COVID-19 Prediction and Tweets Analysis

Gaomin Wu
gw1107

Haoxue Li
hl3664
uploading submissions

Yunya Wang
yw4509

October 26, 2020

1 Introduction

The novel COVID-19 is declared as a pandemic in March 2020. We are interested in the spreading pattern, factors to stop spreading of COVID-19 transmission. Additionally, as it starts slowing down now, we also investigate the change on the social media after the darkest period. Accordingly, we built epidemic model (SEIR) based on US data, provide estimations of the basic reproduction number (R0), analyze the twitter mobility index and its relationship with R_0. Lastly, we conduct some text analysis on COVID-19 related tweets and investigate the topic trend on twitter.

2 SEIR Model

2.1 Methodologies of SEIR

We implement the classical SEIR model used widely in epidemiology studies. In the model, we introduce 6 states, susceptible (S), exposed (E), infected (I), critical (C) (people in critical condition), recovered (R) and dead (D). Fig.1 explains the transition between the six states and their corresponding relationship and parameters used.

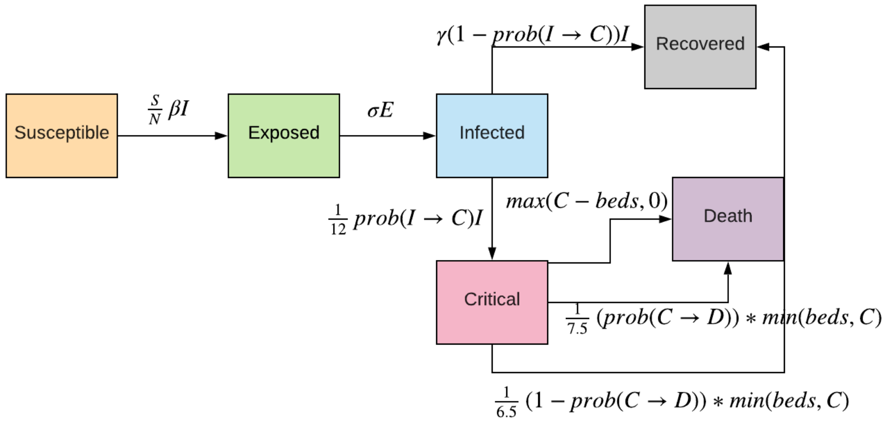

Susceptible
Exposed
Infected
Recovered
Death
Critical
$\frac{S}{N}\beta I$
σE
$\gamma(1-prob(I \to C))I$
$\frac{1}{12}prob(I \to C)I$
$max(C - beds, 0)$
$\frac{1}{7.5}(prob(C \to D)) * min(beds, C)$
$\frac{1}{6.5}(1 - prob(C \to D)) * min(beds, C)$

Figure 1: SEIR Component and Parameter

The following is the explanations of the parameters and relationship used in the chart:

- N (total population of the studied ares)
- From susceptible to exposed: $\frac{dS}{dt} = -\beta(t)I\frac{S}{N}$
- From exposed to infected: $\frac{dE}{dt} = \beta(t)I\frac{S}{N} - \sigma E$
- From infected to critical or recovery: $\frac{dI}{dt} = \sigma E - \frac{1}{12}prob(I \to C)I - \gamma(1 - prob(I \to C))I$
- From critical to death or recovery [4]:
  $\frac{dC}{dt} = \frac{1}{12}prob(I \to C))*I - \frac{1}{7.5}prob(C \to D)min(beds(t), C) - max(0, C - beds(t)) - \frac{1}{6.5}(1 - prob(C \to D))min(beds(t), C)$
- Derivative of recovery to time t: $\frac{dR}{dt} = \gamma(1 - prob(I \to C)) * I + \frac{1}{6.5}(1 - prob(C \to D))min(beds(t), C)$
- Derivative of death to time t: $\frac{dD}{dt} = \frac{1}{7.5}prob(C \to D) * min(beds(t), C) + max(0, C - beds(t))$

  Here we include the scenario of limited resources (ICU beds). If the number of critical patients outweigh the number of ICU, we assume people not getting the ICU beds are going to die due to lack of treatment.

As for ICU beds, we treat it as a time dependent variable; $beds(t) = beds_0 + s * t * beds_0$; and we will search for the best parameter s through model fitting.

