Multi-dimensional Analysis of Complex Networks for Covid-19

Zelin Li, Zhenhua Ning, Zhenyu Tang, and Xiaoli Shi

[1] Student Id: 22S151142 22S151058 22S151085 22S151072
[2] School of Computer Science and Technology, Harbin Institute of Technology
Shenzhen, China
{22S151142}@stu.hit.edu.cn

Abstract. In the first section, we analyzed the latest situation of the Covid-19 in China and construct a complex network. Firstly, we use the number of new Confirmed cases in each city of Chinese provinces as the main statistical indicator. Then we used the geographic locations of different cities as Node, defined the metrics related to growth rate, number of new cases, and city distance to measure the similarity of different cities' epidemic development as Edge of the complex network, and simplified the construction of the relationship matrix by symmetric regularization. Finally we iterated the results for 7 days to obtain several cities with similar epidemic growth, namely, Beijing, Shanghai, Shenzhen, and Hangzhou.

In the second section, we select 22 representative countries and cluster them using K-Means algorithm to classify them into four categories and create a complex network. We defined Node as the country, size as the population, and Edge as the four categories to which it belongs. We conclude that countries with low Confirmed rate and low Death rates are concentrated in Southeast Asia, countries with low Confirmed rate and high Death rates are concentrated in Western Europe, countries with high Confirmed rate and high Death rates are concentrated in the Middle East, and countries with high Confirmed rate and low Death rates are spread all over the world. In addition, we analyzed and concluded that countries with a higher proportion of elderly people are more sensitive to vaccination rates. In other words, higher vaccination rates reduce Death rate more significantly.

In the third section, we introduce the SEIR model to analyze the different rates of epidemic development in each country in terms of the number of Confirmed diagnoses in China, Singapore, US, and Japan. The least squares method was also used to fit realistic Confirmed case growth curves to obtain the epidemic transmission parameters for each country. Finally, it is discussed that if all countries adopt a strict epidemic prevention strategy like China at the beginning of the epidemic, the global epidemic stability will be controlled within 200 days. Then, some recommendations for prevention strategies are given in the hope that they will help the global fight against similar epidemics.

Keywords: Covid-19 · Complex Network · K-Means · SEIR Model

1 Background

Since the outbreak of Covid-19 in 2020, it has caused huge economic losses and population reduction in countries all over the world. The fight against Covid-19 is no longer the responsibility of one country alone, but an important challenge for the survival and development of all humanity that requires the joint participation of countries around the globe. However, the attitude of different countries towards the epidemic shows a large difference, which makes the current form of the Covid-19 very disparate, and it is difficult to respond to the situation in different countries with the same strategy, even different cities in China show different status of the epidemic.

We learned a lot in the Complex Networks course, including basic graph theory knowledge, some small world networks, the concept of Complex Networks and distinctions. In this paper we hope to clearly explain how networks are formed and our rationale for constructing such a graph (Node and Edge notation). Also our hope is to use Python to perform some computer simulations using on the formulated graph to analyze network properties including node degree distribution, average shortest path length, clustering coefficients, etc. Due to the situation of Covid-19, more complex cases such as some dynamic behaviors against complex networks (e.g., robustness to intentional and random attacks, etc.) are not considered by us.

In this paper, we divide the latter into three sections. In the first section, we analyzed the latest situation of the Covid-19 in China and construct a complex network. In the second section, we select 22 representative countries and cluster them using K-Means algorithm to classify them into four categories and create a complex network. In the third section, we introduce the SEIR model to analyze the different rates of epidemic development in each country in terms of the number of Confirmed diagnoses in China, Singapore, US, and Japan. Finally, in Appendix we include the relevant code so that the reader can easily reproduce our experimental results or pursue a more in-depth analysis. You can also view our code repository through this website https://github.com/Ruzzell13/Complex_Network .

Our contribution of this work is divided as follows.

(1) The Section-1 was mainly contributed by Zhenhua Ning, while Zelin Li was involved in the original idea of iteration, the definition of Edge and the adavanced discussion.

(2) The Section-2 was mainly contributed by Zhenyu Tang and Xiaoli Shi, while Zelin Li was involved in the formulation of ideas, the introduction of clustering methods and the discussion of improvement schemes.

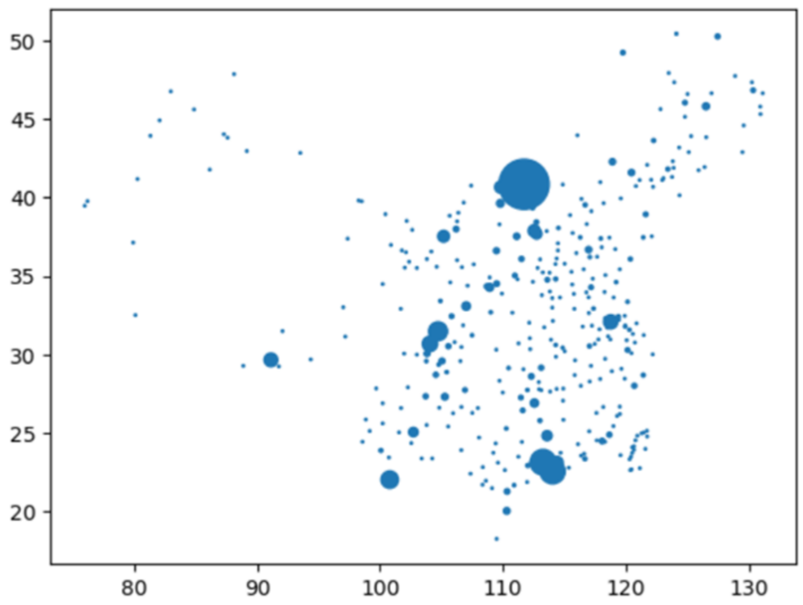
(3) The Section-3 was mainly contributed by Zelin Li, who was also involved in the final presentation of the project, the creation of Slides, the writing of the Paper and the maintenance of the code repository.

2 Section-1: Analysis in China

First we would like to analyze the latest Covid-19 status in China, mainly from https://github.com/BlankerL/DXY-COVID-19-Data, which records the number of Confirmed, suspected Confirmed, cured and Deaths in different cities of China.

