**Simulation of different stages of Covid-19 in Pennsylvania
using logistic growth, SIRD model and branching process**

Abstract:
        The global people have been experiencing serious disasters since 2019 due to the widespread novel disease SARS-CoV-2, known as COVID-19. The infection has caused severe syndrome and becomes a global crisis. Predicting an epidemic will be critical to suspend the spreading and reduce its global impact. Because of lacking concerts and records when the pandemic first broke out, it is difficult to anticipate the size of the spreading using regular methods. In this paper, three epidemic models are presented and the research is focused on the Pennsylvania area. The key concerns for three main aspects of the pandemic are considered and three specific strategies and their parameters are established and compared. By using the reported data in PA, the models simulate the cumulative cases, existing cases, and spreading after vaccination. It gives a constructive understanding of vaccination.

Introduction:
        CoronaVirus Disease 2019, also known as COVID-19, is a new worldwide disease and a challenging epidemic. Since the first a few cases with symptoms like fever, cough, fatigue have been discovered in Wuhan, China from November 2019, this disease quickly spread to the whole country and the neighboring countries. Today, Covid 19 is a continually spreading disease all over the world and a hot topic for people to discuss on the internet. The existence of the disease has a largely negative impact on people's lives and society's functioning. Understanding the trend, spreading development, and the inflection point is of great interest for policymakers and public health organizations to make decisions. There are two most important aspects of the epidemic people would pay attention to 1) the cumulative cases in the pandemic, 2) the existing cases within an area. On the other hand, as the vaccinations of Covid-19 have been put into use since January 2021, we observed that the cases had been slightly decreasing as more people get vaccinated. Exploring the effect and predicting the spreading after the vaccination would be significant. Since then, this paper seeks to model these three stages of the spread of Covid-19 and how the tendency could affect society and individual decision-making.

Method:
1: Using the logistic model to predict the cumulative effect
        Under this method, we examine the application of the logistic model in the pandemic in Pennsylvania, and the least square and curve fit methods were employed to train the model. After estimating the parameters, we did a short prediction over 10 days and 30 days. The logistic model has been widely used to model population growth and bacteria growth nowadays. The assumptions are as follows: the carrying capacity, in this case, would be the maximum number of people who will get infected when time goes to infinity. Since then, by considering external factors like quarantine, isolation, and vaccination, the pandemic will eventually extinct. The model has also been proved to

have a good approximation over China and some other Asian countries. And the estimation would be unreliable for the early stage of outbreaks(Shen, 2020).

The Python image plotting program, called "logistic1.py" and "logistic2.py", was developed to model the overall infected cases of the pandemic. It first inputs the infection data in Pennsylvania and plots cumulative infected cases in terms of time(days) to show the overall trending and spreading of the disease. The image identifies a logistic growth of the cumulative cases as exponential property is shown at the beginning, and the convergence comes afterward. In order to use the logistic model, we solve the differential equation:

$$\frac{dP}{dt} = rP\left(1 - \frac{P}{K}\right)$$

And three parameters carrying capacity K, initial infected cases P_0, and growth rate r are needed to fit the data:

$$P(t) = \frac{KP_0 e^{rt}}{K + P_0\left(e^{rt} - 1\right)} = \frac{K}{1 + \left(\frac{K - P_0}{P_0}\right) e^{-rt}}$$

The program uses the "scipy.optimize" and "sklearn.metrics" package to find the approximated fit curves. The python code "logistic1.py" takes in the data from March 3 2020 to March 7, 2021, and makes a range statement for carrying capacity K and growth rate r within the desired value range. Leveraging the recursion to test the least square error and the using method "curve_fit" to find the most fittable curve with parameter P_0, the code identifies the three most fittable parameters for P(t). Fig.1 shows the data and the prediction line while there are differences at the early stage. Because exponential growth is presented near 200 days after the occurrence of the virus, we focus from there to get a more precise model. Also, due to the lack of attention to the virus and the imperfect information collection system in the early stages, there are missing data before 200 days. The python code "logistic2.py" takes in the data from September 18, 2020, to March 7, 2021, and uses the same strategy for modeling. As a comparison, Fig.2 shows a better approximation. And carrying capacity K, initial cases P_0, and growth rate r present 1040000, 69580, and 0.03, respectively. The prediction gives an approximation for the maximum number of infected people which would be K 1040000. Also, the initial cases are not 1 because we are assuming that the number of cumulative cases after 200 days is 69580. Additionally, in the next section, we develop variance of the SIR model to keep track of the existing cases. As in the logistic growth, the number of cumulative infected people would eventually converge to K as time goes to infinity. K could be a good

