#COVID-19 Discourse on Twitter

Masha Belyi | Cameron Lopez | Joanne Jia

https://mashabelyi.github.io/covid-twitter/

Introduction

The Coronavirus pandemic hit the United States in early Spring 2020. As common in times of crisis, millions of Americans took to Twitter to share and discuss information about COVID-19. In this project we set out to explore how people reacted to this emerging, rapidly evolving situation in their tweets. Towards this goal, we analyzed tweets originating in the United States between March 1 and March 30 to understand how the conversation evolved over time. We consider the following broad research questions:

1. How much did people tweet about COVID-19 on a daily basis? Did the volume of tweetage vary from state to state? And if so, which states contributed to the conversation the most?

2. What was the emotional response to the pandemic, and how did that change over time?

3. What specific topics did Twitter users address in March, and how did that evolve as more information about Coronavirus was floated in the media.

4. What role, if any do social media bots play in the dissemination of information on twitter?

Related Work

Our project is well positioned within the rapidly expanding body of work that aims to characterize online discourse around Coronavirus towards deriving valuable public health insights (Alshaabi et al., 2020; Zhao&Xu, 2020; Ferrara et al, 2020; Shahsavari et al, 2020). The general approach to harnessing social media for pandemic response is not a novel endeavor. Previously, analysis of social media response to N1H1 (Chew&Eysenbach, 2010; Signorini et al 2011), Zika outbreak (Fu et al, 2016; Xinnng et al 2018), and Ebola (Spinney, 2019) successfully derived actionable insights about proper mitigation response for public health responders. In addition, due to the affordances of social media for the rapid spread of information, platforms like Twitter often inadvertently facilitate the spread of misinformation (Grinberg et al, 2019), often instigated via social bots (Ferrara 2018). Thus, analysis of social media has been a popular approach in recent years towards evaluating public response to global events.

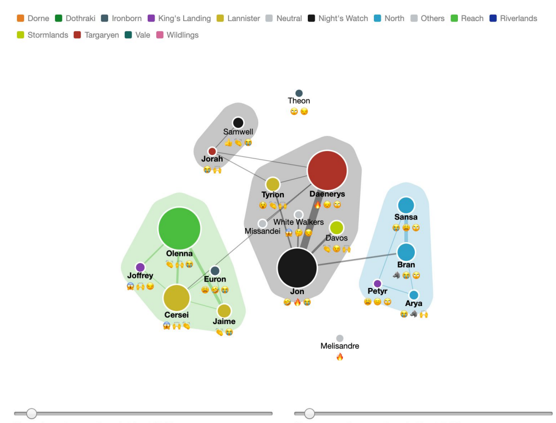
Below we elaborate on select studies that are most relevant to our project, as well as identify approaches to text visualization that served as inspiration for this project.

(Ferrara et al, 2020) closely evaluate the presence of social media bots in the #Covid19 conversation on Twitter. They compare top ngrams used by likely bot vs. non-bot accounts and present their results in a time series plot and word cloud format. The main outcome of the study was an exposure of right-learning polarization in bot-generated posts on Twitter. This work inspired us to differentiate between likely bots accounts and others in our analysis. It also made us vigilant about interpreting raw token and hashtag counts, as we realized that hashtag usage might be artificially inflated by automated accounts.

[Twitter for H1N1 Tracking](#) (Signorini et al 2011)
Sentiment analysis has been used in the past to track social media perspectives on other pandemics, including the H1N1 virus. This study uses Twitter sentiment information on a wide scale to measure public concern, and it shows that estimates of illness derived from Twitter discussion of H1N1 accurately mimics actual disease levels. From this study we realized we could utilize Twitter information as an accurate means of tracking certain aspects of the current COVID-19 pandemic, most importantly public sentiment.

