Super Spreader: How Political Rallies Spread Disease During the Covid-19 Pandemic

Bickston Laenger[*]
bickston@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Yanni Ma[*]
yanni.ma@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

KEYWORDS

Epidemiology, COVID-19, Super Spreader Events, Disease Modeling, Political Rallies, SARIMAX

ACM Reference Format:
Bickston Laenger and Yanni Ma. 2024. Super Spreader: How Political Rallies Spread Disease During the Covid-19 Pandemic. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (EPI '24).* ACM, New York, NY, USA, 8 pages. https://doi.org/N/A

1 INTRODUCTION

A superspreader event (SSE) is an incident in which a single infected individual transmits a contagious disease (in this case COVID-19) to a much larger number of people than the average transmission rate (R_0). There are many different factors like lots of close contact, larger than usual crowds, or prolonged exposure among people that can make an event a SSE. If several of these options are above normal levels, that allows the pathogen to spread more efficiently than in more common transmission scenarios.

In the context of COVID-19, superspreader events were often gatherings like weddings, religious services, or conferences, where it was possible for one infected person to cause an outbreak that affected dozens or even hundreds of others after the event. These events could then be dominoes that started a chain reaction of disease spread which culminated in exponentially greater transmission. Especially towards the beginning of the spread of the disease, when the whole world was trying to find out more about it, many purposefully chose to go against WHO guidelines, leading to the disease spreading and mutating quicker and generally being much worse for all involved.

Superspreader events came into the public view largely through news coverage of several high profile celebrities flaunting COVID restrictions at the time. Notable examples include the White House Rose Garden ceremony which was held to celebrate the appointment of Amy Coney Barrett to the United States Supreme court. This was an event that ended up exposing many high ranking US Government officials including the president. Another notable SSE was the UEFA Champions League Match where over 40,000 fans were in attendance, an event later traced back to a large spike in infections in both Italy and Spain. There were many other superspreader events near the beginning of COVID that also led to a large increase in the total amount of cases and spreading of the disease.

2020, the year when the disease and the pandemic first came to large public awareness, also happened to be an election year in the US. The office of the most powerful person in the world was under contention, and many rallies were being held on both sides to encourage voters to get out and vote. At the same time, there were many protests for the Black Lives Matter (BLM) movement, searching for justice after the death of George Floyd at the hands of Minneapolis police. While the guidelines from the Center for Disease Control (CDC) and World Health Organization (WHO) encouraged staying inside and keeping social distance, political factors instead encouraged people to break those restrictions and in turn make the spread of the disease worse.

In this project, we will take a look at how SSEs of the political variety ended up spreading COVID-19. We plan to obtain data about the size and location of rallies and analyze their connection to spikes in cases. We hope to finish off with a conclusive report on which political rallies were overall more infectious as a whole. We also plan to show the interplay between the dynamics of the pandemic and politics. Finally, we wish to demonstrate the infections and deaths that could have been avoided if better precautions were put into place, assuming that they were followed.

2 RESPONSE TO MILESTONE COMMENTS

(1) Using the new public dataset makes sense
Yes, we decided that it would give us the most accurate results and we believe that we found satisfactory ones because of it.

(2) Data collection setup and initial analysis looks good
We agree, we think it gave us a solid baseline into seeing what the future data insights could hold.

(3) Along with proposed methods maybe also look at more flexible models like Gradient boosting (XGBoost), SVM, Random forests that also provide good interpretability.
While we ran out of time for this project, that was certainly high on the list of things to do in the future steps category. We wanted to put all of our effort into ARIMA and SARIMAX so we could focus in on making those models the best they could be, rather than spread out our modeling on too many different ones.

(4) ARIMA implementation in statsmodel uses numpy backend written in optimized C/C++ so you do not need to do a C++ implementation from scratch. Similarly many scikit-learn

[*] Both authors contributed equally to this proposal.

EPI '24, Fall, 2024, Atlanta, GA

ACM ISBN N/A
https://doi.org/N/A

models are efficient enough to not warrant a rewrite in C++. So I would suggest to not spend time on this aspect.
We did end up using Python for everything, thank you for the recommendation.

(5) Proposed analysis form next steps looks good
Thank you, we believe we executed well on them to give us this final product.

(6) You need to have a control dataset to compare superspreader effect with a baseline; Also study temporal differences between normal and event time.
We were able to incorporate this with the 3 levels of valence. One level for Republican leaning events, one level for Democratic leaning events, and one level for a control. You can see the differences in the 3 at the end. We did also study event happenings and in the end we found no large effects of them happening on the rise in cases.