approximation to the number of initially suspected people as the number of infected cases can never go beyond the S_0.
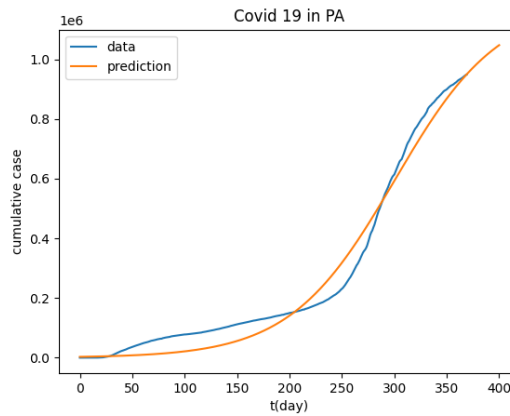

Fig.1 logistic modeling using data from March 3, 2020, to March 7, 2021.
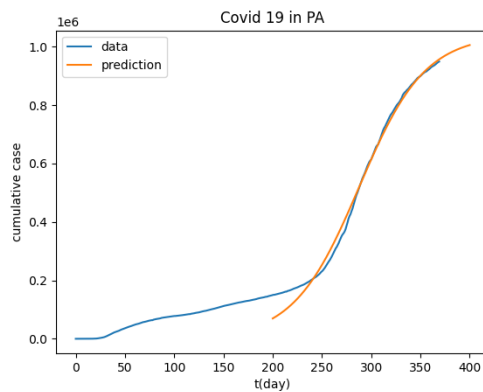

Fig.2 logistic modeling using data from September 18, 2020, to March 7, 2021.

Now, to test the accuracy of the model, we train the model with data from November 8, 2020, to February 25, 2021, and check the prediction error within 10 days. We also extend the range to 30 days and check if the model could be used to do a longer prediction.

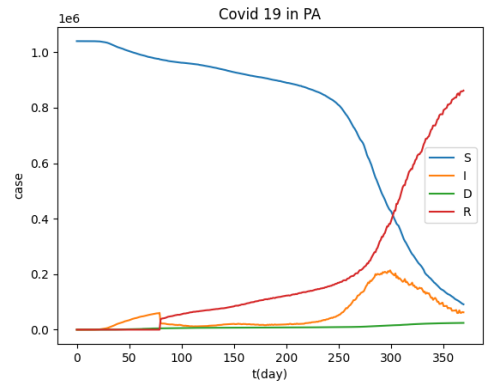2: Using the classic model SIRD to model the existing cases

Fig.3 Change of susceptible, infected, recovered, and dead people in Pennsylvania, marked as S, I, R, D respectively.

Jesús and Charles have proven that the SIRD has a good approximation for the Covid-19 in many states like New York. In the classical dynamic SIRD model, the total population N is divided into several different compartments depicted by the right image, namely, susceptible (S), infected (I), recovered (R), and dead(D). N is a constant and we assume its value won't vary.
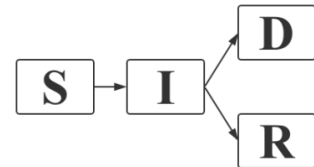
By establishing a system of the SIRD model, we can deal with the number of susceptible, infected, recovered and dead people. The governing equations of this model are:

$$\frac{dS}{dt} = -aIS$$
$$\frac{dI}{dt} = aIS - bI - cI$$
$$\frac{dD}{dt} = bI$$
$$\frac{dR}{dt} = cI$$