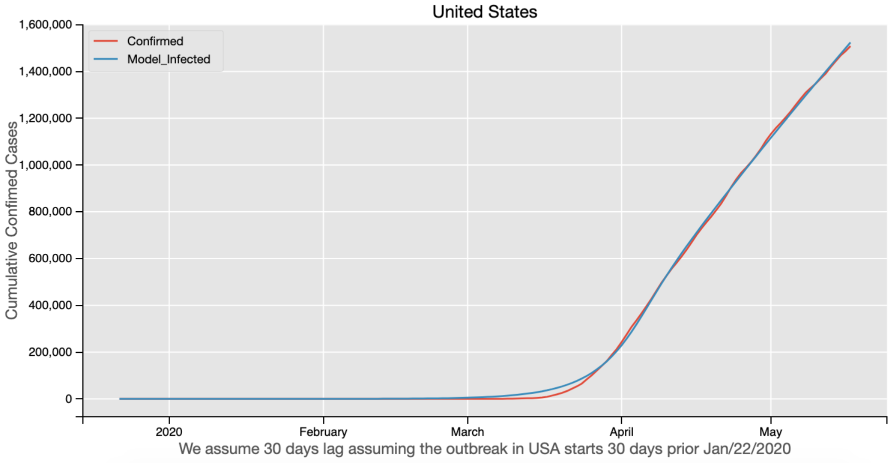
- Another time dependent variable is $\beta(t) = R_0(t) * \gamma$ or $R_0(t) = \frac{\beta(t)}{\gamma}$. R_0 is usually called reproduction number. If $R_0(t) > 1$, the infection speed is greater than the recovery speed and cases will go up; if $R_0(t) < 1$, the pandemic will slow down. Usually there are two ways of determining R_0 [**1**], one is estimating R_0 in an invasion way: $log(I(t)) = log(I_0) + (R_0 - 1)\gamma t$ or estimating R_0 from final size: $R_0(t) = \frac{R_{0\text{start}} - R_{0\text{end}}}{1+e^{-k(-x+x_0)}} + R_{0\text{end}}$. We experiment two ways and decided to use the second one and search for the best $R_{0\text{start}}, R_{0\text{end}}, k, x_0$ through model fitting.

2.2 Data used in SEIR

We use confirmed, death and recovery cases for US only from CSSEGISandData starting from 01/22/2020 to 05/18/2020. We also pull age distribution, number of ICU beds by countries and probability distribution (probability from infected to ICU and probability from ICU to death) by age groups from UN data, which will give us extra computing power to take in to consideration of age distribution and their corresponding death rate in the model.

2.3 SEIR Model Fitting

We use scipy.integrate.odeint to solve the differential equations listed in section 2.1 and use lmfit package to search for the best parameter where we input the initial, minimum and maximum values for each parameter. As for the fitting, we elaborate lmfit with target variable as total confirmed cases which is the sum of infected, deaths, critical and recovery. And we search for the best combination of variables based on least square error.


United States

We assume 30 days lag assuming the outbreak in USA starts 30 days prior Jan/22/2020

Figure 2: SEIR model Fit

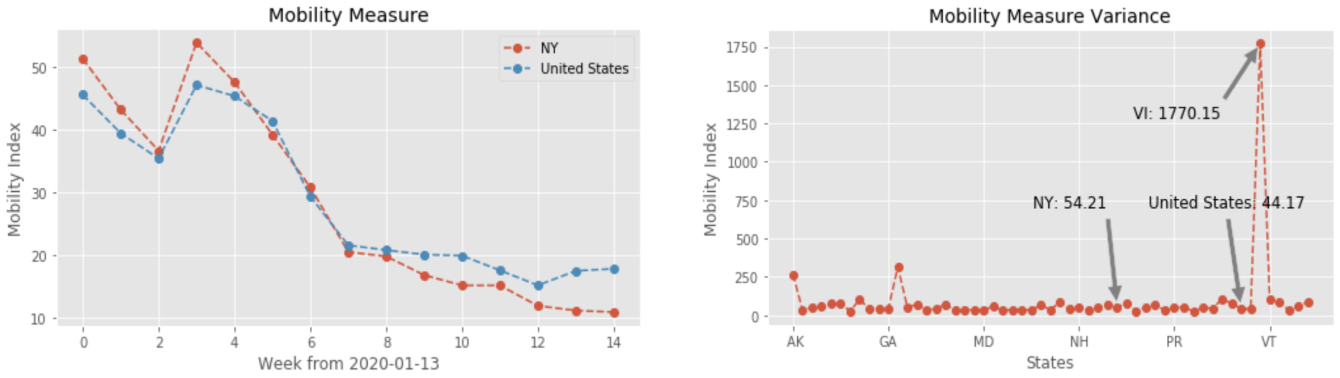
3 Mobility Index and Its Relationship to R_0

3.1 Mobility Index

For each twitter user, we collect all locations in one week period and compute the center of the locations as their home in that week. Then we compute their home location for next week and calculate the distance between this week location and next week location as travelling distance. Accordingly, user with larger travelling distance will have a larger social mobility index.[**2**]

Due to some missing data, we collect locations only for three days: Monday, Wednesday, and Saturday for one week period. And we take data from 2020-01-20 to 2020-05-14. We take the mean of mobility index for all twitter users who reside in one state (i.e. New York) in each week. We also calculate the variance of mobility index. Furthermore, we aggregate all measures and compute the mean and variance of mobility index for all users in United States.

