MA4207 - MACHINE LEARNING AND NETWORK ANALYSIS

Project Report

submitted by

Aditya Kumar Singh
Santanu Das
Balu Shaharica
Siddhikant Mishra

Spring Semester–2020
Department of Mathematics & Statistics
IISER KOLKATA

June 27, 2020

Contents

1 Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). COVID-19 outbreak was first reported in Wuhan, China and has spread to more than 50 countries. WHO declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) on 30 January 2020. Naturally, a rising infectious disease involves fast spreading, endangering the health of large numbers of people, and thus requires immediate actions to prevent the disease at the community level.

1.1 About COVID-19

An infectious disease outbreak is the occurrence of a disease that is not usually expected in a particular community, geographical region, or time period. Typically, a rising infectious disease involves fast spreading, endangering the health of large numbers of people, and thus requires immediate action to prevent the disease at the community level. COVID-19 is caused by a new type of coronavirus which was previously named 2019-nCoV by the World Health Organization (WHO). It is the seventh member of the coronavirus family, together with MERS- nCoV and SARS-nCoV, that can spread to humans. The symptoms of the infection include fever, cough, shortness of breath, and diarrhea. In more severe cases, COVID-19 can cause pneumonia and even death. The incubation period of COVID-19 can last for 2 weeks or longer. During the period of latent infection, the disease may still be infectious. The virus can spread from person to person through respiratory droplets and close contact.

1.2 Obejctive

Our objective is to use the past information(data) about virus spread and model a mathematical differential equation for disease model **SEIR** that will predict the future trend of COVID-19(also called **Predictive modelling**). And besides that we will also a frame our human interaction into a network model called **Barabasi-Albert** and will predict how the infection is spreading within that network.

1.3 About Network Model

Social Networks are a social structure made up of a set of social elements such as individuals or organisation and social contact between them. Social networks are important as one can predict many thing about different social behaviour like **clustering**, **connections** and so many things. In this project, we shall study about one particular epidemic model applied on social networks and see how infection spreads, where edge between nodes(individuals) will form only if they don't maintain the social distancing like if they come within 1 meter radius of infected then they will be exposed.

1.4 Dataset

We will be using dataset from **Kaggle**(here is the link) to fit the model for countries like **Qatar**. The reason why we have chosen this countries is due to their low population(and hence less chaotic behaviour) with high number of cases that implies of how well-mixed their population is.

1.5 Basic Terminology

Susceptible(S): A susceptible individual(sometimes known simply as a susceptible) is a member of a population who is at risk of becoming infected by a disease.

Exposed(E): These are individuals who are already infected but are asymptomatic

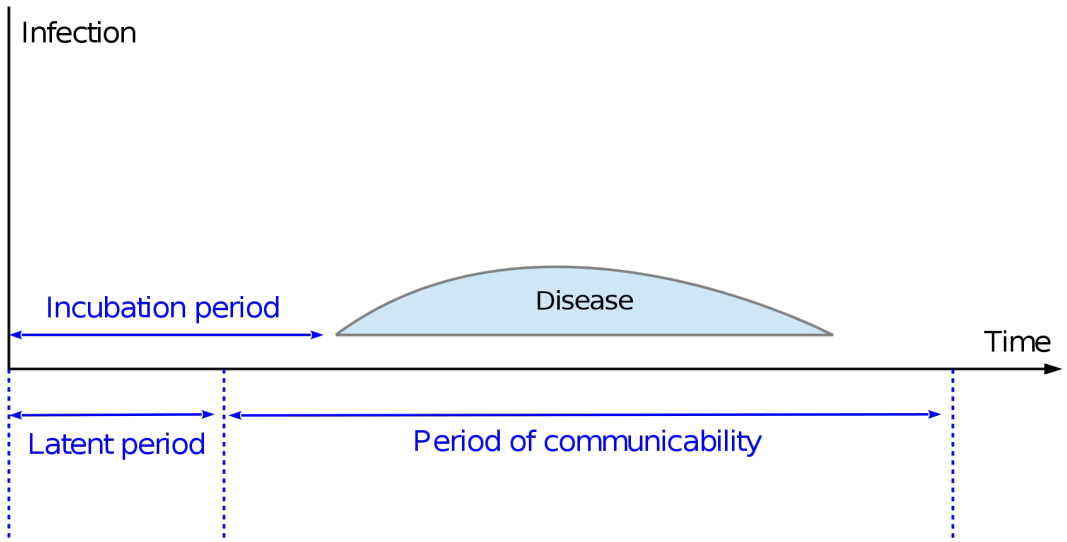
Infected(I): Infectious are individuals who are showing signs of infection and can transmit the virus.

Recovered(R): Recovered are individuals who are previously infected but are no longer infectious and already immune to the virus.

Dead(D): All those individuals who are dead due to the infection and hence are *removed* from the whole population.

Incubation Period: Incubation period is the time elapsed between exposure to a pathogenic organism, a chemical, or radiation, and when symptoms and signs are first apparent.

Latency(or Latent period): The Latent period or the Pre-infectious period is the time interval between when an individual or host is infected by a pathogen and when he or she becomes infectious, i.e. capable of transmitting pathogens to other susceptible individuals.



Basic Reproductive number ($\mathbf{R_0}$) : The basic reproduction number, R_0, is defined as the expected number of secondary cases produced by a single (typical) infection in a completely susceptible population. It is important to note that R_0 is a dimensionless number and not a rate, which would have units of $time^{-1}$. Some authors incorrectly call R_0 the "basic reproductive rate." However, since our model does not include any birth cases, the basic reproductive number, $R_0 = \beta/\gamma$, does not change, where β and λ explained later. If $R_0 < 1$, epidemic fades out and If $R_0 > 1$, epidemic spreads.[Dri17]

2 SEIR Model

SEIR model is a remodelling of SIR disease model where we have just add an extra compartment for *exposed* people to give latency to the infection. So we have four compartments

namely **Susceptible(S)**, **Exposed(E)**, **Infected(I)**, **Recovered or Removed(R)**. People in **S** becomes exposed(i.e. get into **E**) when they come in contact with the infected ones. And later with some probability, exposed people comes in compartment **I** infected and then loss infection(removed) again with some probability to come in compartment **R**.