3 PROBLEM DEFINITION

We plan to model the spread of an infectious disease (in this case Covid-19) at different political rallies and controls to see if there are any characteristics unique to political rallies compared to any other Super Spreader Event. We will do this using ARIMA and SEIR modeling and plan to be able to find the difference, if there is any, between a political event and any other given SSE. Our initial hypothesis is that Republican events will be more likely to spread the disease because many Republican lawmakers encouraged taking fewer precautions to mitigate disease spread during the pandemic and as such should spread the disease more. We also predict that there should be some correlation between any given event happening and a large spike happening right after.

4 LITERATURE SURVEY

SSEs have contributed to the progression of various pandemics throughout the ages. Ones such as influenza and even the black death, not just COVID-19. Due to the unique timing and scale of COVID, some research has been conducted on the effect of SSEs on the transmission of COVID-19. However, existing research on SSEs has large amounts of both strengths and weaknesses, in part due to the ambiguity present on what exactly a SSE is. Understanding the existing literature on SSEs is incredibly important for preventing future pandemics from having as large of an impact as COVID did.

Pressman and Choi-Fitzpatrick's study on the impact of COVID-19 on U.S. protest patterns highlights that the pandemic did not significantly alter the core of protest. Acts such as street marches, chanting, and speeches remained the primary way that the public protested. However, the pandemic did lead to some notable protest-related changes. First, the subjects of protests shifted, with an increased emphasis on public health and economic policies. Some protestors even adapted by incorporating social distancing measures or choosing formats like car caravans to minimize close contact. Curiously, protests were often held near medical facilities during this period. The researchers note that while there was an increase in the use of online tools for organizing, it wasn't much different than pre-pandemic digital strategies [6].

Similar to how we are planning on using the Crowd Counting Consortium (CCC) to analyze protests, Pressman and Choi-Fitzpatrick based their study on that repository. Their work differed from ours since they only analyzed how COVID-19 changed protest behaviors rather than looking at how protests and other similar SSEs impacted case transmissions.

4.1 Strengths

Many papers have contributed useful insights into the characteristics and implications of SSEs. For example, Kumar et al. (2020) reveals that understanding the role of SSEs in COVID-19 can improve epidemic control strategies [2]. They document how super-spreaders, including a bartender at an Austrian ski resort, contributed to outbreaks across Europe, and they bring up the importance of early identification of super-spreaders. Additionally, Stein (2011) discusses how SSEs have been observed in various infectious diseases, such as measles and tuberculosis [7]. This study provides a broad understanding of how SSEs occur and their role in accelerating disease transmission, but there are opportunities to expand upon the work, and investigate SSEs in more specificity.

Teicher (2023) provides a historical review, offering insight into how the term "super-spreader" has evolved. They explain that the term was initially related to gastrointestinal diseases but later expanded to airborne diseases like tuberculosis, which is closer to the way that most of us think of SSEs now [8]. Majra et al. (2021) does a well-supported discussion of the relevance of SSEs in the spread of SARS-CoV-2. They note that large-scale gatherings, such as political rallies, played a significant role virus transmission [5].

4.2 Weaknesses

Our literature review also shows that there are weaknesses in current research. One issue brought up by Kumar et al. (2020) is that there are not many consistent definitions for what qualifies as a superspreader event [2]. Without a clear, universally accepted definition, the impact of SSEs on outbreak control measures remains unclear. We will be sure to strictly define what a superspreader event is in our work so there are no uncertainties in our final evaluation. Another limitation comes from the focus on retrospective analysis. Lots of studies, such as those by Stein (2011), identify SSEs only after they have occurred, making it difficult to proactively prevent such events [7].

Kyriakopoulos et al. talk about the important role of managing SSEs for controlling the spread of COVID. They show that early diagnosis of pre/asymptomatic individuals and efficient monitoring are essential to prevent outbreaks. They argue that epidemiological models that exclude SSEs produce uncertain and often misleading results, which is why it is important to integrate these events into disease prediction models to better understand pandemics [3]. The Kyriakopoulos paper was solely a literature review, without any technical implementation, so we aim to innovate further by incorporating SSEs into modeling.

In addition to this, the existing models like those discussed by Luhar et al. (2022) often improperly account for the unique social and behavioral factors at play during political rallies or protests [4]. While mathematical models like the SEIR framework can predict disease spread, they do not always capture the nuances of real-world gatherings, where non-compliance with safety measures (such as mask-wearing) can make transmission rates much much worse. The current literature generally tends to provide a high-level overview of what SSEs can do to transmission rates and disease spread without analyzing the specific details of how different types of events can lead to different disease outcomes. For example, research rarely looks into whether events like weddings or parties have a different impact than political ones, showing the opportunity to differentiate SSEs further for further research.