The parameters a, b, c are infection rate, death rate, and recovery rate, respectively. The python code "SIRD.py" plots the spreading data of susceptible, infected, recovered, and dead people in Pennsylvania. Then it simulates the pandemic with SIRD models using the package "scipy" and method "curve_fit".
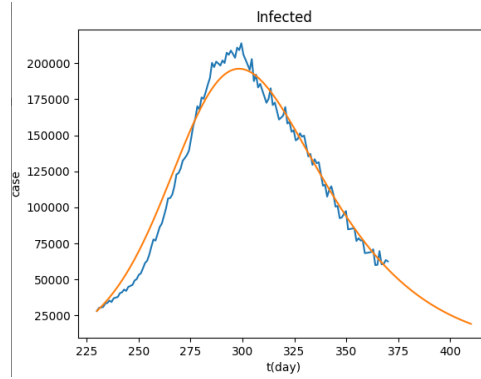
Fig.3 Prediction model for existing infected people in Pennsylvania using data from 230 days to 370 days

Varying the values for the parameters, we find the best fit curve for the existing infected population. The errors were minimized during the fitting process. Here we also want to test the accuracy of the model. As we did before, we test the error for 10 days and 30 days intervals.

3: Using Branching process to develop the model for vaccination

The branching process is a kind of stochastic individual-based process. After Galton and Watson first came out of the idea of exploring the probability of family name extinction, the branching process has been studied and learned for years to describe population dynamics. Now more researches use it to model the early stage of virus spreading. The benefit of using a stochastic model than deterministic models discussed above is that it tends to have variability and can create a better prediction.

In a branching model, each generation size is based on the size of the last generation. And we collect the i.i.d random variable for each individual of the last generation and count the number of the current size:

$$Z_{n+1} = \sum_{i=1}^{Z_n} X_{n,i}$$

Zn = 0 is considered as absorbing state. Considering each point in the branching process as an infected individual and its offspring state as the people infected by the previous person, we model the pandemic with the branching process.
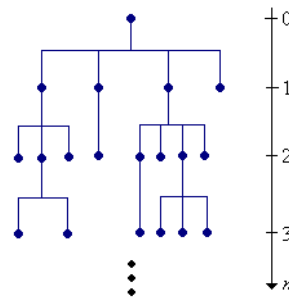
Fig.4 Branching process with generation 0, $Z_0 = 1$, generation 1, $Z_1 = 4$, generation2, $Z_2 = 8$ ...