For cities with high correlation, we believe that there is a link between their epidemic prevention policies and the status of the epidemic. Specifically, based on the epidemic data, we consider that the actual epidemic prevention policies as well as the epidemic status are similar for cities with similar daily additions as well as growth rates. Also, we consider the distance relationship between cities, but we give less weight to the distance relationship because of the modern transportation. We can guess that the first-tier cities should have a higher correlation with the epidemic situation based on population density and population movement, and this does match the existing epidemic situation. Therefore, we will finally use this to verify the credibility of our network relationships.

Then we construct the Complex Network of Covid-19 in China During the early 7 days in October, 2022. We define Node is every city in China and its Size is the number of newly Confirmed cases. We got the figure as below without Edges.



To introduce Edge, we defined it as a metric to measure the similarity of the growth rate of the epidemic between different cities, mainly calculated by the number of new Confirmed, the number of population, and the distance between cities. The calculation formula is shown below

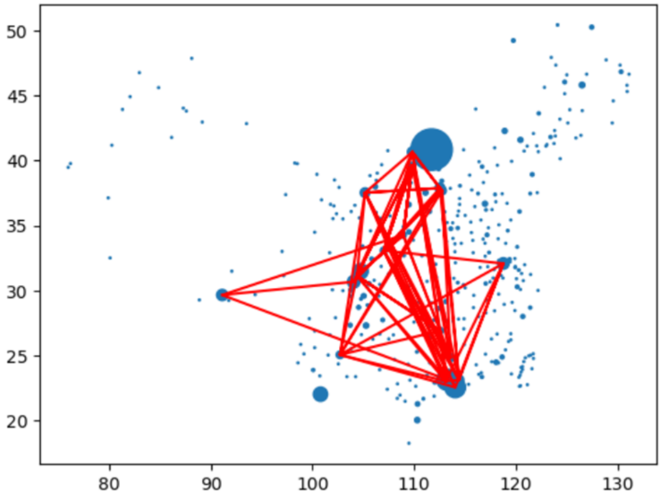
$$Edge_{A,B} = e^{-\frac{\left((new_A + \lambda \times GR_A) - (new_B + \lambda \times GR_B)\right)^2}{dist_{A,B} \times \sigma^2}}$$

This is where ***new*** denotes the number of newly Confirmed and ***GR*** denotes the growth rate of Confirmed, so that the population size can be taken into account as an influencing factor. ***dist*** denotes the distance between two cities, while and are hyperparameters manually adjusted by us, and the final result is a weight between zero and one, which is used to measure the similarity of the current status of the epidemic infection in the two cities.

- $similarity\ matrix\ M \in R^{n \times n},$

$$M = \begin{bmatrix} 1 & City_1, City_2 & \cdots & City_1, City_n \\ City_2, City_1 & 1 & \cdots & City_2, City_2 \\ \vdots & \vdots & \ddots & \vdots \\ City_n, City_1 & City_n, City_2 & \cdots & 1 \end{bmatrix}$$

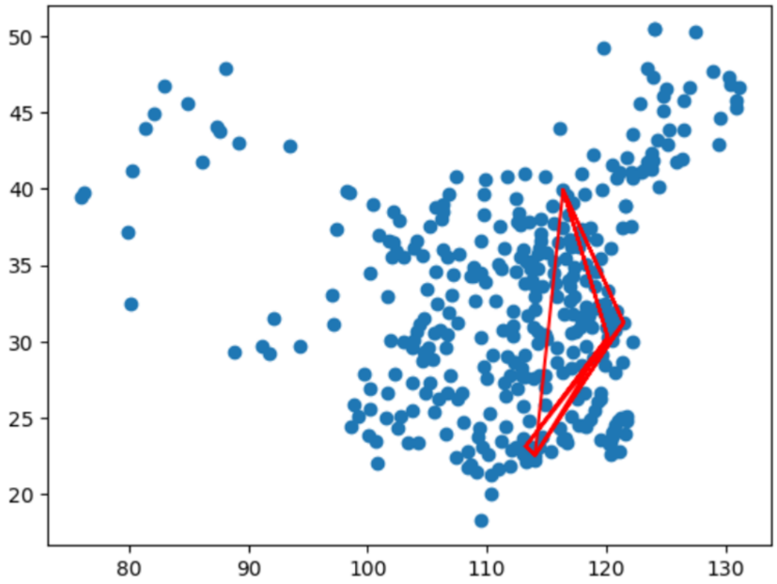
- Symmetric Normalization:
- $M = S^{-\frac{1}{2}} M S^{-\frac{1}{2}},\ \ S = diag(sum(M))$



We use the M-matrix to measure the similarity of different cities for subsequent iterations to calculate the similarity, with a total of n cities taken into account. We compute the new M-matrix simply and quickly using symmetric normalization.

From the picture above we can see that Shenzhen and Guangzhou (big population cities) seem to be more similar, but Edges of Complex Network are heterogeneous. So we use the 7 days' date to iterate the change of this cities, and for convenience we set the Node in same size to demonstrate the Edges of each city more clearly.

We then updated the similarity matrix using M = 0.5 × M + 0.5 × new M, where new M is the similarity matrix calculated from the data of the second day. After iterating through all the data for the seven days of the National Day, we obtain the graph shown below (only the edge relationships are shown for ease of observation).



Main Result

As we can see, with the iteration of the similarity matrix, the cluttered relationships among cities become regular, and we have reasons to believe that cities with edge relationships shown in the figure have high epidemic similarity. Among the first-tier cities, Beijing, Shenzhen, Shanghai and Hangzhou all have high similarity, which is consistent with the high population flow in first-tier cities during the National Day holiday.

Future Work

With more data involved in the computation and a more reasonable similarity calculation, we can produce a more reasonable similarity map of epidemics in cities across the country. In this way, classical algorithms such as conditional random field, CRF can be used to predict the risk of epidemics in each city.