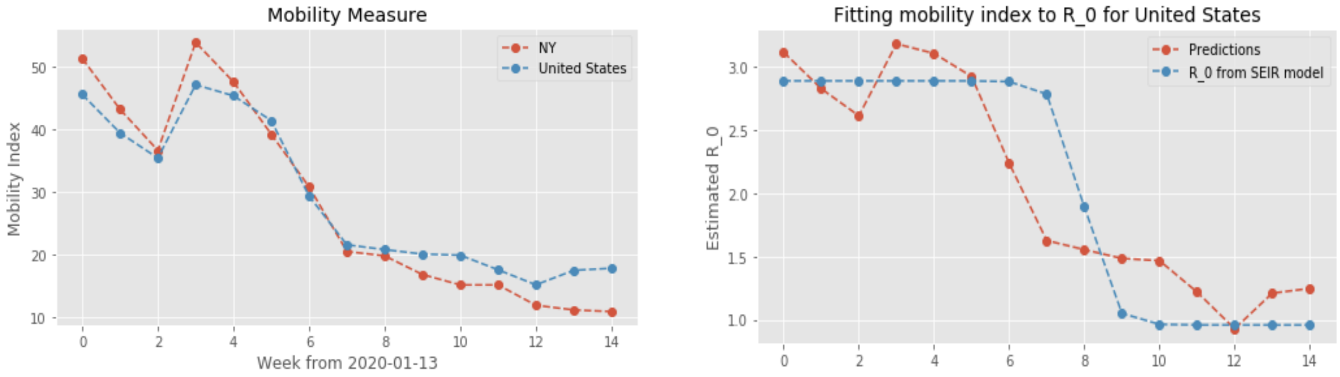
The mobility index of New York and United States is shown in Fig.3(a). The countrywide mobility variance is shown in Fig.3(b). The overall trend for New York and United States are similar, with a steep decrease at first and later a flatten period with really low mobility index. Due to the policy of partially re-opening some of the states, the country wide mobility index picks up a bit from the beginning of May.



(a) Mobility Index for NY and United States from 2020-01-13 to 2020-05-14

(b) Mobility Measure Variance for all states from 2020-01-13 to 2020-05-14

Figure 3: Mobility Index and it's Variance



(a) Fitting mobility index to R_0 for NY

(b) Fitting mobility index to R_0 for United States

Figure 4: Fitting mobility index to R_0

3.2 Relationship between Mobility Index and R$_0$

We fit the mobility index (dependent variable) to R_0 (response variable) using log linear regression [3] as approximation for New York and Country Wide respectively; and we compare this approximated R_0 with the R_0 obtained from SEIR model. From Fig.4(a), the predicted R_0 from mobility index for New York is pretty consistent with our estimation from SEIR model; From Fig.4(b), the fitted country wide R_0 from mobility index is within reasonable range from estimation from SEIR model with some minor fluctuations.

From the above comparison, we could clearly observe direct relationship between the mobility index and reproduction number R_0. With the enforcement of the social distancing, we successfully control and reduce R_0 through time and by our model indication, R_0 are in the hope to fall below 1 in the up coming month, which indicates the turning point of our growth curve after this almost 5 month's long battle against COVID-19 .

4 Tweets Topic Trend Analysis

4.1 Data for Tweets Topic Trend Analysis

We use two public resources of aggregated tweets data as our source in this section: COVID-19-TweetIDs and covid19-twitter With daily tweets ID filtered with COVID-19 related keywords.

4.2 Findings

We first investigate the number of tweets related with COVID-19, shown in Fig.6. We consider all tweets without filtering out the retweets. We can see two spikes at the beginning of Feb and March, matching the timeline of the first confirmed case in US and the sharp increase in cases in US. We are surprised to find that both the number decrease quite fast to a constant value.