Normally, it takes an average of 2-6 weeks to recover from the disease, while the average death period is about one month for critical patients. So we assume that after 6 weeks, 42 days, the infected people will be removed, either by recovery or death, which means that the existing infected populations are newly infected people other than those who have been infected 42 days ago. Starting from 42 days after the first case has been recorded, we would minimize the uncertainty in the early stage. Then we consider the 42-day interval as a new generation period and count the number of offspring for each round by dividing the news recorded cases by the existing cases at the beginning of each period. Fig.6 shows the results and the average value is roughly 2.5, which means, on average, 2.5 people would be infected by one person infected 42 days ago. Also, we calculate the average value before the vaccination which is 2.72 and it could convey a more informative message in the later study. It also indicates that using the branching process to simulate the pandemic would be reasonable within a short period of time since the offspring means are stable within that time.
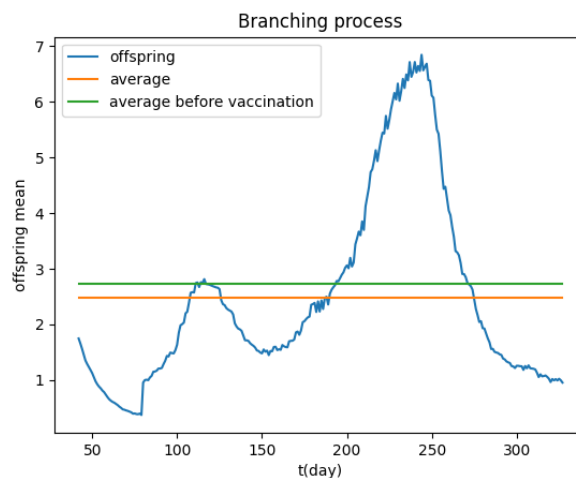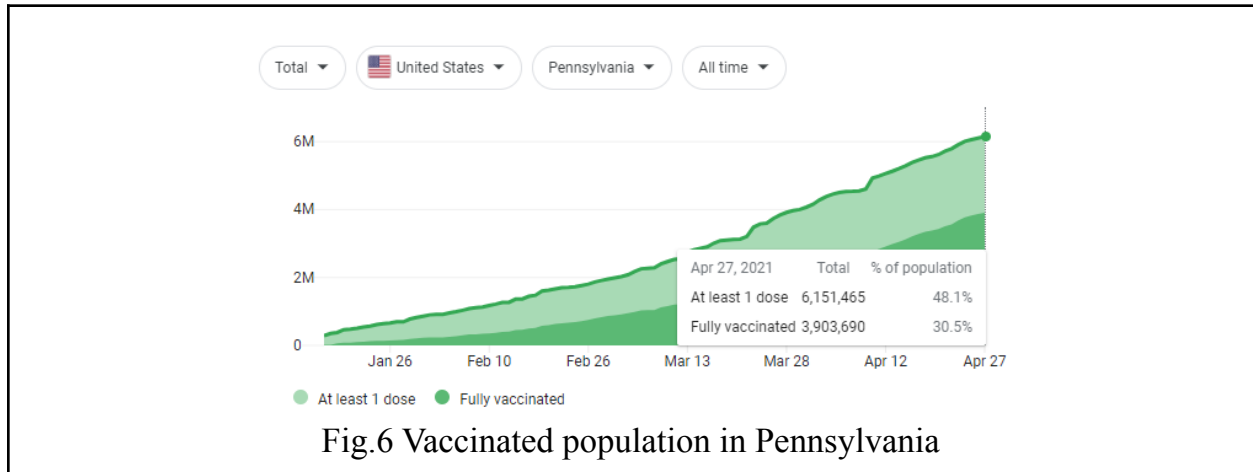

Fig.5 Branching process offspring mean using 42-day interval and the average value in PA

Now, we could use the lastest offspring mean,0.956, to shortly predict the pandemic within a week after that. We also see that the extinction probability of the disease would be 0.044 after 42 days.

Fig.6 Vaccinated population in Pennsylvania

Another vision of the branching process is that it leads to a better understanding of vaccination. Up to April 27, 2021, about 30% of populations in PA have been vaccinated. Fig4 models the branching process model with a mean equals to 2 for disease spreading before vaccination. Fig.6shows the same model as the scenario after vaccination. As a comparison, it could be seen that with a 30% Immune crowd, the infected groups of people have reduced by more than half. Although the model only gives an approximation for 3 generations with a hypothetical mean value, it presents certain insights for us to identify the model for long-term behavior.
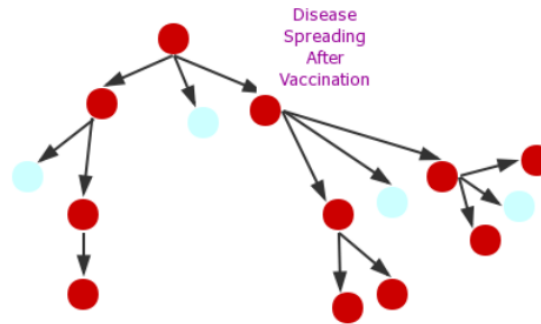