3 Section-2: Analysis in World

3.1 Different types of countries

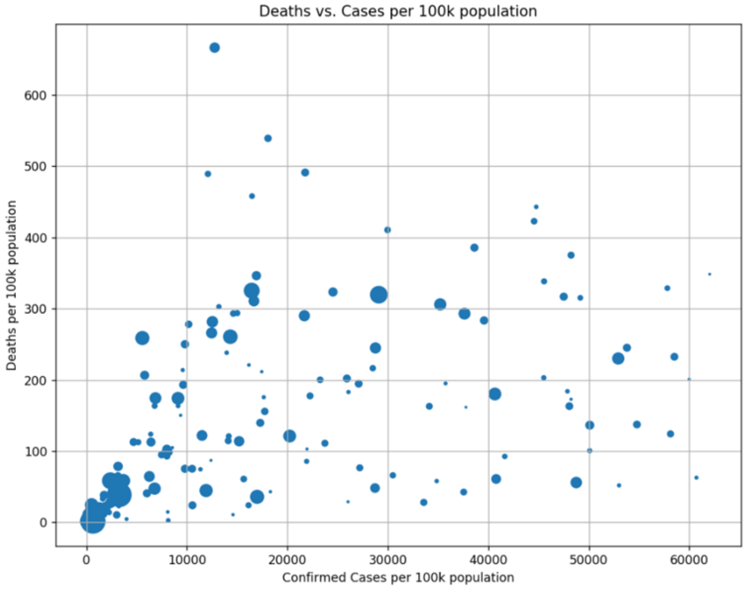
By comparing data since the outbreak of Covid-19 in various countries around the world, we hope to find correlations between the outbreaks in each country and try to find correlations between the outbreaks and the geographic location of each country by comparing their geographic relative locations.

Our data sources are mainly from Johns Hopkins University statistics of Covid-19 https://github.com/CSSEGISandData/COVID-19 and Wikipedia's population numbers for different countries https://zh.wikipedia.org/wiki .

Among the many data of Covid-19, we selected Confirmed cases and Deaths as indicators to analyze the correlation of the epidemic in each country. We selected the global epidemic data of Covid-19 for each region from the database up to 2022/10/9 and processed the data by country.

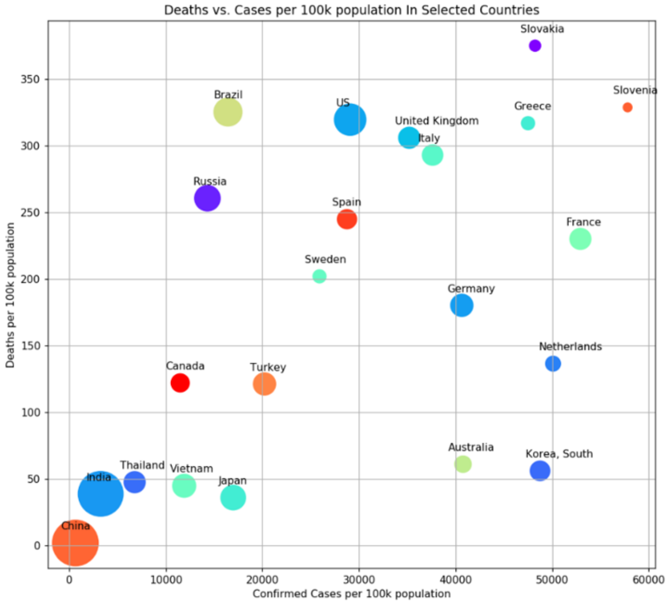
Considering that the population base of each country is too different, resulting in an order of magnitude difference between Confirmed cases and Deaths in each country as well. Therefore, we consider the ratio of Confirmed cases and Deaths to the total population of the country as a better indication of the severity of the epidemic in that country. Using them as horizontal and vertical coordinates, we obtain the following scatter plot.

Each of these Nodes represents a country (the names of the countries are not shown here due to the large number of countries), and the larger the radius of the dot indicates the larger its population base.



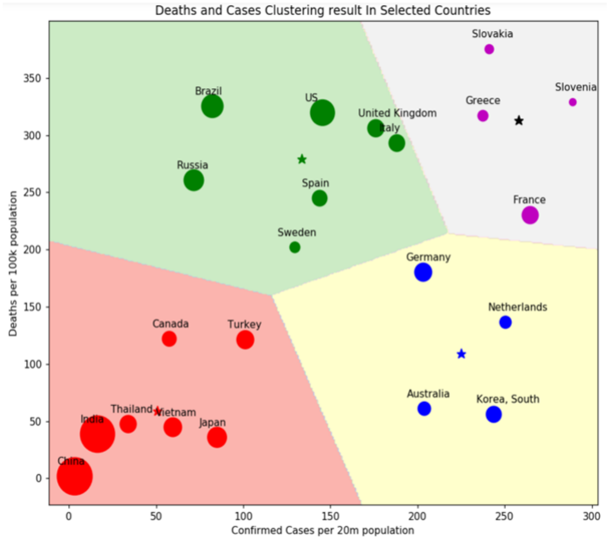
As can be roughly seen, the Nodes in the lower left corner are more dense, with several larger Nodes. It can be inferred that several countries with larger population bases and several with smaller population bases have a smaller proportion of Deaths and Cases, and they have a better control of the epidemic. Most of the remaining points are more evenly distributed, with only a few countries having excessively high Confirmation and mortality rates.

However, because the graph includes all countries worldwide, it will contain more noise and redundancy. For example, many countries with very small populations are included, but if the population size is too small, the impact of the epidemic has a large randomness; the authenticity of the data of some underdeveloped countries is not easy to determine; and many countries with close populations and geographic locations also have very similar data. Therefore, we selected 22 more representative countries and focused on them for analysis.