5 PROPOSED METHOD

5.1 Intuition

Our research addresses a significant gap in understanding the impact of political SSEs on the spread of COVID-19. By combining robust datasets and models such as ARIMA and SEIR, along with incorporating political valence and events as an exogenous variable, we aim to provide an insightful and novel analysis in this underexplored area.

This work offers both a novel perspective on political SSEs and a foundation for future research to explore correlations between political events and disease spread. There are many intriguing options for extending the research done here. The insights gained can help determine whether various political groups and events contributed to differing levels of transmission, offering a starting point for further studies in this field.

5.2 Approach

We plan to take our datesets and plug them into the ARIMA and SEIR models to give us accurate results on how Covid-19 would have most likely spread in those scenarios.

We are looking to see if there are any gross outliers either in the difference between political groups or between the groups and the control group. Using 2 models should give us a better chance at finding some intuition into this problem, or at the very least be able to point us along the right track.

After we run the models on the data we have collected, we plan to compare the results and graph them out to see if there are any noticeable trends in them. We will make many different kinds of graphs but one in particular should be more telling than most. That is the graph comparing all of the rallies together with the control, so it will have 3 lines total, one for Republican events, one for Democratic events, and one for control events. This will be able to give us insights into not only whether or not the political events have greater (or at the very least different) spread patterns than the control group but also the differences between the groups as well. Our initial hypothesis was that the Republican events would have the highest rate of spread, then the Democratic events, then the control.

Once we have the graphs from both the models we should be able to find patterns and trends and report on them from there.

5.3 Algorithms and Models

5.3.1 ARIMA Model.

Used commonly for time-series forecasting and analyzing the trends and seasonality in data, the Autoregressive Integrated Moving Average (ARIMA) model combines three components that can help us isolate the effects of superspreader events on case spikes:

- **Autoregressive (AR)**: This component captures the relationship between an observation and a number of lagged observations (previous time points). The AR term can be represented as:

  $$AR(p) : Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \epsilon_t$$

  where Y_t is the value at time t, ϕ_i are the parameters to be estimated, and ϵ_t is white noise.
- **Integrated (I)**: This component involves differencing the time series to make it stationary. The order of differencing is denoted by d.
- **Moving Average (MA)**: This part models the relationship between an observation and residual error from a moving average model applied to lagged observations:

  $$MA(q) : Y_t = \theta_0 + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}$$

  where θ_i are the parameters of the model.

The overall ARIMA model is denoted as ARIMA(p, d, q), where p is the number of lag observations, d is the degree of differencing, and q is the size of the moving average window. The general ARIMA equation can be represented as:

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \theta_0 + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t$$

5.3.2 SEIR Model.

The Susceptible-Exposed-Infectious-Recovered (SEIR) model is a model widely used in epidemiology to simulate the spread of infectious diseases. It divides the population into four compartments:

- **Susceptible (S)**: Individuals who are not infected but are at risk of infection.
- **Exposed (E)**: Individuals who have been infected but are not yet infectious (latent period).
- **Infectious (I)**: Individuals who are infected and capable of spreading the disease.
- **Recovered (R)**: Individuals who have recovered from the disease and are assumed to be immune.

The transitions between these compartments are described by the following set of ordinary differential equations (ODEs):

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

where:

- N is the total population, $N = S + E + I + R$.
- β is the transmission rate.
- σ is the rate of progression from exposed to infectious.
- γ is the recovery rate.

Both of these approaches together should help us to get a more in depth look at how the political events shaped the rise of the disease and while imperfect they should be a fairly close representation.

6 EXPERIMENTS

6.1 Questions to be answered

- Do large events happening affect the subsequent spikes in Covid cases?
- What, if any, differences are there between Political Rallies as SSEs compared to any other given SSE?
- What, if any, differences are there in the spread of disease between the different kinds of political rallies of different parties?

6.2 Experiment Details

6.2.1 Data Collection. As we finished up work on this project, we have identified the data sources that align best with our goals. These data sources provide the needed granularity for us to draw accurate conclusions about political SSEs and their specific impacts.