Fig.4 Branching process model with **μ**=2 for Disease spreading before vaccination
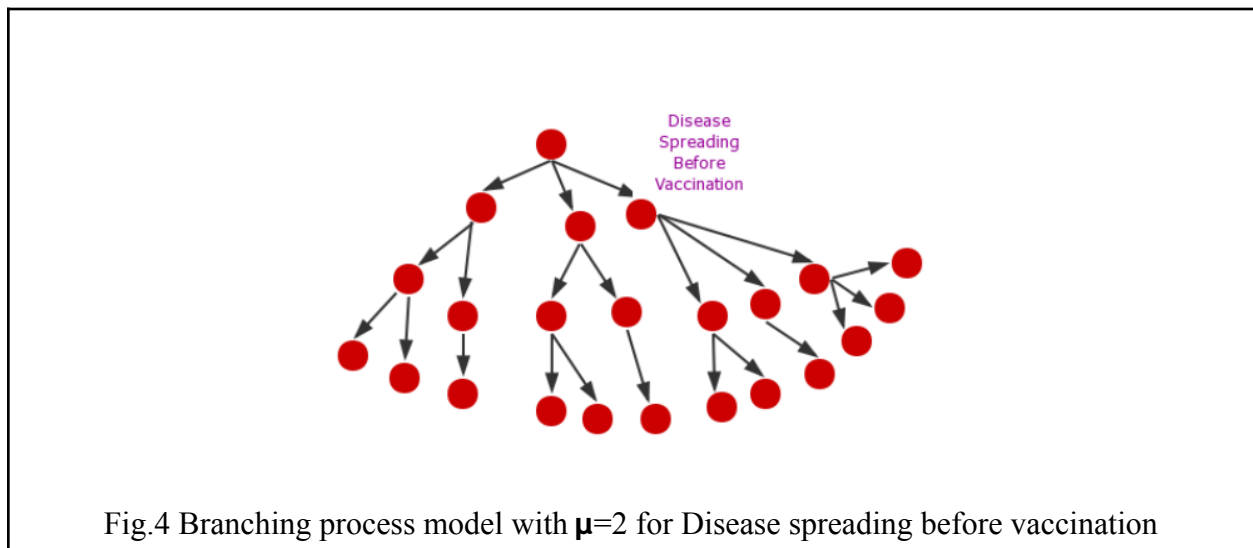
Fig.5 Branching process model with **μ**=2 for Disease spreading after vaccination

MRC Centre for Global Infectious Disease Analysis at the Imperial College London reported that the coronavirus has an R_0 of 1.5 to 3.5, which means one infected person can infect, on average, from 1.5 to 3.5 people. The report could prove the previous works of the paper while we found the average value before vaccination is 2.72. Normally in a branching process, if the mean value **μ** is less than the 1, we would assume that it would 100% extinct in the long term. Otherwise, the transmission of the coronavirus will increase in the population if **μ** is greater than the 1. Also, the vaccination rate shows a linear growth with a rate of 1% per 3 days. Then we solve the equation $2.72 - 2.72p < 1$, it could be seen that the virus would eventually stop spreading if the proportion of vaccinated people has been larger than or equal to 63.24%. And that happens after 100 days since April 27.

In order to test the approximation, short-term and long-term behavior are being considered. Python code "Branchingtest.py" and "Branchingtest2.py" subtract the data 42 days ago from the current data and divided it by the existing data 42 days ago to calculate the offspring using a 42-day interval. Then it uses the predicted offspring value to predict the value of offspring on the next day, and the new prediction is used again as the current cumulative data to do the prediction 10 times, and 30 times after that.

Once all the errors have been calculated and visualized, we compare the percentage of error in three models.

Results and discussion:
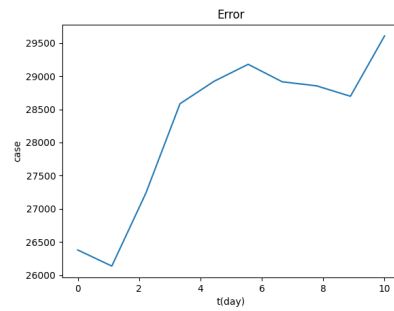
Fig.6 Error in the logistic model
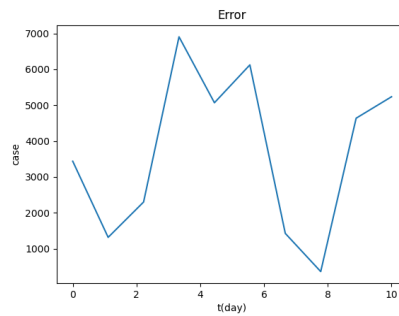
Fig.7 Error in the SIRD model
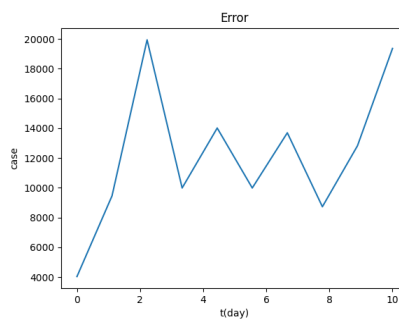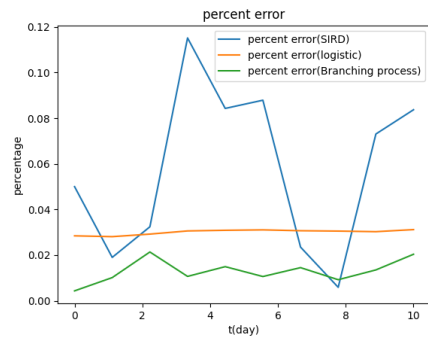