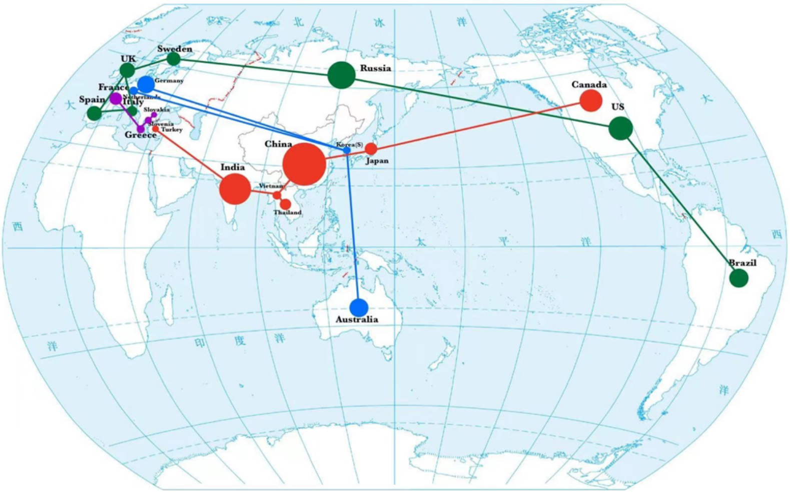
It can be seen that these countries are more evenly distributed. In order to find the correlation between individual countries on the data of the proportion of Deaths and Confirmed cases. We chose to cluster these countries using a simple K-means clustering algorithm.

Since the value of Confirmed cases is very different from the value of Deaths, which has a great impact on the clustering effect, we normalize them and give the same weight to both to show the severity of the national epidemic. The final clustering effect is as follows.



It can be seen that these representative countries are divided into four categories and by the shape of their distribution, they correspond to exactly four scenarios: Low Confirmed rate and Low Deaths rate; High Confirmed rate and Low Deaths rate; Low Confirmed rate and High Deaths rate; and High Confirmed rate and High Deaths rate.

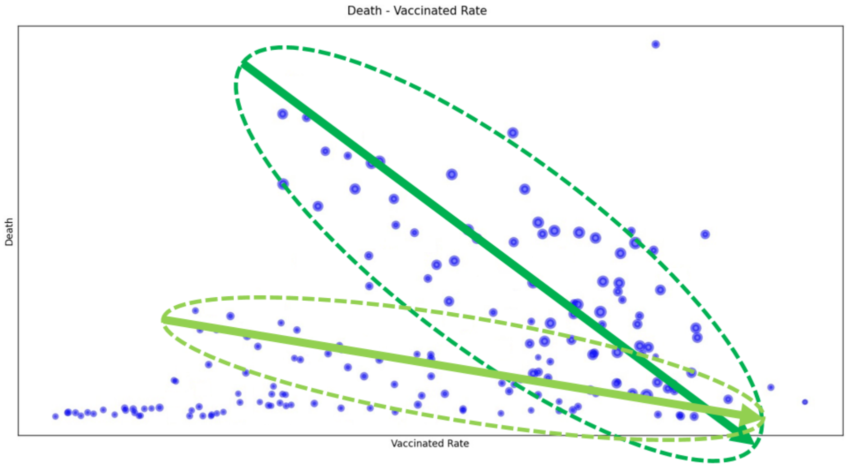
We visualized the results obtained above on a world map, expecting to find the relationship between their clustering results and geographical location.



We mark the Nodes of the same class with the same color and connect them sequentially to obtain a complex network. Each connected component of this complex network represents a class, indicating that they have similar Confirmed cases rate and Deaths rate.

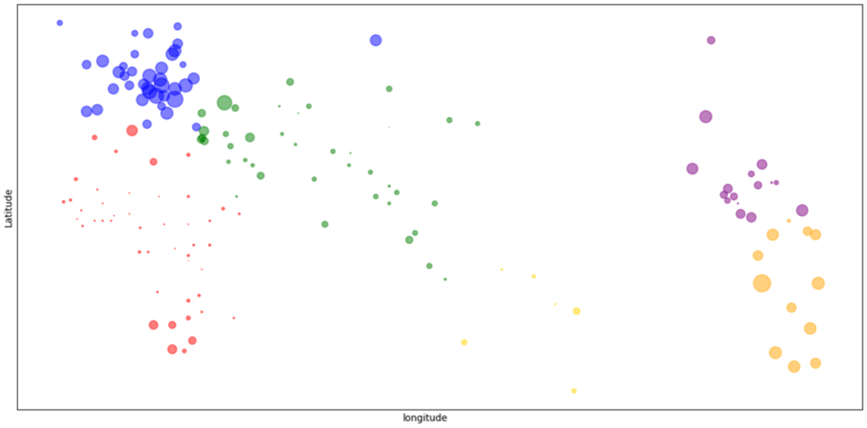
3.2 Elders more rely on Vaccination

We collected the percentage of the population over 70 years of age as well as Vaccination rates and Death rates for each country. In the complex network we designed, there is no explicit Edge, Node represents the different countries, and Size represents the percentage of the population over 70 years old.



From the above figure, it is obvious that the distribution of most Nodes can reflect the negative correlation between Death rate and Vaccination rate, except for many small Nodes close to the x-axis indicating some small countries.In addition, we can also find that for countries with a larger proportion of people over 70 years old, higher Vaccination rates lead to a sharp decrease in Death rates. In contrast, in countries with a smaller population aged 70+, this downward relationship is more even, and is reflected in the picture with a flatter trend. We can conclude that a higher Vaccination rate is a crucial part of reducing the Death rate for the aging population and requires more attention than for younger people

After that we reflected these Nodes onto a world map to try to find connections in terms of geographic location and came to the conclusion. In Europe more people die much and concentrated, while South America and North America die much but not concentrated. Asia, Africa, Australia died not so many people.



Main Result

The results of the visualization can somewhat verify our speculation that the countries with Low Confirmed rate and Low Deaths rate are mainly located in Southeast Asia; the countries with Low Confirmed rate and High Deaths rate are mostly located in Western Europe; the countries with High Confirmed rate and High Deaths rate are located near the Middle East. The countries with High Confirmed rate and High Deaths rate are located near the Middle East. The countries with High Confirmed rate and Low Deaths rate are widely distributed and not concentrated.However, it can be seen that there are countries in each category that are geographically far apart, which may be related to various factors such as the nature of each country, epidemic prevention policies and economic and medical strength. Besides, elder people are more necessary to receive higher Vaccination rate to against the Covid-19.