Our first data source is the Crowd Counting Consortium (CCC) data repository, a collection of metrics hosted on GitHub by the Nonviolent Action Lab [1]. This repository collects data on various political protest events in the USA, defining protest broadly. Various csv files from the CCC include information about protests, demonstrations, rallies, and more. The full dataset is split into two files, one representing the period 2017 - 2020, which is the one we will be focusing on, and one from 2021 - present. Available on GitHub, the csv files include entries for number of participants, event type, and geographic location. The authors of this data source recognize that no large-scale dataset can ever fully capture all relevant events with complete accuracy; however, they mitigate this risk by implementing a robust review process that takes in suggestions and error corrections through their website, so we think this source is reputable enough to use.

Here are some of the notable columns found in the CCC dataset:

- **date**: event date in YYYY-MM-DD format
- **online**: indicates online-only events. 1 for yes, 0 for no.
- **type**: type of event (e.g., march, protest, sit-in)
- **title**: name of the event, if available.
- **valence**: political leaning of the event (2 for right-wing/pro-Trump, 1 for left-wing/anti-Trump, 0 for neutral).
- **issues**: tags representing political issues related to the event (e.g., democracy, women's rights).
- **size_low, size_high, size_mean, size_cat**: various metrics related to the size of the event, such as the lowest, highest, and average estimated count of participants, and a categorical size indicator.
- **fips_code**: unique five-digit code identifying each county and state, used for easy matching and integration with other datasets

The second data source, accessible from the NY Times, is a comprehensive dataset for COVID-19 case and mortality data for the US [9]. For this dataset, we are planning on analyzing changes in case numbers, investigating how local infection rates evolved in response to specific SSEs.

The dataset provides cumulative counts of COVID-19 cases and deaths at the national, state, and county levels. The repository also provides a live dataset and a historical one, with the former representing only the current day, a snapshot of reported cases and deaths on that day, with updates throughout the day. Once that day is over, then the live data for that day is finalized and archived in the historical dataset. On the day they are first announced, that is when the cases and deaths are counted. Since the dataset stopped being updated on March 24, 2023, the distinction between live and historical data is not relevant anymore. We plan to use the historical county-level dataset for 2020 primarily, with the possibility of using 2021 data as well.

The goal of this dataset is to compile comprehensive time series data, offering a detailed record of the pandemic's progression. Data is sourced from state and local governments as well as health departments across the country. However, a notable limitation of this dataset is the occasional shortage of COVID-19 testing throughout the pandemic, which may restrict how fully the data reflects the outbreak's actual spread.

Here are the columns of the county-level dataset:

- **Date**: date of reporting for each entry
- **County and State**: geographic information
- **FIPS Code**: same as FIPS Code above
- **Cases and Deaths**: total cumulative number of reported cases and deaths as of the date of reporting

The dataset was created by a team of NYT journalists who monitored government updates, data releases, and press conferences, ensuring accuracy by corresponding and following up with public officials. Due to inconsistencies in the U.S. public health reporting system, they updated entries as new information emerged, adjusting for corrections in state or county records. The team assigned cases to the patient's treatment location rather than their residence, but not all state governments did the same. The dataset includes both "confirmed" cases (validated by COVID tests) and "probable" cases (based on symptom assessments without testing), but some governments didn't do the same, or made it so there was no way of separating the confirmed from the probable. They aimed for the most accuracy possible, but discrepancies may still be present due to the difficulty of patch-working together data from 50+ state and territorial governments, since they all release and categorize data in different ways. The NYT team emphasized that the American public health system was overwhelmed, and thus had difficulties reporting information with accuracy, timeliness, and consistency.

Using these two datasets, we will apply both the ARIMA and SEIR models to analyze the impact of various political gatherings on COVID-19 spread. The Crowd Counting Consortium data will allow us to estimate the size and location of events, while the NY Times dataset will provide the necessary COVID-19 case trends for those locations.

6.2.2 Experiment Description. We began by cleaning and preparing our data. To start, we began with essential tasks such as removing duplicates and filtering out anomalies, but were quickly able to move to the data preparation part. For most of our more complex modeling, we combined the datasets on their FIPS codes and then sorted by date to create a complete look at the events happening in each county and the associated case counts. We also had to modify the case data slightly to fill in days with zero case counts, interpolating cases to limit the number of days with zero cases recorded.

Having zero cases in a day in any given county is extremely unlikely, so these measures were taken to ensure better results when modeling.
Below is a Python code snippet used to adjust the data for the cases:

```
county_data_2021['adjusted_cases'] =
    county_data_2021['new_cases'].replace(0, pd.NA
    ).interpolate(method='linear').fillna(0)
```

This line processes the daily new cases data to handle missing or zero values. It replaces zeros by applying linear interpolation to estimate missing values based on neighboring data points, and fills any remaining gaps with zeros to ensure a complete dataset for analysis. If there were several events in one day, for ease of adding the exogenous variable, we kept the one with the largest attendance. In the future, other strategies should be explored for handling several events in one day.