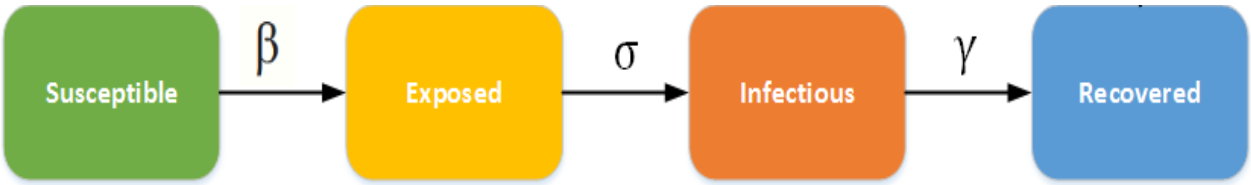
2.1 Assumptions

1. Population is closed(without "vital dynamics") i.e. the births, deaths, and immigration are therefore ignored and the count is set to a fixed number, say **N**, where $N = S + E + I + R$ at any time during the epidemic.
2. The individuals in the Exposed state are infected but not yet infectious.
3. Well-mixed population.
4. SEIR model assumes that the latent and infectious times of the pathogen are exponentially distributed.
5. The person once recover from infection will never catch it again and hence will be immune to it life-long.[Ham20]

2.2 Parameters

Let's define:

1. Incubation rate σ, is the rate of latent individuals becoming infectious. Given the known average duration of incubation Y, $\sigma = 1/Y$.
2. Recovery rate $\gamma = 1/D$, is determined by the average duration of recovery D, of infection. After this period, they enter the removed phase.
3. The infectious rate, β, controls the rate of spread which represents the probability of transmitting disease between a susceptible and an infectious individual. In other words, each infected individual has, on average, β contacts with randomly chosen susceptible people per unit time(day). Mathematically, $\beta = R_0\gamma$ in closed population scenario.



2.3 Working of Model

1. Start the epidemic with a group of Susceptible individuals and atleast one Infected individual.
2. Infectious class mix with the Susceptible class and create Exposed individuals following a probabilistic process($= \beta$).
3. Exposed individuals spend some days without spreading the virus and based on another probabilistic process($= \sigma$) become additional Infectious class.
4. Newly Infectious class repeat 2 and create more Exposed class.

5. Infectious individuals based on a probabilistic process($= \gamma$) either recover or die and become Removed class.

2.4 Mathematical Representation of SEIR Model

1. The disease is transmitted only when an infected person has contact with a susceptible one. If the total population consists of N people, then the average probability of a person you meet at random being susceptible is S/N, and hence an infected person has contact with an average of $\beta S/N$ susceptible people per unit time. Since there are on average I infected individuals in total that means the overall average rate of new exposed will be $\beta SI/N$. And we can write a differential equation for the rate of change of E thus:
$$\frac{dE}{dt} = \beta \frac{SI}{N} \tag{1}$$
Since at the same time the number of Susceptible individuals goes down at the same rate:
$$\frac{dS}{dt} = -\beta \frac{SI}{N} \tag{2}$$

2. Also it's not like that Exposed group doesn't deplete, they also go down at some constant rate by getting Infectious, hence corresponding changes to (1) would be:
$$\frac{dE}{dt} = \beta \frac{SI}{N} - \sigma E \tag{3}$$

3. Now Infected group increases at the rate as Exposed group depletes and also goes down at some constant rate due to Recovery or Death(Removed):
$$\frac{dI}{dt} = \sigma E - \gamma I \tag{4}$$

4. Again Removed group increases at the rate same as to the depletion rate of Infected group. But here the Removed group never goes down as there is no transformation to other state hereafter. Hence:
$$\frac{dR}{dt} = \gamma I \tag{5}$$

Putting (2), (3), (4), and (5) together, we get:

$$\begin{aligned}
\frac{dS}{dt} &= -\beta \frac{SI}{N} \\
\frac{dE}{dt} &= \beta \frac{SI}{N} - \sigma E \\
\frac{dI}{dt} &= \sigma E - \gamma I \\
\frac{dR}{dt} &= \gamma I
\end{aligned}$$

where $S + E + I + R = N$.
Define: $s = S/N$, $e = E/N$, $i = I/N$, $r = R/N$, then system of non-linear differential equation would be after dividing N throughout:

$$\begin{aligned}
\frac{ds}{dt} &= -\beta si \\
\frac{de}{dt} &= \beta si - \sigma e
\end{aligned}$$

$$\frac{di}{dt} = \sigma e - \gamma i$$
$$\frac{dr}{dt} = \gamma i$$

where $s + e + i + r = 1$.
At time, $t = 0$, we have:

- $S(0) = N =$ the whole population.
- $E(0) = 0$ i.e. zero exposed individuals which will get exposed in next time step.
- $I(0) = 1$(atleast 1).
- And $R(0) = 0$.

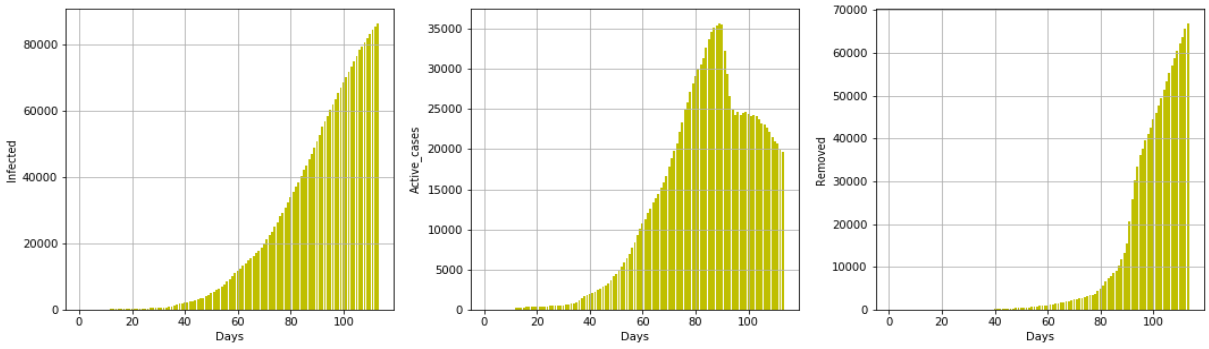
Now we know a linear differential equation has structure whereas a non-linear differential equation can really be anything. Standard methods like *Reduction of Order*, *Separation*, *Bernoulli's method*, *Laplace*, *Anninilator*, etc. usually don't apply and even proving existence is difficult. So instead of trying to solve this analytically we will approach it *computationally*.

3 SEIR Model: Solving and Fitting

In this section, we will model the global trajectory of the infection counts using the SEIR model. We will be using data from **Kaggle** and will focus on Qatar. Before using it, we will clean it and use.
Using *Scipy's* implementation in *Python* of the numerical integration of ODE, *odeint*, we will plot S, E, I, and R with respect to time and will fit it against the real data for **Qatar**. But before doing so, let's visualize our data and distribution.

3.1 Data Visulaization

We have done the bar plot for real data for Qatar.



