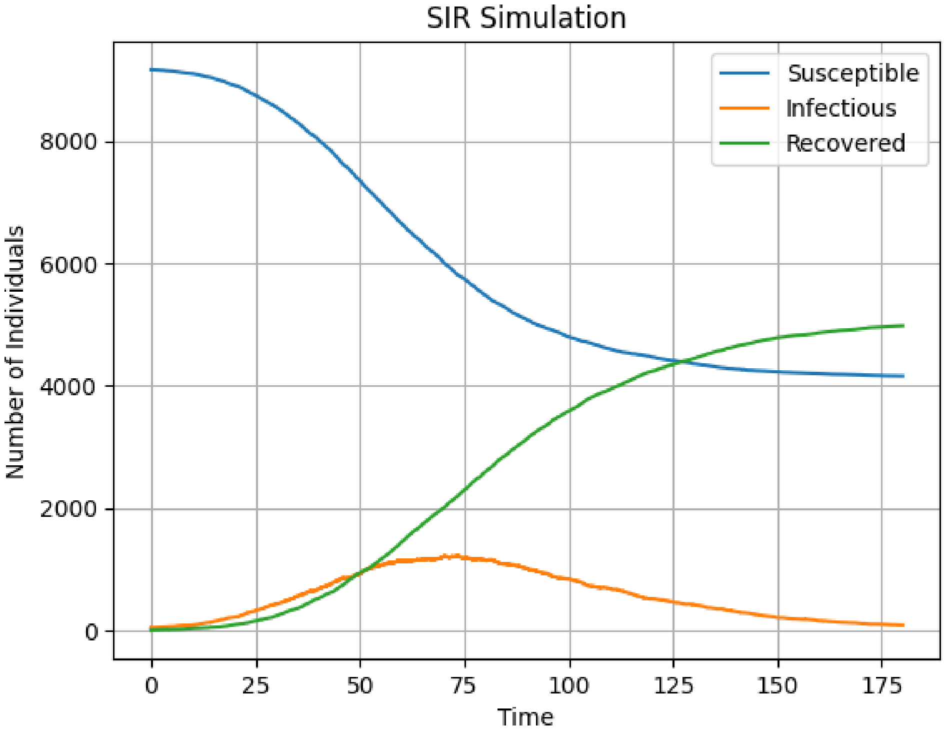

Figure 5. Sample SIR model simulation on graph *G* with Closeness Centrality isolation policy.