We were originally planning on including data about the Black Lives Matter movement, but in almost all cases it ended up being nearly identical to the Democratic outcomes, so we ended up combining them together in our final results. If we split BLM events and other left-leaning events, then the dataset for all events wasn't as complete.

We then started the experimental part by implementing the SEIR model to get a simple model that could give us a quick look and see if there was an easy correlation between event count and number of cases. Here is one such graph:

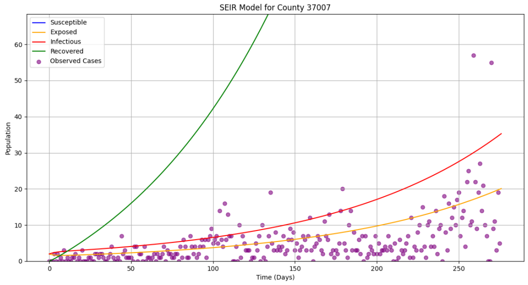

Figure 1: SEIR Initial plot

However, we quickly found out that there is no good correlation between number of events or even event size and how the cases progressed over time so we decided to move away from it. There was no sweet spot for β that we could find that was even within an error bar for most counties. This is most likely due to the fact that reported case data can be very different county by county due to differing methodologies for testing or reporting. Still, in this instance it led to a result that was less clear than we would like to be able to make conclusions from so we moved on.

Next, we tried to correlate events happening to a spike shortly thereafter. We did this by graphing when events happened in certain counties along with their daily cases and spikes measured by comparing the z score of a given value to a z threshold. If any given day was a spike, we marked it along with the days of events as shown in this graph:

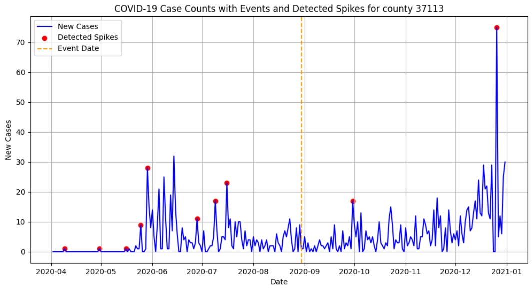

Figure 2: Spike graph with an event and spikes all around

Next, we implemented the mechanistical ARIMA model. Specifically, we implemented the SARIMAX model so it could both include seasonality (because Covid, like flu, gets much worse in the Winter) and the exogenous variable of what political leaning each event had. The models incorporate event-based data such as their political leanings and their sizes to study the potential impacts of the events on the overall case trends.

We started by using unsmoothed models but due to the data being very variable (again due to case reporting inconsistencies) we found that a smoothed model worked much better. For instance, here is an initial model run without smoothing:

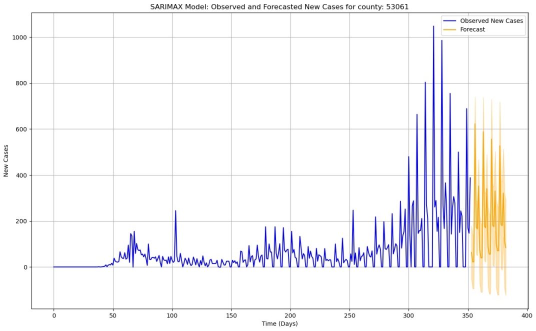

Figure 3: SARIMAX graph without smoothing

As you can see, the hypothesized case count has several confidence intervals getting predicted to be negative at some of their points, not something that could ever happen in real life. Because of this we changed to a smoothed model that fixed a lot of the problems we ran into as you can see in this much more realistic predicted graph:

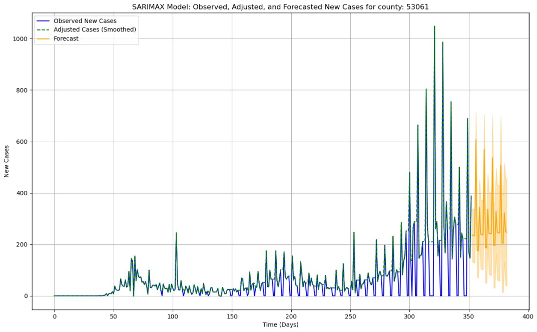

Figure 4: SARIMAX graph with smoothing

Once we had a working SARIMAX model we then focused on improving it. We worked on minimizing Root Mean Square Error (RMSE) for the model so it could also be generalized on more than just one county from our dataset. We implemented grid search to find the best hyperparameters to train the SARIMAX model on and once we had them trained our model once more to ensure it was the best it could be. We did this for 2 models, one with exogenous variables and one without. The only exogenous variable we used for this project was the valence variable, which showed the leaning of each different kind of event. The majority of events were neutral but there were enough both Republican and Democratic leaning events to give us a solid look into their differences.