Fig.8 Error in the Branching process model

Fig.9 Comparison of the three models' percent error. Blueline represents the error in the SIRD model; Redline represents the error in the logistic model; Greenline represents the error in the branching process when predicting for 10 days.
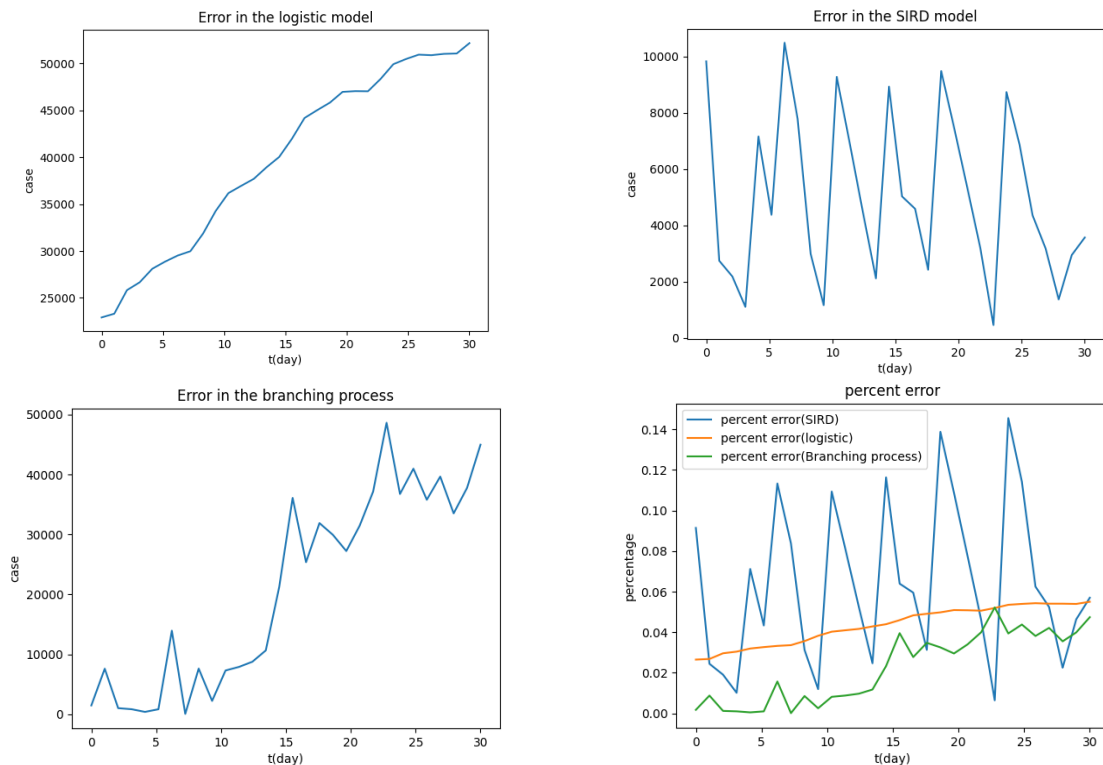