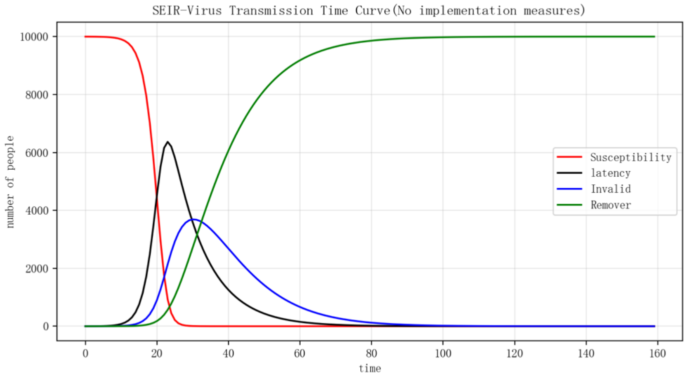
Future Work

Because the visualization was done manually, the number of countries selected was insufficient. In the future, an automated algorithm can be designed for visualization, and then more country data can be added to get a more comprehensive data. Also categorize each country in the world according to its geographical location. The distribution of the two categories is compared to better determine the relationship between the severity of the epidemic and the geographical location.

4 Section-3: Analysis of different Strategies

4.1 Using of SEIR model

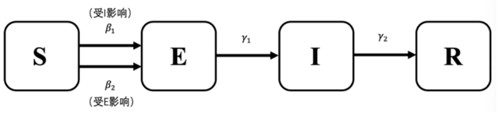
An attempt is made to describe the minimum level that could be reached in the real world with global cooperation against the epidemic, the range of problems that must be addressed at present, and any circumstances that might accelerate or hinder the achievement of the goal.



Introducing the SEIR model, we fit data to the global epidemic population, find the minimum level that can be reached in the future by linear regression, try to show the problems that exist in each country during an epidemic outbreak, and give the corresponding solutions that can accelerate the fight against the epidemic and the mistakes that may accelerate its deterioration.

Firstly, assuming that the total number of people in the city remains constant, i.e. without considering birth and death rates, we have N = S(t) + E(t) + I(t) + R(t), where N is the total number of people in the city and is a constant, and S(t), E(t), I(t) and R(t) denote the number of individuals in the network who are in the susceptible, latent, infected and removed states at time t respectively. Second, if a susceptible individual comes into physical contact with an infected individual, the probability of the susceptible individual becoming a latent individual (contagion rate) is β_1, the probability of an individual in the latent state infecting others per unit time is β_2, while he will move into the infected state with probability γ_1, and similarly, the infected individual becomes a remover with probability γ_2.

r_i is the number of people who are in contact with the infected each day, r_e is the number of people who are in contact each day, and t_free is the time point when the government takes effective measures. C_1, C_2, and C_3 are constants, and $C_1 > C_2$, $C_1 > C_3$. As the government takes active countermeasures, the sick people will be quarantined and effectively treated in time, and the number of



$$\frac{dS(t)}{dt} = -\beta_1 \frac{I(t)}{N} S(t) r_i - \beta_2 \frac{E(t)}{N} S(t) r_e$$

$$\frac{dE(t)}{dt} = \beta_1 \frac{I(t)}{N} S(t) r_i + \beta_2 \frac{E(t)}{N} S(t) r_e - \gamma_1 E(t)$$

$$\frac{dI(t)}{dt} = \gamma_1 E(t) - \gamma_2 I(t)$$

$$\frac{dR(t)}{dt} = \gamma_2 I(t)$$

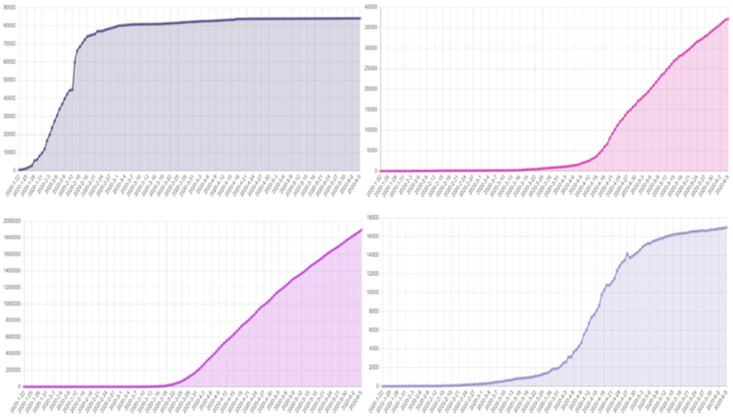
$$r_i = \begin{cases} C_1 & t < t_{free} \\ C_2 & t \geq t_{free} \end{cases},$$

$$r_e = \begin{cases} C_1 & t < t_{free} \\ C_3 & t \geq t_{free} \end{cases}$$

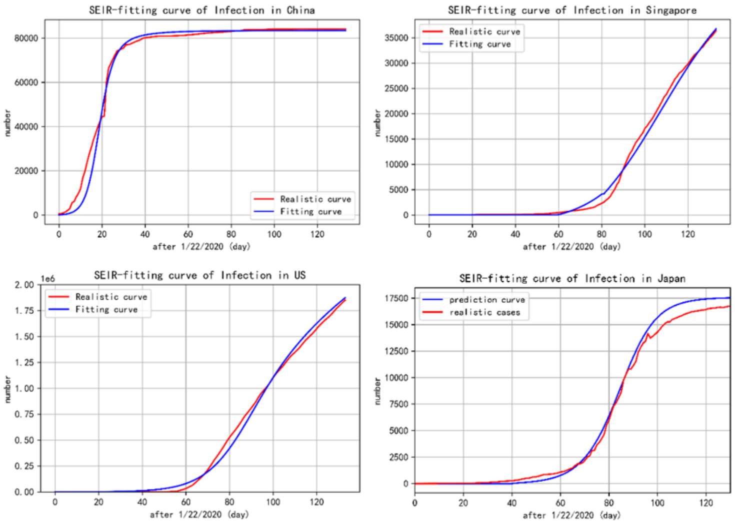
people in contact with the sick will decrease. At the same time, because people actively reduce the frequency of going out, stay away from crowded places, and wear masks to protect themselves from the virus, the number of people that susceptible people come in contact with has also decreased a lot.