Once we had all of our SARIMAX models, we then set them against each other to see which ones were best. Once we found the most optimal model, that was the one we used to then predict outcomes based on event valence. We ran it 2 separate times, once weighted for the number of people at each event and once weighted just for the total number of events for each county. This part limited the total number of counties that we could graph because there were over 1700 counties in our dataset and the majority of them only had zero or one recorded event in them. Even fewer yet had 1 event from each side to be able to compare. Still, this let us compare at the county level and not across counties which should ensure greater consistency in our results.

6.3 Observations

We had initially hypothesized that there should be some correlation between an event happening and a spike in cases right after the fact. One such graph to support that view would be this one right here:

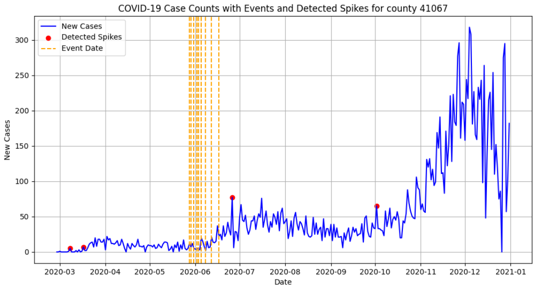

Figure 5: A spike graph with a correlation between events and spikes

The exact opposite of that hypothesis would be that perhaps events should happen right after spikes. There is even a graph that supports that presumption:

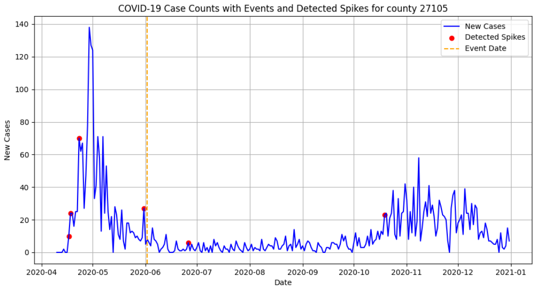

Figure 6: A spike graph without a correlation between events and spikes

However, in looking through several hundred of the graphs for each of the counties revealed that there is simply no correlation between events and spikes. Most graphs of this sort looked like Figure 2 above, where there was an event neither close to the front nor the back of any spike, instead it is almost completely random. While this did not support our initial hypothesis, it is not entirely surprising either. There are plenty of rational explanations for why events did not have any relation to spikes. One such explanation is that our dataset simply cannot contain every possible event nor perfectly get the political leanings of each one. Further, the line in the graph could be an event of 10 people or one of 1000, with each of them most likely having a very different outcome on being a SSE. Finally, in many of the counties they are simply too large for any given event to make a sizable dent in the daily case count more than any other given meeting of infected people with susceptible ones.

This now brings us to our main hypothesis: did Republicans spread Covid-19 more than Democrats did? We can look at a few graphs to tell. After we had tuned the graphs to be the best they could be we ran them on all the counties which had events for both Republicans and Democrats and compared them. We trained the SARIMAX model on the 365 days of data in 2020 then had it predict the next 30 days for each of the 3 valences of Republican leaning events, Democratic leaning events, and the control. You can see the graph of that here:

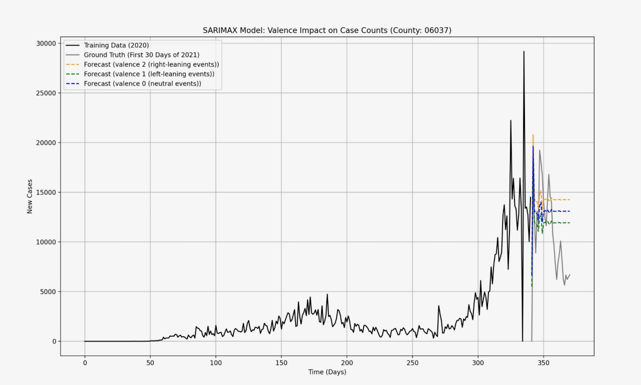

Figure 7: SARIMAX for LA County

It is plain to see in that graph that our hypothesis was proven correct, that because the main people at Republican political events, Republicans, were more likely to spread Covid. In essence, this graph shows that if there were only Republican events held for the first 30 days of 2021 then the predicted cases line would have ended up at the top predicted line. If only neutral events were held, it would predict the middle line, and if only Democratic events were held, it would predict the lowest line. This trend is shown even more so when you weigh for the number of people at each event:

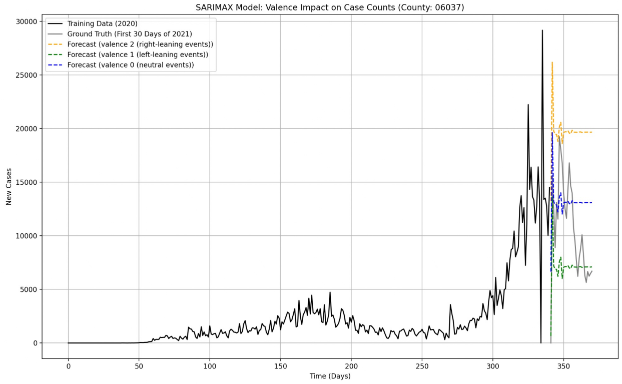

Figure 8: SARIMAX for LA County weighted by people at event

This further exacerbates the trends shown in the original graph. This also is not a case of data skewing because the average number of people at both Democratic and Republican events were about the same. This just shows that the model learned that the more Republican events that were held, the more one can assume that the daily cases will go up in the subsequent time. This trend was also repeated for other large counties that held both Republican and Democratic events, this was not just held to LA county.

6.4 Findings

Our initial hypothesis that Republican events and therefore Republicans themselves would be more likely to spread Covid-19 was shown to be true. It was shown over a large variety of counties in both red and blue states so we can at the very least say that the evidence is in our favor for that point. With that said, the part of the hypothesis where we predicted Democratic events to spread more than the control was also untrue. We had originally predicted that the Democratic events would most likely spread more than the control because the control events had smaller turnout in most cases and as such should not contribute to a rise in cases as much but that turned out not to be the case. Our main hypothesis is that Democratic events at this time pushed heavily for Covid safety measures such as social distancing and mask wearing, while control events would most likely be a more casual gathering where masks and social distancing may generally be less cared about.

Our second hypothesis of the events being large causes for spikes was also proven to be false simply due to the lack of any amount of correlation. This is also confusing because one would assume that if the events did not often lead to spikes in Covid then how did the Republican events differentiate themselves from the Democratic events in the SARIMAX model? Our main hypothesis is that the Republican event itself did not cause the spike, but instead used the compounding probability of disease spread over time to end up with the lead in the end. For instance, if a Republican only has a 1% greater chance of spreading Covid on any given day, over a year that results them being 37 times more likely to spread the disease than a Democrat. While the real data is not that stark, we do believe that the little percentages of unsafe measures such as avoiding social distancing and masking added up to lead to the differences we can see in the LA graph.

7 CONCLUSION

7.1 Limitations

The first and largest limitation is that of time. We were given 2 months and a 2 person team to build up several reports, do data collection, run the experiment, then make a final report, website, poster, and presentation. All of these things coupled with the other classes we were in meant we had to make the most of our time when we were working on this. With that said, we put the best amount of effort we could have into this and are satisfied with the end result. The condensed timeline also helped us to not waste too much time on items that were not necessary while still encouraging us to do the best work we could with the time provided. If we had more time we could also most likely implement several items in the Future Steps category as well.

Another limitation is that of model imperfection. We started this class with a very famous quote: "All models are wrong, but some are useful." This held especially true here when our entire project was built around taking results from models we created. This will inherently mean that our data cannot be trusted 100%, but instead should be taken as a data point and then combined with others calculated in different ways to see if a pattern emerges. Still, our individual data should also at the very least provide a baseline to be built off of as well. In the end, we did the best we could with the input we were given and made the models to the best of our abilities.

7.2 Future Work

Tying into the limitations, with more time we could go into a host of new ideas and directions in exploring what is and is not feasible or at the very least understandable from the data we took. We could expand both our data palates as well as what we ended up doing with it. We chose ARIMA and SEIR because they were proven models that have strong applications in Epidimiology. In the future we could try some different models with what we have right now such as a hybrid Machine Learning + SEIR to combine the epidemiological grounding of SEIR with a machine learning model like random forest to optimize parameter estimation and improve prediction. We could also get geographic data and implement spatial regression models to see how political events spread better or worse than other superspreader events such as weddings. Getting more specific into types of events would be a high priority on the list of things to continue with.