Fig.10 Error when predicting for 30 days

In the short term, we predict the pandemic and show the error in 10 days prediction. The SIRD model has a larger variability (depicted in Fig.7) and the percentage of error is between 0.59% and 11.52%, which seems to fluctuate around 6%; While the logistic model's accuracy decreases as time goes by (depicted in Fig.6) and the percentage of error is between 2.84% and 3.12%. The branching process identifies a relatively smaller variability but fluctuation could be seen. The percentage of error in the model is between 0.4% and 2.1%. The branching model performs pretty well(depicted in Fig.8) and is more stable compared with the other two models. It could be seen that the intersections occur frequently so three models could be used together to make better predictions while the branching model, in the short term, is the most reliable as we assumed. In the long term, we do the same process except that we show the error in 30 days prediction(depicted in Fig.10). The results would not be surprising. Linear growth could be seen in both the error of the logistic model and the branching model while the increasing rate of the error in the branching process is greater. As a comparison, the percent error of the SIRD model would fluctuate around a stable and reasonable value. As time goes by, we would expect that the performance of the SIRD model would increase over a longer period of time. However, in a short time, the branching model makes more sense.

Accordingly, all models have been proved to have varying degrees of accuracy. The main issue found during the building of the models was the lack of additional data about the exposed or hospitalized people: if more compartments could be extended to describe the pandemic, the prediction could be more precise. Because of the large variability in the pandemic, every model has some drawbacks and limitations. To reemphasize, the logistic model simulates the overall cumulative cases in PA; SIDR shows the dynamic pandemic and predicts the existing cases; Branching process helps to model the spreading within a relatively short period of time and indicates the extinction time and probability, which creates a better understanding of vaccination effect.

Limitation:

The branching process can only approximate the very short time period of spreading and the accuracy will decrease as pandemics continues to spread. Vaccinations are not 100% effective as reported by CDC. Irregular vaccine supply will also produce uncertainty. Also, in the long term, the offspring mean μ in the branching process would change as people's awareness of epidemic prevention is strengthened or more people get vaccinated. More factors like birth and immigration could be taken into account. Future research could be focused on adding time intervals to the edge of the branching process. The continuous Markov chains are worthy of further investigation and μ could be made into a variable instead of a constant to predict the longer period of time. I also expect to use the offspring mean found in the paper with a specific offspring distribution to train the data and make simulations. Additionally, because of the lack of attention to the virus

and the imperfect information collection system in the early stages when the virus first spread out, many cases were not taken into consideration. So the early detection contains much missing important information. SIRD models could also be refined by adding more parameters and groups of interest. And time series could also be taken into account in future research.

The region of study is limited to Pennsylvania, the models could be applied in other areas for further tests. We believe that homogeneously spread would not be seen across states and the parameters and errors would be different over different regions of the country.

Reference:

[1]Jacob, Christine. "Branching processes: their role in epidemiology." *International journal of environmental research and public health* 7.3 (2010): 1186-1204

[2]Shen, Christopher Y. "Logistic growth modelling of COVID-19 proliferation in China and its international implications." *International Journal of Infectious Diseases* 96 (2020): 582-589.

[3]Fernández-Villaverde, Jesús, and Charles I. Jones. *Estimating and simulating a SIRD model of COVID-19 for many countries, states, and cities*. No. w27128. National Bureau of Economic Research, 2020.

[4]Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

[5] Virtanen, P., Gommers, R., Oliphant, T.E. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2

[6]The Atlantic Monthly Group. "Data API." *The COVID Tracking Project*, covidtracking.com/data/api. Accessed 4 May 2021.

[7]Global Change Data Lab. "Coronavirus (COVID-19) Vaccinations - Statistics and Research." *Our World in Data*, ourworldindata.org/covid-vaccinations?country=CHN. Accessed 4 May 2021.

[8]CDC. "COVID-19 Vaccination." *Centers for Disease Control and Prevention*, 11 Feb. 2020, www.cdc.gov/coronavirus/2019-ncov/vaccines/reporting-systems.html.

[9]"MRC Centre for Global Infectious Disease Analysis." *MRC Centre for Global Infectious Disease Analysis*, Imperial College London, www.imperial.ac.uk/mrc-global-infectious-disease-analysis. Accessed 5 May 2021.