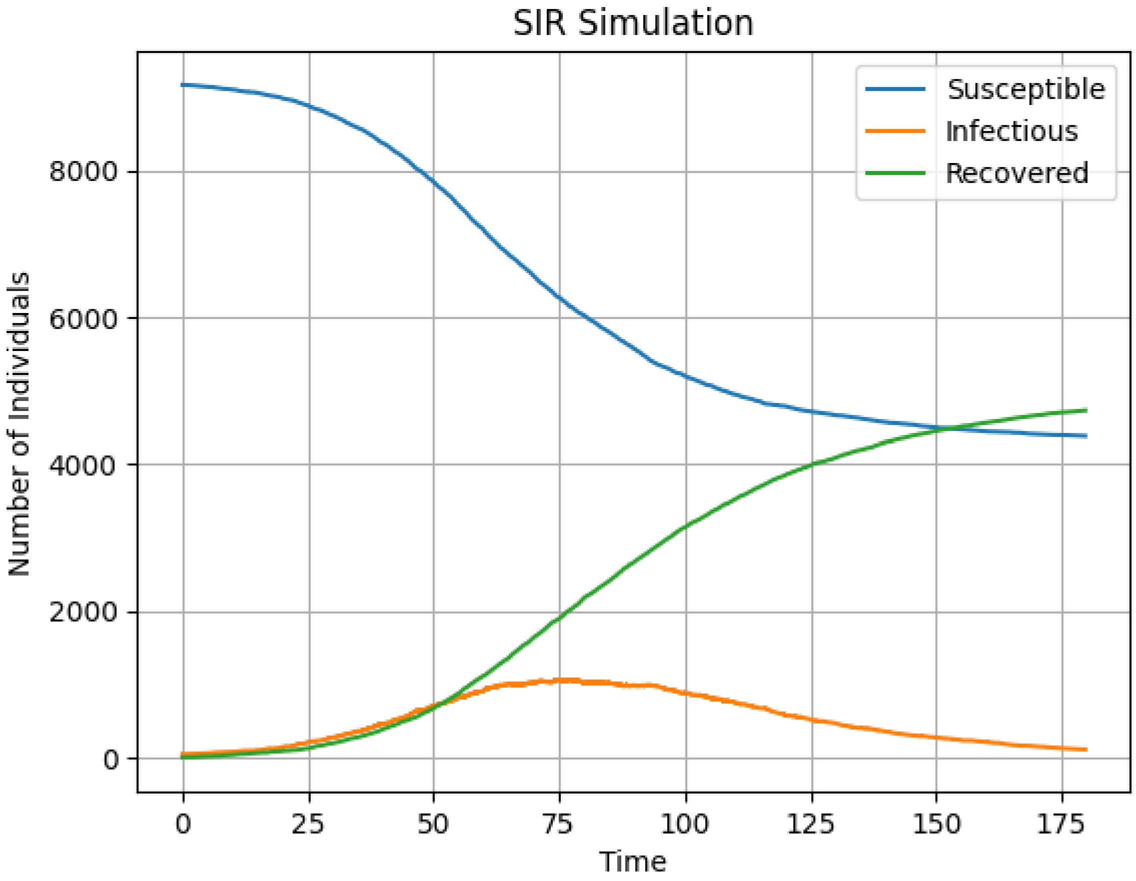

Figure 6. Sample SIR model simulation on graph *G* with Betweenness Centrality isolation policy.

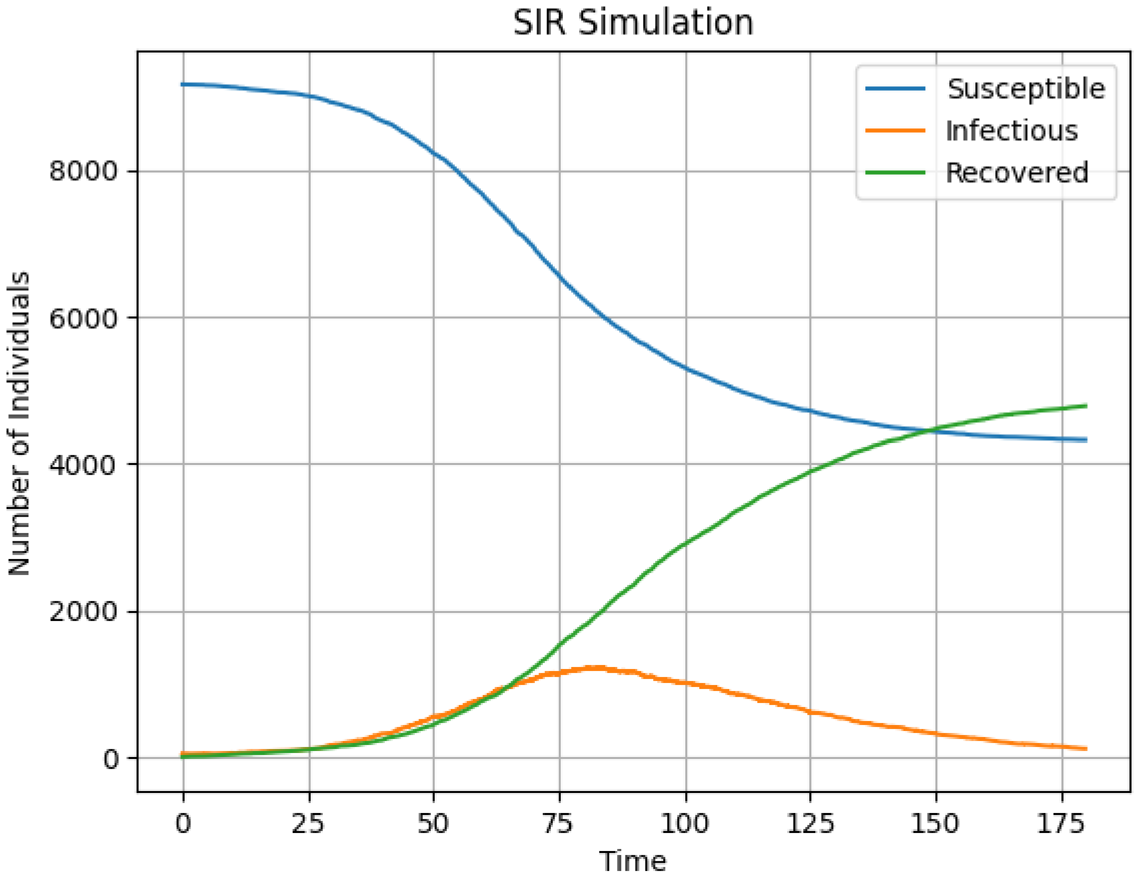

Figure 7. Sample SIR model simulation on graph *G* with Eigenvector Centrality isolation policy.

In all of the SIR model plots implementing an isolation policy we can clearly observe that the curve of infections is severely flattened in contrast to the no policy scenario.