More than just trying different models or getting better data to put in them it would also be possible to continue this work by incorporating real time data into it. While Covid is nowhere near its fear levels or heights of the 2020-2021 season, it is still very rampant today. Even now it continues to have spikes when many people now almost fully dismiss it. If we were able to get reliable real time data and put it into the models to continually monitor and see how political rallies to this day are affecting its spread it would not only help stem the spread of the disease more but would also be further increase the amount of research in this area. As an election year, 2024 saw a multitude of political rallies on both sides, still most likely spreading Covid even if the focus on the disease is no longer what it once was. As such, being able to see the real time data on this might be a good way to not only expand the project but also spread more public awareness of how disease is constantly around us each and every day even if people don't give it nearly the attention it deserves.

7.3 Conclusion

The findings of this study show a straightforward but important point: those who did not follow expert recommendations in regards to the pandemic and public health likely contributed the most to increased case counts. Our results suggest that no single event was a major driver of significant spikes in case counts. It's likely that each event contributed incrementally, compounding the case counts over time. In the end, it is always important to treat infectious and deadly diseases with the caution they deserve, no matter your partisan leanings.

The process of analysis presented several challenges, particularly in working with the datasets. For instance, the valence variable had to be recoded to properly differentiate between the effects of different political events as an exogenous variable in the ARIMA model. The COVID-19 case data was incomplete, with dates showing zero case counts. To address this, we applied data cleaning and imputation strategies to interpolate missing values, aiming to produce more realistic projections. Another complexity was tuning the ARIMA model. The various underlying mechanisms of COVID made it challenging to select parameters that captured the full dynamics of the data, so model optimization took longer than expected.

Looking ahead, we aim to expand our analysis to include a broader set of counties, as time constraints limited us to a subset of the 1,736 counties available. We also plan to explore alternative models to validate whether our conclusions remain consistent across different methodologies. By refining both our approach and the data, we hope to gain deeper insights into the relationships between political events, public health behaviors, and disease spread.

8 EXPECTED WORK DISTRIBUTION

Bickston will take care of the data cleaning and the coding of the SEIR model.

Yanni will do the research in what has been done previously as well as sourcing the data.

They will both work equally in doing the project proposal, milestone report, final report, and final presentation. They will both work on insight generation from the data, evaluating the data from the models, and small tasks here and there to help ensure the project comes along smoothly.

REFERENCES

[1] Crowd Counting Consortium. 2024. Crowd Counting Consortium Dataset on Political Protest Events in the United States. https://github.com/nonviolent-action-lab/crowd-counting-consortium Includes data on protests, rallies, demonstrations, and similar events in the United States, with datasets covering 2017-2020 and 2021-present..

[2] Sanjiv Kumar, Shreya Jha, and Sanjay Kumar Rai. 2020. Significance of super spreader events in COVID-19. *Indian journal of public health* 64, 6 (2020), 139–141.

[3] Anthony M Kyriakopoulos, Apostolis Papaefthymiou, Nikolaos Georgilas, Michael Doulberis, and Jannis Kountouras. 2020. The potential role of super spread events in SARS-COV-2 pandemic; a narrative review. *Archives of Academic Emergency Medicine* 8, 1 (2020).

[4] Mitul Luhar, Assad A Oberai, Athanassios S Fokas, and Yannis C Yortsos. 2022. Accounting for super-spreader events and algebraic decay in SIR models. *Computer Methods in Applied Mechanics and Engineering* 401 (2022), 115286.

[5] Dasha Majra, Jayme Benson, Jennifer Pitts, and Justin Stebbing. 2021. SARS-CoV-2 (COVID-19) superspreader events. *Journal of Infection* 82, 1 (2021), 36–40.

[6] Jeremy Pressman and Austin Choi-Fitzpatrick. 2020. Covid19 and protest repertoires in the United States: an initial description of limited change. *Social Movement Studies* 20, 6 (2020), 766–773. https://doi.org/10.1080/14742837.2020.1860743

[7] Richard A Stein. 2011. Super-spreaders in infectious diseases. *International Journal of Infectious Diseases* 15, 8 (2011), e510–e513.

[8] Amir Teicher. 2023. Super-spreaders: a historical review. *The Lancet Infectious Diseases* 23, 10 (2023), e409–e417.

[9] The New York Times. 2021. Coronavirus (COVID-19) Data in the United States. https://github.com/nytimes/covid-19-data Data from The New York Times, based on reports from state and local health agencies. Includes cumulative case and death counts at various geographic levels across the U.S..

Received 8 October 2024