3.2 Model plot and fitting

Below is the given graph for parameters $\beta = 1.05$, $\sigma = 1/5.2$, $\gamma = 1/9.5$ and for certain initial conditions i.e.

- $S(0) = 2 \times 10^7$,
- $E(0) = 0$,

- $I(0) = 1$,
- $R(0) = 0$.

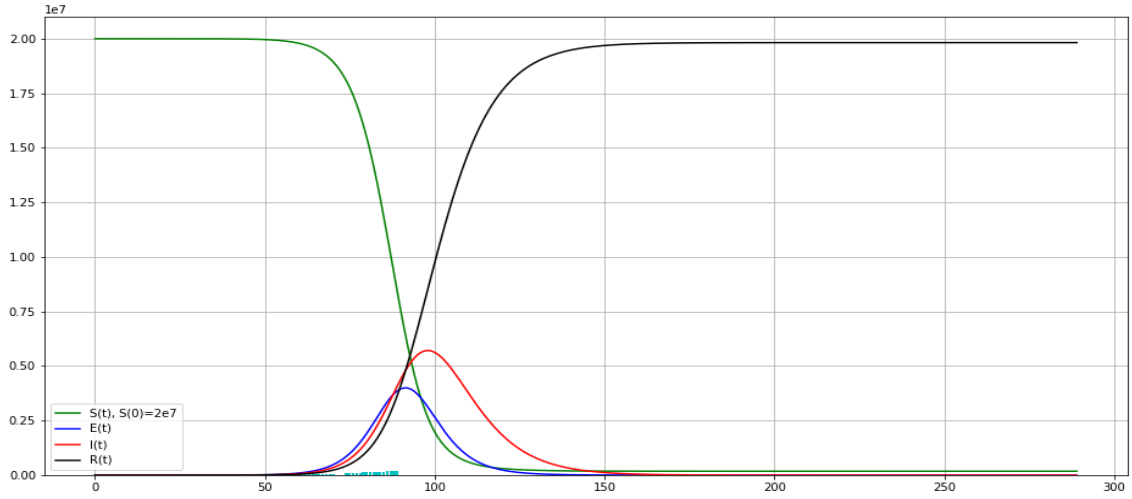


Figure 1: Plot from solution of system of Diff. equations

3.2.1 SEIR Parameter trends

	S	**E**	**I**	**R**
β	**Saturates lower**	**Left and higher**	**Left and higher**	**Saturates higher**
σ	**Saturates lower**	**Left and higher**	**Left and higher**	**Saturates higher**
γ	**Saturates lower**	**Left and lower**	**Left and lower**	**Saturates higher**

If $\sigma < \sigma_{\text{threshold}}$, and $\gamma < \gamma_{\text{threshold}}$, then exposed peak is lower as compared to infected and vice-versa. And always exposed will peak before infected.

3.2.2 Parameter estimation

For estimating the parameters the following are assumed:

- Mean latent period $1/\sigma$ and mean infectious period $1/\gamma$ are assumed to be constant even though the virus mutates very fast.

$$\sigma = 1/10 \quad , \quad \gamma = 1/15$$

- Initial Reproduction Number $R_0 = 4$.
- Now we define $\beta := R_0\gamma K$. Since reproduction number is β/γ for SEIR Model. Also R_0 and γ are assumed to be constant, so any change in β is due to K. Therefore, to fit data we will change K only.
  Note that $R_{\text{eff}} = R_0 K$(*Effective reproduction number*).
  We have tried '*lmfit*' module of *Python* to fit data using least squares method. Unfortunately the error reported was very high, hence unreliable.

3.2.3 Fitting

Below is fit for **Qatar** data for parameters $\beta = (R_0\gamma)K = (R_0\gamma)(0.311) = 0.083$, $\sigma = 1/10$, $\gamma = 1/15$, $r_{\text{eff}} = 1.245$ and initial conditions are as follows:

- $S(0) = 2807805$ (Population of Qatar),
- $E(0) = 0$, $I(0) = 1$,
- $R(0) = 0$.

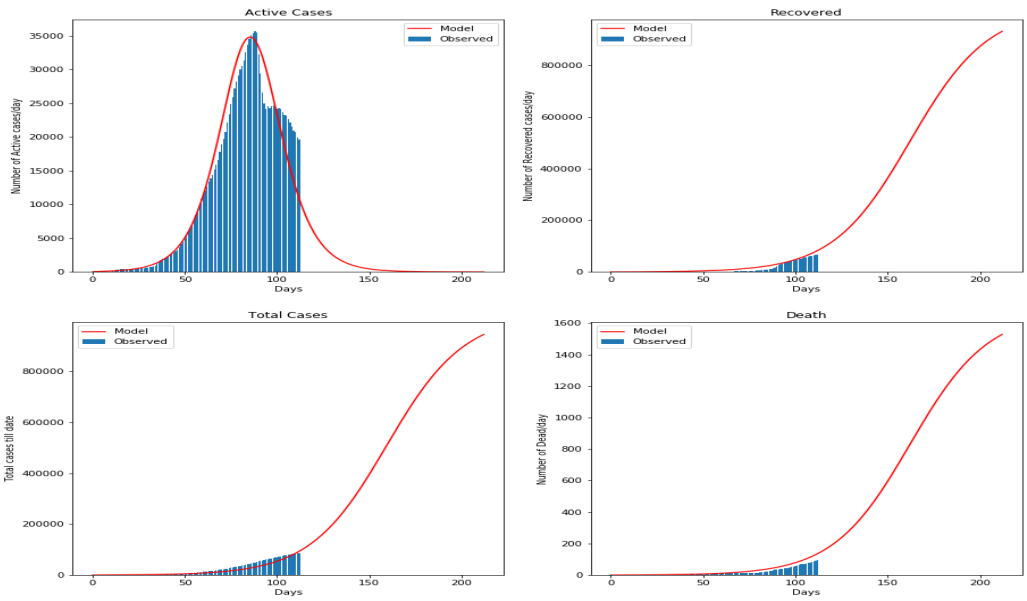
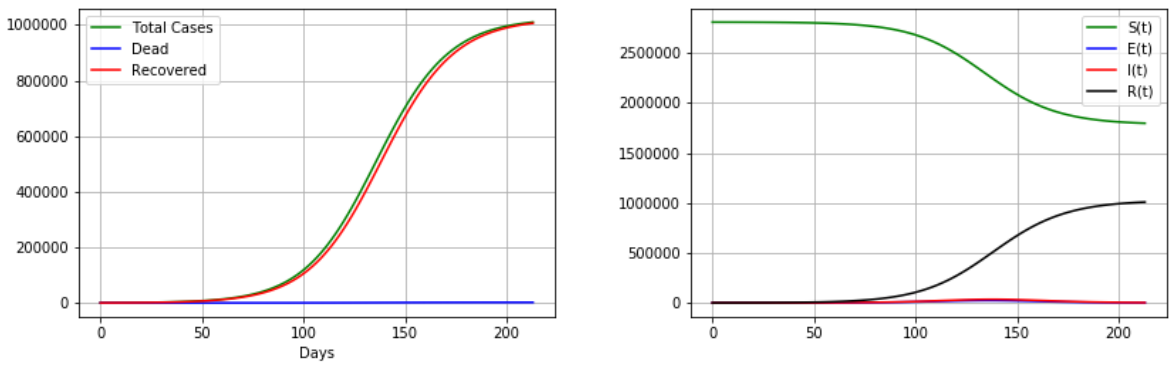