Isolation Policy	δ - Economic Loss Indicator	$\hat{x}$ - Estimated Deaths (rounded)
No Policy	347.95	31
Degree Centrality	154.12	20
Closeness Centrality	167.32	20
Betweenness Centrality	149.16	19
Eigenvector Centrality	151.72	19

Table 4. Performance metrics per isolation policy. For robustness, a SIR model was run 10 times per isolation policy and average values are presented.

By reviewing the results (shown in Table 4) of applying each of the isolation policies, we can state that the Betweenness Centrality isolation policy minimizes $f(G)$. This is a very interesting result since the betweenness centrality used an approximated method for ranking the influence of the nodes. However, from a theoretical point of view, this makes a lot of sense, since the betweenness policy aims to remove the bridge nodes in the network. Bridge nodes are the most likely to carry infections between communities. Eigenvector isolation policy is a very close competitor.

6. Conclusion

In this study, we conducted experiments to investigate the impact of removing influential nodes from a graph representing a population in order to mitigate the impact of a COVID outbreak. We used a SIR epidemic model to simulate the spread of the disease, and we estimated the number of deaths and an economic loss metric as indicators of the pandemic's impact.

Our results indicate that targeted removal of influential nodes can have a significant effect on mitigating the spread of the epidemic. The experiments demonstrated that this method disrupts the transmission pathways and connectivity within the population, effectively reducing the chances of infection and the subsequent spread of the disease. By identifying and removing nodes with high centrality measures, especially when based on betweenness centrality, we observed a decrease in the number of deaths and economic losses.

Further research and analysis can be conducted to refine the selection process for influential nodes, considering dynamic network structures and real-time data. Additionally, exploring the trade-off between the number of nodes removed and the resulting impact on the epidemic's mitigation can provide valuable insights for designing more refined targeted intervention strategies.

Overall, by understanding the network structure and identifying influential nodes, we can propose to the director of the Health Ministry an effective early targeted intervention plan, based on individuals with high betweenness centrality, that can minimize the number of deaths and mitigate the economic losses caused by the COVID pandemic.

7. References

1. Bhasin, J. (2019, August 19). Graph analytics-introduction and concepts of centrality. Medium. towardsdatascience.com/graph-analytics-introduction-and-concepts-of-centrality
2. House, T. (2011). Modeling epidemics on networks. preprint http://arxiv.org/abs/1111.4875.
3. INSEE. (2020). *Point de Conjoncture Du 26 Mars 2020*, p. 4.
4. Manivel, Christian. COVID-19 : Détermination des paramètres R°, β et γ d'un modèle SIR de l'épidémie à partir des décès à l'hôpital en France sur la période du 19 février au 1er juillet 2020. 2020. ⟨hal-02923715⟩
5. Menczer, F., Fortunato, S., & Davis, C. (2020). *A First Course in Network Science*. Cambridge: Cambridge University Press. doi:10.1017/9781108653947
6. Miller et al., (2019). EoN (Epidemics on Networks): a fast, flexible Python package for simulation, analytic approximation, and analysis of epidemics on networks. Journal of Open Source Software, 4(44), 1731. https://doi.org/10.21105/joss.0173
7. Lionel Roques, Etienne Klein, Julien Papaix, Samuel Soubeyrand. Modèle SIR mécanistico-statistique pour l'estimation du nombre d'infectés et du taux de mortalité par COVID-19. [Rapport de recherche] INRAE. 2020. ffhal-02514569v1f
8. Sasaki, K. (2020, March 11). *Covid-19 dynamics with SIR model*. The First Cry of Atom. https://www.lewuathe.com/covid-19-dynamics-with-sir-model.html
9. World Health Organization. (n.d.-a). *France: WHO coronavirus disease (Covid-19) dashboard with vaccination data*. World Health Organization. https://covid19.who.int/region/euro/country/fr
10. W. O. Kermack and A. McKendrick, "A Contribution to the Mathematical Theory of Epidemics," Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character, vol. 115, no. 772, pp. 700–721, Aug. 1927
11. Zhou, F., T. Yu, R. Du, G. Fan, Y. Liu, Z. Liu, J. Xiang, Y. Wang, B. Song, X. Gu, et al. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China : a retrospective cohort study. The Lancet.

Appendix

The code and datasets to reproduce the experiments presented in this article can be found in the following link: https://github.com/mikeguzman1294/SocialNetworkAnalysis/tree/main/challenge_2

There, the notebook named 'Graph_Construction' implements the construction of graph G, whereas the notebook named 'COVID_Simulation' implements the isolation policies, the SIR simulation and the minimization problem metric evaluation.