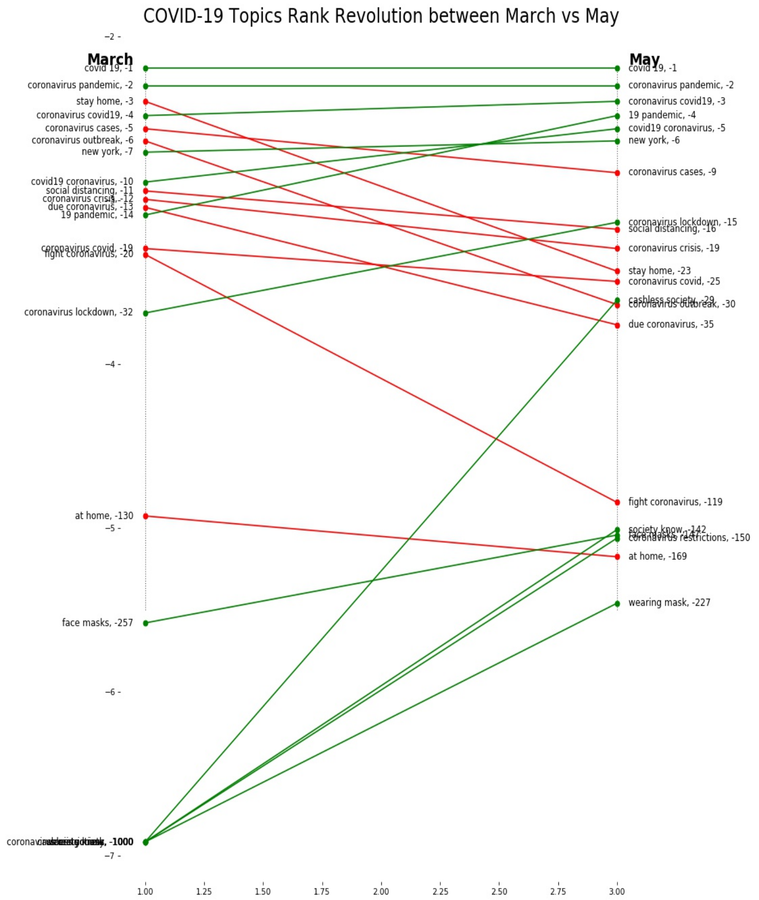


Figure 5: Most Frequent "bi-gram" rank change of tweets from 2020 March to 2020 May, red means rank decrease, green means rank increases. Left down corner -1000 means not in the top 1000 frequent bi-gram in March

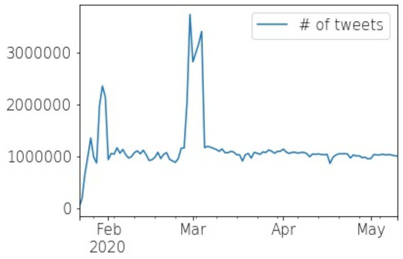


Figure 6: Number of COVID-19 related tweets from 2020-01-22 to 2020-05-10

This leads us to further investigate the change of topics on social media from March to May. We compare the rank of the most frequent bi-gram in tweets between March and May , shown in the slope chart Fig.5. We can see that social distancing related words like "stay home", "social distancing" and words like "fight coronavirus", "coronavirus outbreak", "coronavirus crisis" starting to lose people's attention in May. However, some new words start to pop up, phrases related to re-open and going out: "wearing mask", "coronavirus restrictions"; and words related to new trends in the society after pandemic: "cashless society", "society know" etc. Although being a hot topic, the contents related to COVID-19 have clearly changed from anxiety about staying home and fighting the virus to getting out safely and starting post-pandemic life.

5 Discussion

Our analysis shows that the R_0 is the most crucial variable driving the growth curve and it is closely related with the mobility index. As the social distancing policy being in place across the country starting March, the sharp decrease in the mobility drives down the reproduction number and allow us to approach the turning point of the curve faster. This indicates the stay home, stay healthy order was timely and necessary. Also the tweets analysis clearly demonstrates an information shift in the topic discussion from virus related ones to topics more related to re-opening.

It will be more rewarding if we could further conduct scenario simulations including vaccinations and possible recurrence of the covid-19 in the winter of 2020.

References

[1] King, Aaron A. “Introduction to Inference: Parameter Estimation.” Accessed May 19, 2020. Introduction to inference: parameter estimation.

[2] Xu, Paiheng, Mark Dredze, and David A. Broniatowski. ”The twitter social mobility index: Measuring social distancing practices from geolocated tweets.” arXiv preprint arXiv:2004.02397 (2020).

[3] Thakkar, N., R. Burstein, H. Hu, P. Selvaraj, and D. Klein. ”Social distancing and mobility reductions have reduced COVID-19 transmission in King County, WA.” Institute for Disease Modeling (2020).

[4] En.wikipedia.org. 2020. Compartmental Models In Epidemiology. [online] Available at: ¡https://en.wikipedia.org/wiki/Compartmental$_{m}odels_{i}n_{e}pidemiology > [Accessed 20 May 2020]$.