4.2 Singapore, China, the US and Japan

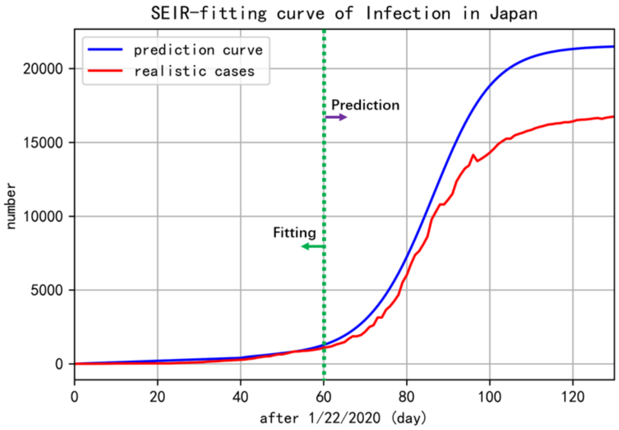
As it is well known that the best time to prevent and control an outbreak is at the beginning of the outbreak, this paper selects four representative countries, Singapore, China, the United States and Japan, as the initial outbreak indicators as of 5 June 2020. As four countries with widely varying epidemic control strategies, their analysis can more visually highlight the impact of epidemic control strategies on the development of the Covid-19, with the total number of people affected by the disease shown in the images below.



The following is a graph of the total number of patients in these four countries after fitting the SEIR model:



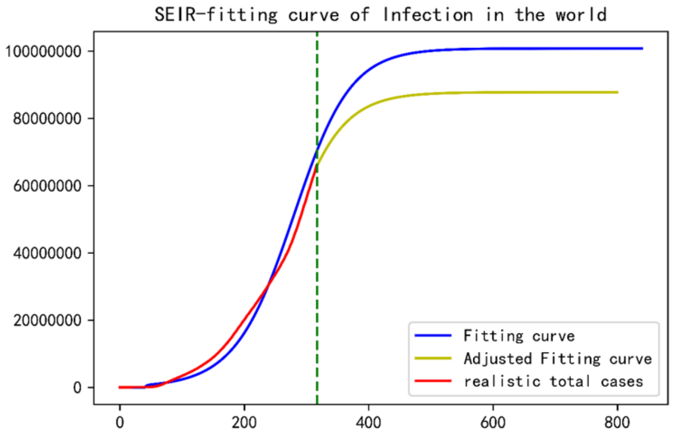
Based on the statistics of the total number of sick people in each country after June 5, 2020, and the real data from news reports, it is clear that about 60 days after the outbreak, countries have started to exchange and transport supplies to better fight the epidemic . Therefore, the pure SEIR model is not applicable after 60 days. In this paper, the SEIR model is used to analyze the time period of the first 60 days after the outbreak, when the influence of external disturbances is less.



In the above picture, we ensure that the model can be well fitted in the first 60 days, and continue to simulate the development of the epidemic without changing the parameters. As you can see, after 60 days, due to the help of epidemic materials or Under other interventions, both the growth rate and the cumulative total are lower than the curve predicted by the previous situation. This reflects that countries have adopted strong corresponding policies to fight the epidemic in the middle of the epidemic. Among them, the international The delivery of medical supplies between countries and the efforts of various countries to work together to fight the epidemic.

4.3 What If Others Be Like China

We first constructed the SEIR model. The data source was the cumulative number of confirmed cases of the global epidemic as of 4 December 2020, with the world as the union population. First, we fitted the model curves to realistic data and tried to derive the subsequent trend of the global epidemic. Afterwards, in order to find the minimum target that can be achieved to effectively control the epidemic, we assume that if all countries in the world are aware of the seriousness of the epidemic, China, which has a better control of the epidemic, is undoubtedly a worthy candidate for study. after 4 December 2020, we introduce into the model the parameters of the curve that previously fitted the accumulation of cases within China, representing the fact that all countries in the world, like China, have adopted strict control and preventive measures as China has done. The results obtained are as follows.



As you can see from the graph above, based on the simulation results of our model, at the current rate of development of the epidemic, it may take more than 300 days for the global epidemic to stabilise, and if China's strict control strategy is followed today, the epidemic may be stabilised in around 200 days. There are, of course, many practical factors that we have not taken into account. For example, it is becoming increasingly difficult for countries to restrict people's travel due to the large base of sick people. However, given the increasing research into the delivery of medical supplies and vaccine drugs between countries, we still believe that a strategy such as the one used in China to reduce person-to-person contact or the wearing of qualified medical masks would be the best and could have very good results. To ensure that the outbreak can be effectively controlled as the model suggests, we have listed some measures for reference.

Main Result

Based on an in-depth analysis of the proposed model, we have highlighted the following insights. The proposed model refines global targets for the minimum achievable level of global cooperation in the fight against epidemics.

The following scenarios may accelerate the achievement of global cooperation in the fight against epidemics. A number of organisations working together against the epidemic are established or relevant agreements are proposed, which may increase the exchange of cooperation between regions. The development of effective vaccines may significantly reduce the probability of healthy people being infected. More medical supplies and medical personnel are sent to those in need.

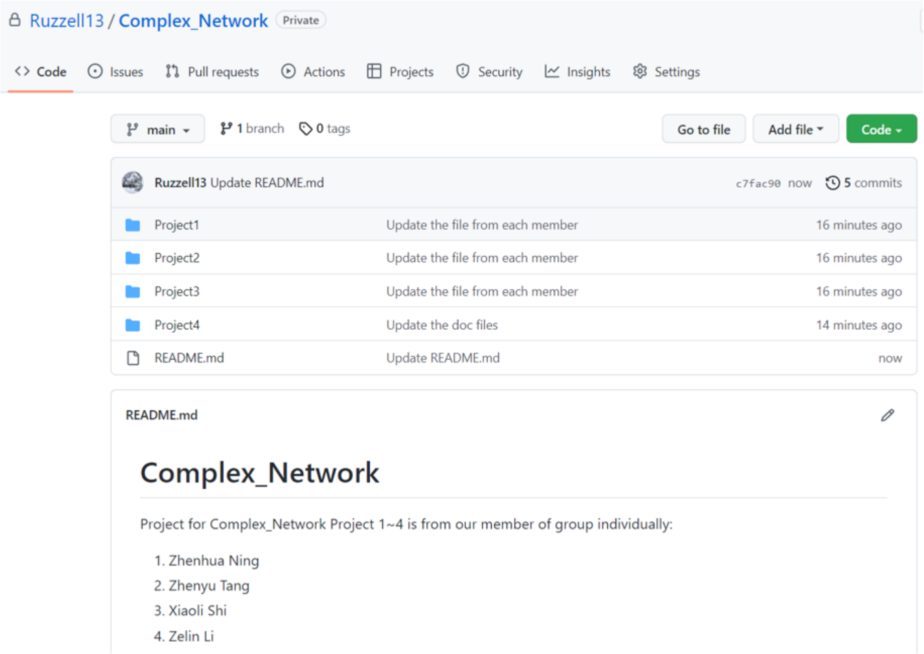
In contrast, situations that may prevent the achievement of global coopera tion in the fight against pandemics are as follows. If there is news or media spreading inappropriate statements, such as rumours, this may lead to people distrusting their government, giving up on themselves, choosing not to wear masks on the streets and other irrational behaviour, inefficient direct cooperation between government agencies will also deteriorated it.