Figure 2: Model fit on Real "Qatar" data



(a) Death, Recovered, Total cases

(b) S,E,I,R

Figure 3: Fit of SEIR Model for "Qatar"

The modeled death curve is got by finding out the fraction of dead cases w.r.t removed cases form real data.

$$\text{Total cases} = I(t) + R(t)$$
$$\text{Recovered} = (1 - \text{frac}_{\text{dead}})R(t)$$

$$\text{Dead} = \text{frac}_{\text{dead}} R(t)$$

$$\text{frac}_{\text{dead}} = \frac{\#\text{ dead cases in observed data}}{\#\text{ dead caes} + \#\text{ recovered cases in observed data}}$$

3.3 Prediction using SEIR Model

In the initial days of the epidemic we would want to know when and where the cases peak and when it will die out. We will fit training data to find value of β. The fitting is done in the similar way as we did for the entire data. Here to check if the model works, we will use first 75 days of the Qatar data. We have done this because we know how the epidemic has progressed in Qatar. Below is the predicted graph for the training data.

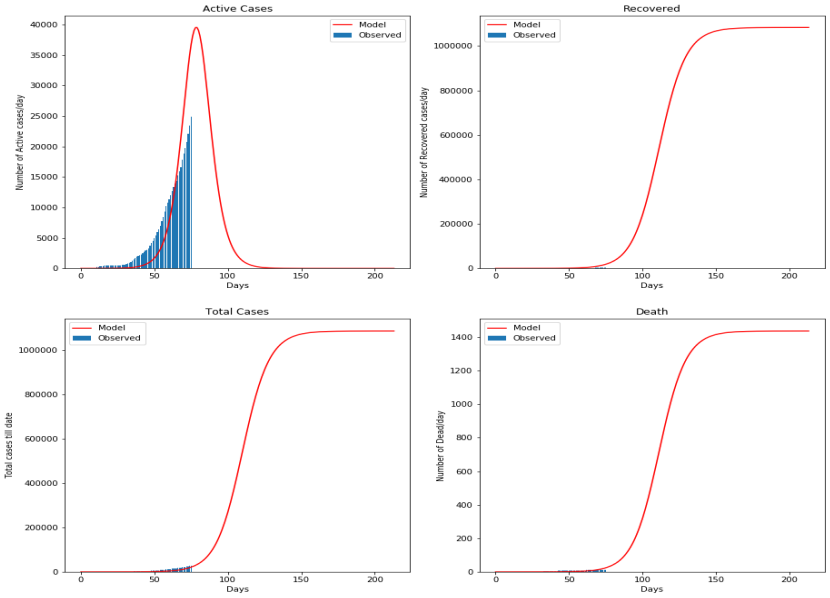
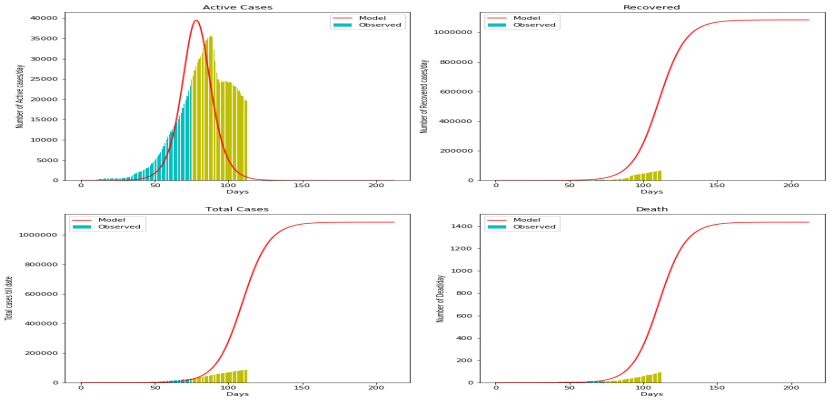


Figure 4: Fitting on training set

With the same initial conditions, we get the following parameters:
$\beta = (R_0\gamma)K = (R_0\gamma)(0.318) = 0.0848$, $R_{\text{eff}} = 1.272$
Next, we will plot the rest of the data along with the training data. It is observed that peak we have predicted is higher and towards the left as compared to the peak of data. From this it is inferred that there has been a change in β after day 75.



3.4 Lockdown Analysis

Since σ and γ are constant. Any change in lockdown status will only change β. Stricter the lockdown, lesses the β.

Let us define a "lockdown parameter" $= \alpha \ \in [0, 1]$, where $\begin{cases} \alpha = 1, \text{ is 100\% lock down} \\ \alpha = 0, \text{ is 0\% lock down} \end{cases}$

We have to model β as a function of α.

Input:

1. Training Data:= Whatever data we have till date. We want to see what will happen if lockdown is relaxed or strictened from the next day only.

2. Day of lockdown change/imposed (i.e. change in α) $= t_{LD}$

3. For the training data there will be an 'α_0' corresponding to 'β_0' we fit it with. 'α_0' has to be user defined. But cannot be 1 as this would imply there will be no spread of disease.

By fitting the training set, we get,

$$\beta_0 = R_0 \gamma K_0$$

3.4.1 Modelling $\beta(\alpha)$

Since β decreases with increase in α, we take

$$\beta \propto (1 - \alpha)$$

Since $\beta = R_0 \gamma K$, where $R_0 \gamma$ is constant. So,

$$K \propto (1 - \alpha)$$

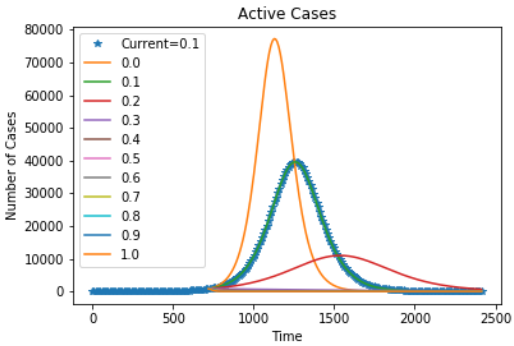
Hence given K_0 and α_0, we can have K from unitary method

$$K = K_0 \frac{(1 - \alpha)}{(1 - \alpha_0)}$$

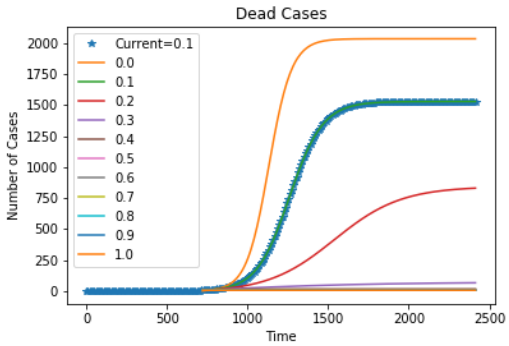
where α is the lockdown parameter we want to impose on t_{LD}.

$\implies \beta(\alpha) = R_0 \gamma K_0 \frac{(1 - \alpha)}{(1 - \alpha_0)} = R_0 \gamma K$, and $R_{\text{eff}} = R_0 K_0 \frac{(1 - \alpha)}{(1 - \alpha_0)}$ is the effective reproduction number when lockdown is imposed.

Now let's see plot for different α_0 value for $t_{LD} = 75$ for Qatar data.