Future Work

In the future, we hope that it will contribute to the fight against the epidemic in the real world. Until then, there are still issues that must be overcome, including applying this analytical model to more national levels to more finely differentiate epidemic prevention policies and assess the strength of the fight against the epidemic.

5 Appendix

We have opened all the code source to Github in order to make it easier for readers to reproduce. Furthermore, you can clone it locally to run and modify it, or to study it more deeply.



Ruzzell13 Update README.md c7fac90 now 5 commits

Project1	Update the file from each member	16 minutes ago
Project2	Update the file from each member	16 minutes ago
Project3	Update the file from each member	16 minutes ago
Project4	Update the doc files	14 minutes ago
README.md	Update README.md	now

README.md

Complex_Network

Project for Complex_Network Project 1~4 is from our member of group individually:

1. Zhenhua Ning
2. Zhenyu Tang
3. Xiaoli Shi
4. Zelin Li

5.1 section-1:In$_{C}hina.ipynb$

```
import wget
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
df_city = pd.read_json("city_geo_list_utf8.json")
city_name = df_city.columns.values.tolist()
print(city_name[:3], len(city_name))
lat, lng = df_city[:1].to_numpy()[0], df_city[1:2].to_numpy()[0]
similarity = np.zeros((len(city_name), len(city_name)))
print(similarity.shape)
df_city = pd.read_json("city_geo_list_utf8.json")
```

```
bsgsh = []
for city in citylist:
    bsgsh.append(df_city.columns.get_loc(city))
bsgsh
def new_confirm(file_name1, file_name2):
    df1, df2 = pd.read_csv(file_name1), pd.read_csv(file_name2)
    df1 = df1[df1[”cityName”].isin(city_name)].drop_duplicates(subset=[”cityName”]).set_index([”cityName”])[”city_confirmedCount”].to_dict()
    df2 = df2[df2[”cityName”].isin(city_name)].drop_duplicates(subset=[”cityName”]).set_index([”cityName”])[”city_confirmedCount”].to_dict()
    sizes = np.array([df2[name] - df1[name] if name in df2.keys() and name in df1.keys() else 0 for name in city_name])
    return sizes
    df1, df2 = pd.read_csv(”DXYArea1.csv”), pd.read_csv(”DXYArea2.csv”)
df1 = df1[df1[”cityName”].isin(city_name)].drop_duplicates(subset=[”cityName”]).set_index([”cityName”])[”city_confirmedCount”].to_dict()
df2 = df2[df2[”cityName”].isin(city_name)].drop_duplicates(subset=[”cityName”]).set_index([”cityName”])[”city_confirmedCount”].to_dict()
pd.read_csv(”DXYArea1.csv”)[:20]
sizes = new_confirm(”DXYArea1.csv”, ”DXYArea2.csv”)
idx = np.array(np.where(sizes > 1))[0]
si = np.expand_dims(sizes[idx], axis=-1)
n = si.shape[0]
si = 1 - (np.square(si - si.T)) / (si + si.T)
si[np.eye(n, dtype=np.bool_)] = 0
idx_mask = np.where(si > 0.2)
edge = np.array([idx[idx_mask[0]], idx[idx_mask[1]]]).T
print(edge.shape)
print(edge)

plt.plot(lat[edge], lng[edge], color=’r’)
plt.scatter(lat, lng, sizes+1)
plt.show()
bsgsh_edge = []
for e in edge:
    if e[0] in bsgsh and e[1] in bsgsh:
        bsgsh_edge.append(e)
bsgsh_edge = np.array(bsgsh_edge)
```

```
bsgsh_edge.shape
n = len(bsgsh)
bsgsh_edge = []
for i in range(n):
    for j in range(i + 1, n):
        bsgsh_edge.append(np.array([bsgsh[i], bsgsh[j]]))
bsgsh_edge = np.array(bsgsh_edge)
bsgsh_edge
lng[bsgsh_edge]
```