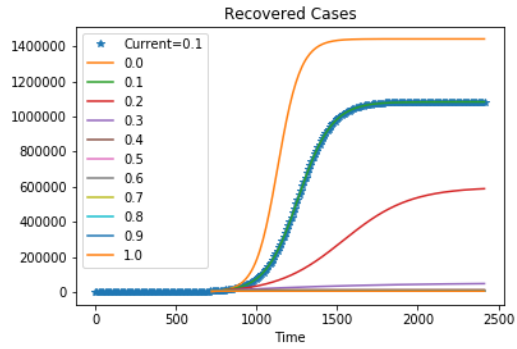


(a) $\alpha = 0.1$, Active cases

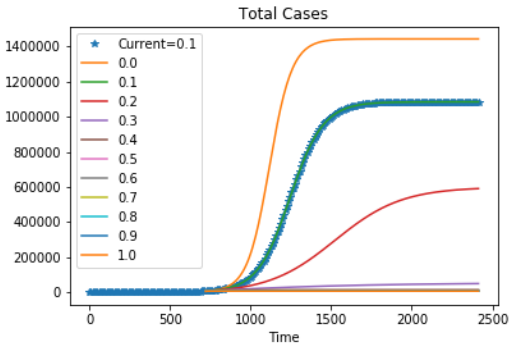


(b) $\alpha = 0.1$, Death Cases

Figure 5: Analysis for Active and Dead cases for $\alpha = 0.1$

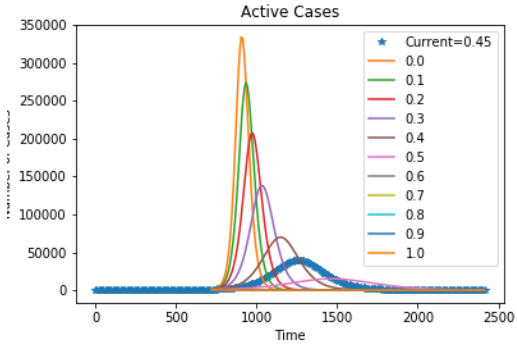


(a) $\alpha = 0.1$, Recovered cases

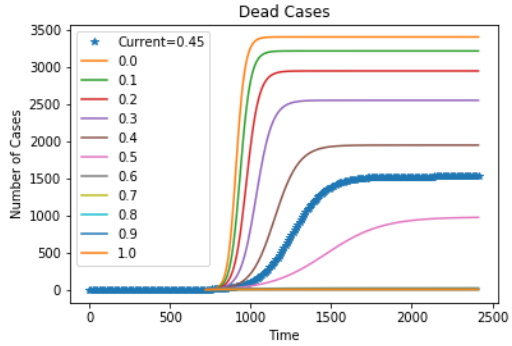


(b) $\alpha = 0.1$, Total Cases

Figure 6: Analysis for Recovered and Total cases for $\alpha = 0.1$

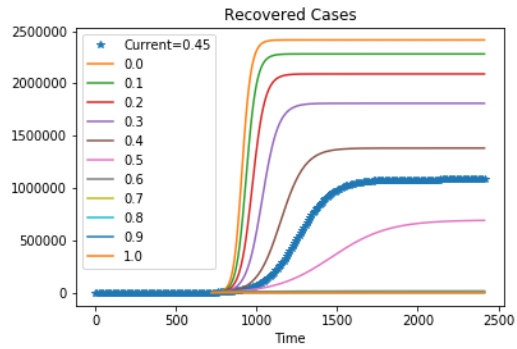


(a) $\alpha = 0.45$, Active cases

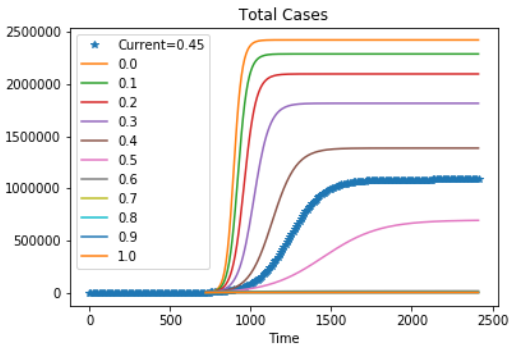


(b) $\alpha = 0.45$, Death Cases

Figure 7: Analysis for Active and Dead cases for $\alpha = 0.45$

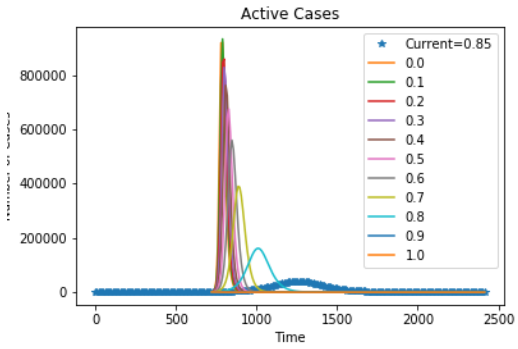


(a) $\alpha = 0.45$, Recovered cases

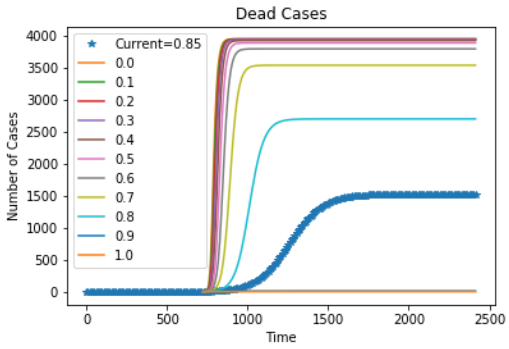


(b) $\alpha = 0.45$, Total Cases

Figure 8: Analysis for Recovered and Total cases for $\alpha = 0.45$

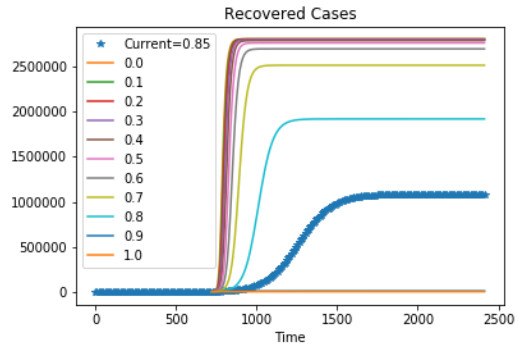


(a) $\alpha = 0.85$, Active cases

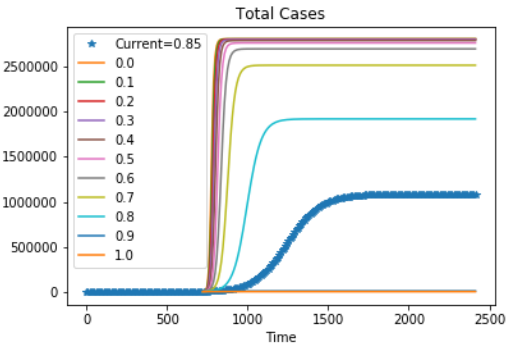


(b) $\alpha = 0.85$, Death Cases

Figure 9: Analysis for Active and Dead cases for $\alpha = 0.85$



(a) $\alpha = 0.85$, Recovered cases



(b) $\alpha = 0.85$, Total Cases

Figure 10: Analysis for Recovered and Total cases for $\alpha = 0.85$

3.5 Discussion

- Change in α is actually the consequence of lockdown. Also, $\alpha = 0$ and $\alpha = 1$ is very unlikely to happen.

- The initial value α_0 is very significant. It gives insights about the country's lockdown efficiency and how the disease spreads in that country.

- Note that when we do not change lockdown status at t_{LD},

$$\beta(\alpha) = R_0 \gamma K_0$$

3.6 Shortcomings for SEIR Model

- Since this mathematical model has parameters which are based on the average only, it can not really capture human interaction. We are working with a collection of average values of parameter. So, it is bound to have error.

- Measuring α is problem. We can't know what exactly α means, policy wise.

- Implementation of the lockdown. α has to be continuously monitored.

- Since the pandemic is going on, we do not have definite data to test and use for future pandemics.

- N isn't constant for a given country.

- The parameters affecting β aren't exhaustive.

4 SEIR Model on Barabasi-Albert Network

4.1 Generating Barabasi-Albert Network

The Barabási–Albert (BA) model is an algorithm for generating random scale-free networks using a **preferential attachment mechanism**. Several natural and human-made systems, including the Internet, the **world wide web**, **citation networks**, and some **social networks** are thought to be approximately scale-free and certainly contain few nodes (called hubs) with unusually high degree as compared to the other nodes of the network. The BA model tries to explain the existence of such nodes in real networks. The algorithm is named for its inventors **Albert-László Barabási** and **Réka Albert** and is a special case of an earlier and more general model called Price's model.
Preferential attachment means that the more connected a node is, the more likely it is to receive new links. Nodes with a higher degree have a stronger ability to grab links added to the network.
Let's discuss the algorithm for obtaining Barabasi-Albert Network.

Algorithm 1: Algorithm for Barabasi-Albert Network

Result: We obtain a scale-free Barabasi-Albert network
The network begins with an initial connected network of m_0 nodes;
while $\{Total\ Nodes < n\}$ **do**
- New nodes are added to the network connects to $m < m_0$ existing nodes with a probability that is proportional to the number of links that the existing nodes already have. Formally, the probability p_i that the new node is connected to node i is $p_i = \dfrac{k_i}{\sum_j k_j}$ where k_i is the degree of node i and the sum is made over all pre-existing nodes j;

end

4.2 Applying SEIR Model on BA Network

Basic Idea:
We will name our nodes(indexing) and divide them into 4 compartments(list) namely, Exposed(E), Infected(I), Removed(R), Susceptible(S) and track them eventually with time. Before we will discuss the algorithm, let's know some important terms that we've used in our algorithms.

- **σ-Influence Model**: It shows the fact that the infection probability increases faster than a linear infection. In σ-influence model, the probability that a exposed node will be infected by one of its infected neighbors is σ, and the infections from all its infected neighbors are independent. So if a exposed node has n infected nodes at time 't', the probability that it will be infected at the next time by its n infected neighbors is[ZHA15]

$$p = 1 - (1 - \sigma)^n$$

- γ: This gives us the "chance" of Infected entering into Removed group. It is a threshold kept between 0 and 1 and for each infected individual we generate a random number between 0 and 1. If that is less than the threshold , then infected goes in removed group.

Let's see in brief the algorithm we have used for tracking each compartment's size with respect to time.

Algorithm 2: Algorithm for Tracking each Compartment's size

Result: A 4×time array where 4 = # of Compartments and time = Total time
Initialization;
- Indexing of nodes of network;
- Randomly choose one node and put it in Infected Group;
- At time = 0, take # of nodes in Exposed = 0;
- Also # of individuals in Removed Group = 0;
- Susceptible Group = Indexing set - (Exposed + Infected + Removed);
- Time = t = 0;

while {# *of Infected* $\neq 0$} **do**
  if {# *of exposed* $\neq 0$} **then**
    - Randomly choose nodes (person) from Exposed with probability being different for each node following σ-**influence model**;
    - Put selected ones in I;
  end
  - Randomly choose individuals from I with some probability threshold(γ) and put it in Removed Group;
  - Now put all those nodes, connected to nodes in Infected Group, in Exposed;
  - Set Susceptible = Indexing Set - Exposed - Infected - Removed;
  - $t+ = 1$;
end

4.3 Plot from Networked SEIR Model

- For $n = 10000$, $m = 25$, $\sigma = 0.5$, $\gamma = 0.2$

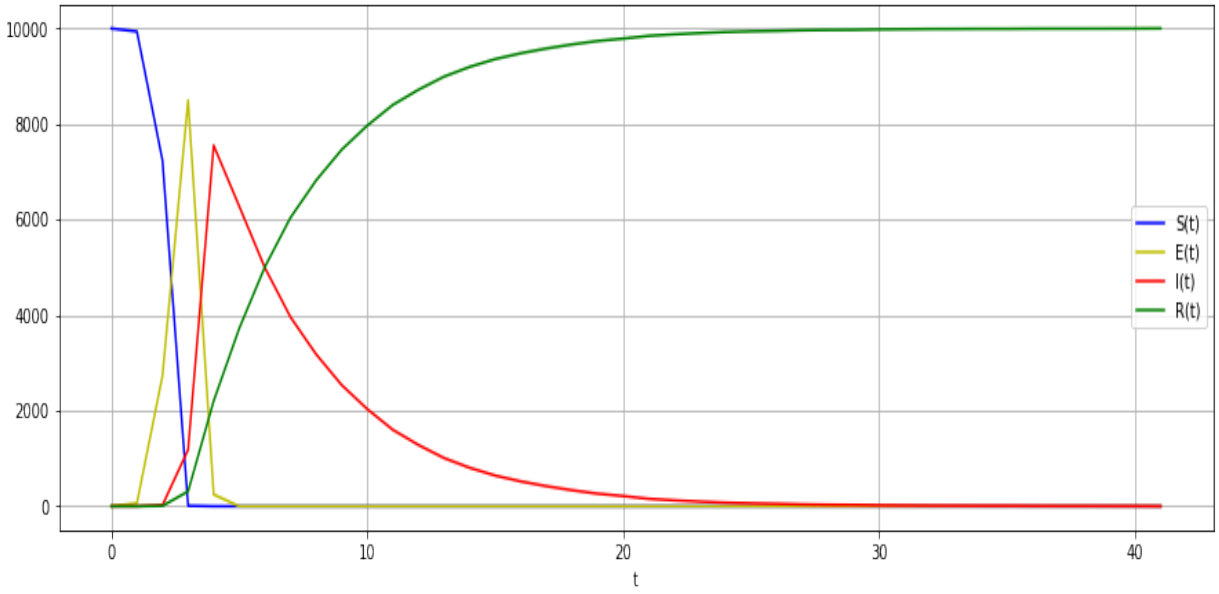


Figure 11: Plot of Compartment's size with time obtained from BA

Now we calculate β from network by taking **average degree** of network and the **total**

nodes,

$$\beta = \frac{\text{average degree}}{\text{total \# of nodes}}$$

Since each invidual (node) can contact β individuals per unit time, then for n number of nodes, it would be $n\beta$. So each node is contacting $n\beta$ individuals (per unit time) which means each node has $n\beta$ edges on average.
Using that β value, and σ and γ and putting in the system of differential equations we have for SEIR, the plot we will get from it's solution is as follows:

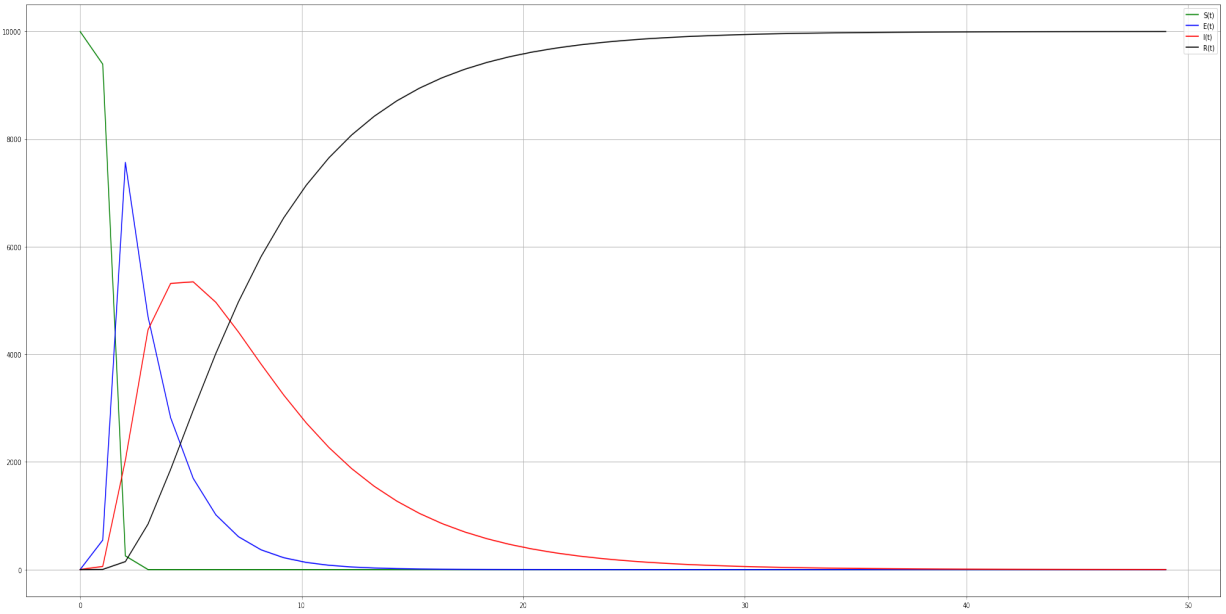


Figure 12: Plot for corresponding parameters obatined from ODE's solution

- For $n = 10000$, $m = 50$, $\sigma = 0.35$, $\gamma = 0.2$

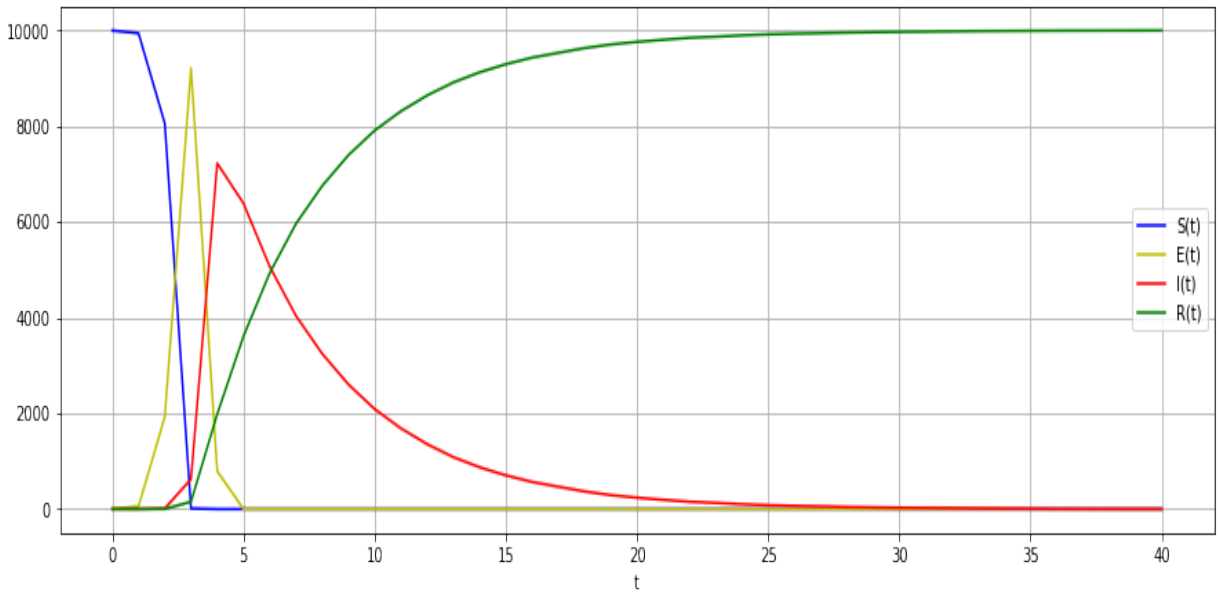


Figure 13: Plot of Compartment's size with time obtained from BA

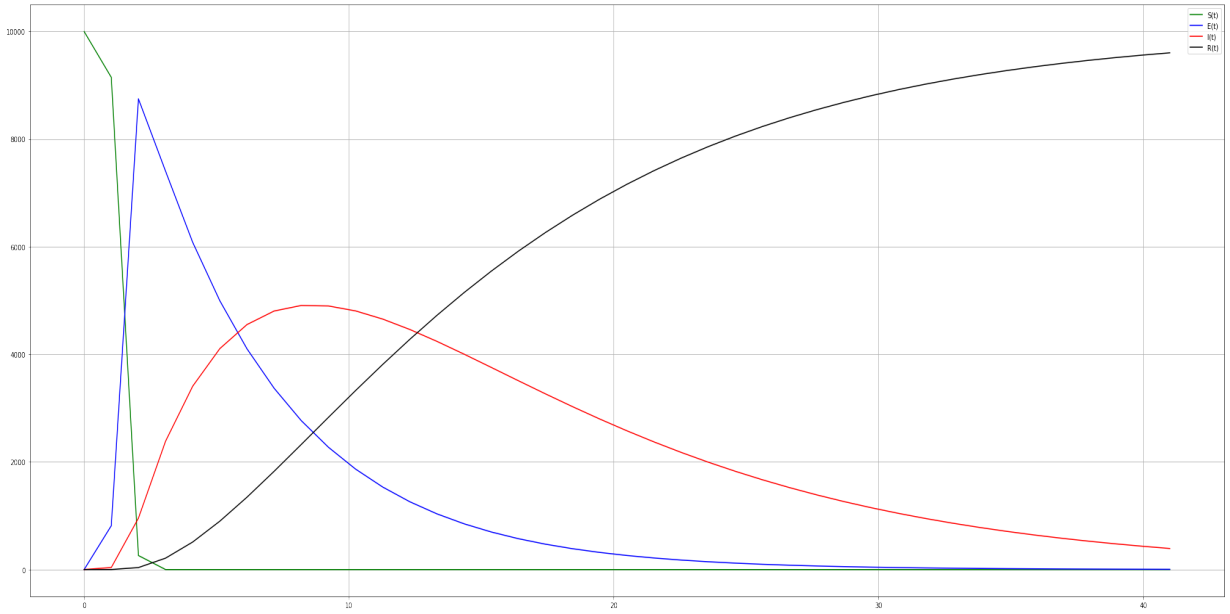


Figure 14: Plot for corresponding parameters obatined from ODE's solution

4.4 Shortcomings for Netwoked SEIR Model and for Whole Project

1. Large country's data pose many more complexities while fitting because it involves the following reasons:

   - Large countries consists of large number of communities, where some of them are isolated and we take them into account while constructing Susceptible group. So our assumption of well-mixed population fails.
   - Each region/province is not equally affected which again inculcates non-uniformity in data.
   - Pandemic is not over yet, and from the data we have, since it follows a non-linear pattern, it's hard to say when it will over. So fitting our model on real-time data put a big challenge before us. If we have complete data, then we are sure to capture non-linearity in our model by adjusting parameters accordingly.

2. Now coming to Network Part:

   - The Barabasi-Albert network we have used represents a single well-mixed community, which don't capture the whole city/country human interactions due to different communities. Hence our effort to fit New-York data, which involves high density cases, fails.
   - And to take multiple Barabasi-Albert model in order to represent human-society, we're not fortunate enough to have sufficient computational power. Hardly we can do computations for 10,000 nodes, which again takes 20 min on average.

5 Conclusion

- In pandemic like this, providing timely information to the public is paramount. A better modelling of COVID-19 will assist govt. and authorities to disseminate verified articles, provide updates to the situation and advocate good personal hygiene to the people.
- And also this will spread awareness to the common people by providing scientific-based data analysis, prediction and verified news.

- Coming up with good models is the need of the hour. One has to have a sound knowledge of Economics, Public-Policy along with the necessary tools from Machine Learning and Network analysis to improve the existing ones.

References

[ZHA15] HONGGANG WANG ZHAOYANG ZHANG. "Modeling Epidemics Spreading on Social Contact Networks". In: *IEEE Transactions on Emerging Topics in Computing* (2015), pp. 410–419. DOI: https://ieeexplore.ieee.org/document/7029011.

[Dri17] Pauline van den Driessche. "Reproduction numbers of infectious disease models". In: *Department of Mathematics and Statistics, University of Victoria, Victoria, BC, V8W 2Y2, Canada* (2017). DOI: https://doi.org/10.1016/j.idm.2017.06.002.

[Ham20] Fairoza Amira Binti Hamzah. "CoronaTracker: World-wide COVID-19 Outbreak Data Analysis and Prediction". In: *Bulletin of the World Health Organization* (2020), pp. 1–32. DOI: http://dx.doi.org/10.2471/BLT.20.255695.

6 Acknowledgements - Team Work

Glad to say that our every team member worked whole-heartedly and co-operated throughout. Everyone has contribution in every field. So here names of those, who contributed most in which area, are mentioned.

6.1 Report Writing

Initial part for report writing, from Introduction to Mathematical Representation of SEIR Model has been done by **Santanu** and **Siddhikant** and the rest are by **Aditya**.

6.2 Presentation

It's done by **Siddhikant**, **Santanu** and **Shaharica** mostly.

6.3 Coding

One part of coding i.e. SEIR Model fitting on data through ODE equations has been done by **Shaharica**, while the Networked Model has been done **Aditya**.