5.2 section-2:In$_Wo$*rld.ipynb*

```
import matplotlib.pyplot as plt
import numpy as np
import openpyxl
excel = openpyxl.load_workbook('E:\\Hitsz\\homework\\
                                \\2022        .xlsx')
print(type(excel))
sheets_names = excel.sheetnames
print(sheets_names)
sheet = excel['Sheet2']
print(sheet)
c1 = sheet['C1']
content = c1.value
print(content)
from openpyxl.utils import get_column_letter,
    column_index_from_string
max_column = sheet.max_column
column = get_column_letter(max_column)
row2 = sheet['A1':'%s1' % column]
for row_cells in row2:
    for cell in row_cells:
        print(cell.coordinate, cell.value)
country = []
data = np.zeros((188, 5))
row_num = 2
while row_num <= sheet.max_row :
    country.append(sheet.cell(row=row_num, column=1).
        value)
    row_num = row_num + 1
col_num = 2
row_num = 2
while col_num <= sheet.max_column :
    while row_num <= sheet.max_row :
        data[row_num-2][col_num-2]=sheet.cell(row=row_num, column=col_num).value
        row_num = row_num + 1
    col_num = col_num + 1
    row_num =2
import matplotlib.pyplot as plt
dia = pow(data[:,2],0.5)*0.01
plt.scatter(data[:,3],data[:,4],dia)
plt.title('Deaths vs. Cases per 100k population')
plt.xlabel('Confirmed Cases per 100k population')
plt.ylabel('Deaths per 100k population')
plt.grid(True)
plt.show()
excel2 = openpyxl.load_workbook('E:\\Hitsz\\homework\\ \\          .xlsx')

sheets_names = excel2.sheetnames
sheet_2 = excel2['Sheet1']
country_2 = []
data_2 = np.zeros((sheet_2.max_row, 5))
row_num = 1
while row_num <= sheet_2.max_row :
    country_2.append(sheet_2.cell(row=row_num, column=1).value)
    row_num = row_num + 1
print(country_2)
col_num = 2
row_num = 1
while col_num <= sheet_2.max_column :
    while row_num <= sheet_2.max_row :
        data_2[row_num-1][col_num-2]=sheet_2.cell(row=row_num, column=col_num).value
        row_num = row_num + 1
    col_num = col_num + 1
    row_num =1
dia_2 = pow(data_2[:,2],0.5)*0.05
N = sheet_2.max_row
colors = np.random.rand(N)
plt.figure()

plt.scatter(data_2[:,3], data_2[:,4], s=dia_2, c=colors, cmap='rainbow')
plt.title('Deaths vs. Cases per 100k population In Selected Countries')
plt.xlabel('Confirmed Cases per 100k population')
plt.ylabel('Deaths per 100k population')
for i in range(0,N):
    plt.annotate(country_2[i], xy = (data_2[i,3], data_2[i,4]), xytext = (data_2[i,3]-1500, data_2[i,4]+10), weight = 'light')
plt.grid(True)
plt.show()

data3 = np.zeros((22, 2))
data3[:,0] = data_2[:,3]/200
data3[:,1] = data_2[:,4]
from sklearn.cluster import KMeans

plt.figure()
k = 4
model = KMeans(n_clusters=k)
model.fit(data3)
centers = model.cluster_centers_
result = model.predict(data3)

x_min, x_max = data3[:, 0].min() - 15, data3[:, 0].max() + 15
y_min, y_max = data3[:, 1].min() - 25, data3[:, 1].max() + 25

xx, yy = np.meshgrid(np.arange(x_min, x_max, 1), np.arange(y_min, y_max, 1))
z = model.predict(np.c_[xx.ravel(), yy.ravel()])
z = z.reshape(xx.shape)
cs = plt.contourf(xx, yy, z, cmap = plt.cm.Pastel1)

mark = ['r', 'g', 'b', 'm']
cresult = [[] for i in range(22)]

for i, d in enumerate(data3):
    cresult[i] = mark[result[i]]
plt.scatter(data3[:,0], data3[:,1], s = dia_2/1.5, c = cresult)
plt.title('Deaths and Cases Clustering result In Selected Countries')
plt.xlabel('Confirmed Cases per 20m population')
plt.ylabel('Deaths per 100k population')
for i in range(0,N):
    plt.annotate(country_2[i], xy = (data3[i,0], data3[i,1]), xytext = (data3[i,0]-10, data3[i,1]+10), weight = 'light')
mark = ['*r', '*g', '*b', '*k']
for i, center in enumerate(centers):
    plt.plot(center[0], center[1], mark[i], markersize=10)

plt.show()
```

5.3 section-3:Analysis.py

```
import pandas as pd
from functools import reduce
import networkx as nx
import matplotlib.pyplot as plt
import json
import numpy as np

def data_precess():
    data = pd.read_csv('owid-covid-data.csv')
    #data = pd.read_csv('https://covid.ourworldindata.org/data/owid-covid-data.csv')
    d1 = data.drop_duplicates(subset=['iso_code'], keep='last')
    d1 = d1[['iso_code', 'location', 'continent', 'total_deaths_per_million', 'aged_65_older']]
    d1 = d1.sort_values(by='iso_code')


    d2 = data.groupby('iso_code')['people_fully_vaccinated_per_hundred'].max()
    d3 = data.groupby('iso_code')['stringency_index'].mean()

    with open('allCountriesGeojson.json', encoding='utf-8') as f:
        country_loc = json.load(f)
    location = pd.json_normalize(country_loc, 'features',
                                ['id', ['properties', 'latitude'], ['properties', 'longitude']],
                                meta_prefix = '_',
                                errors='ignore')
    location = location[['id','properties.latitude', 'properties.longitude']]
    location.columns=['iso_code','latitude', 'longitude']

    dfs = [d1, d2, d3, location]
    res = reduce(lambda x, y: pd.merge(x, y, on='iso_code', how='inner'), dfs)
    res = res.dropna()
    res = res.drop_duplicates(subset=['iso_code'], keep='last')
    res.to_csv('latest_info.csv')

def main():
    data_precess()
    data = pd.read_csv('latest_info.csv')
    data1 = data[['aged_65_older', 'stringency_index', 'people_fully_vaccinated_per_hundred']]

    G = nx.Graph()
    G.add_nodes_from(data['iso_code'])

    pos = {}
    node_size = []
    node_color = []
    for i in range(len(data)):
        pos[data['iso_code'][i]] = [data['longitude'][i], data['latitude'][i]]
        node_size.append(data['total_deaths_per_million'][i]*0.05)
        if data['continent'][i] == 'Asia':
            node_color.append('green')
        elif data['continent'][i] == 'Europe':
            node_color.append('blue')
        elif data['continent'][i] == 'Africa':
            node_color.append('red')
        elif data['continent'][i] == 'North␣America':
            node_color.append('purple')
        elif data['continent'][i] == 'South␣America':
            node_color.append('orange')
        else:
            node_color.append('gold')
        ''' 'people_fully_vaccinated_per_hundred'
        for j in range(i+1, len(data)):
            if np.linalg.norm(data1.loc[i] - data1.loc[j]) < 2:
                G.add_edge(data['iso_code'][i], data['iso_code'][j])
            '''

    nx.draw_networkx(G,pos,with_labels=False,alpha=0.5,node_size= node_size, node_color = node_color,
                        linewidths=1, width=1,edge_color='gray')

    plt.xlabel('longitude')
    plt.ylabel('Latitude')
    plt.show()

if __name__ == '__main__':
    main()
```