[To Tweet or to Retweet?](#) (Lee&Sundar,2013)
The credibility of a tweet is perceived by basis of the expertise of the tweeter, while the credibility of retweets is perceived by basis of the trustworthiness of the retweeter. This paper was part of the inspiration for our decision to include retweets in our tweets per day visualization as well as differentiate between tweets and retweets for sentiment analysis.

[Force-Directed Graph of Game of Thrones](#). Krist Wongsuphasawat's work showed us how tweets could potentially reveal interesting patterns between subjects that were often mentioned together. Even though the algorithm knew nothing about the show, it was able to "make sense" of the relationships between them. Inspired by his force-directed graph in D3, we explored a similar graph that was aimed at showing relationships clearly.



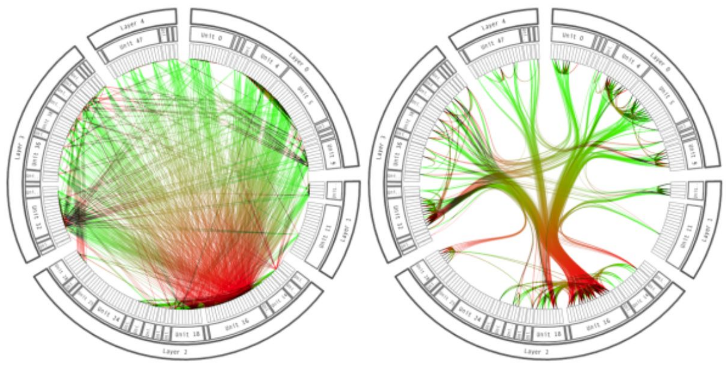
[Selecting Semantically-Resonant Colors for Data Visualization](#)
Studies show that using semantically-resonant colors improves participants' reading speed. The semantically-resonant colors are selected from representative images associated with the values. This motivated our selection of blue and yellow colors for our sentiment plot. We encode blue to encode negative sentiment, as in "feeling blue"; and we encode yellow to denote positive sentiment, as yellow is generally a happy (sunny!) color.

[Visualizing Conversations over Time.](#) We drew a lot of inspiration from the Twitter Interactive web page[1]. Two visualization of the #Oscars and #Election2016 conversations particularly stood out to us as very effective in visualizing the amount of Twitter "buzz" devoted to select subtopics with the main Twitter conversation. We liked this approach so much that we implemented our

[1] https://interactive.twitter.com/

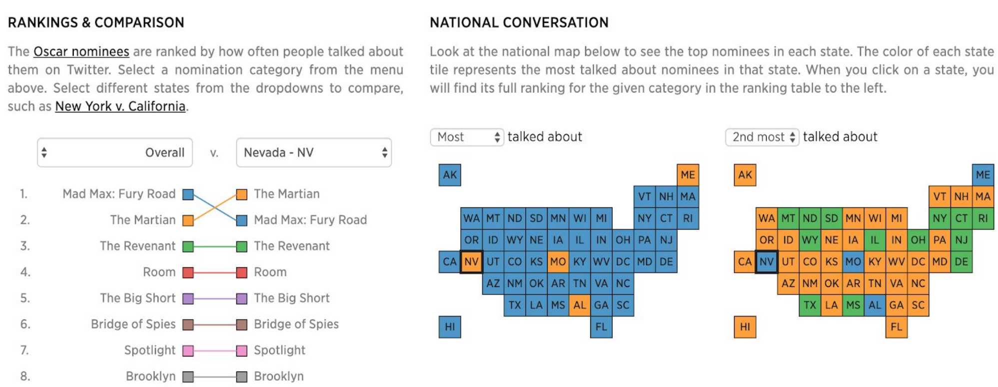
own version of the #Election2016 visualization design. We used the Twitter design as a template, but implemented the code ourselves in d3.

[Hierarchical Edge Bundling]



One difficulty in showing relationships through a network graph is it very quickly becomes too complex and unreadable. What an hierarchical edge bundling technique does is it groups the adjacent edges together to decrease the visual clutter. This point is clear through the comparison below. *Left*: Without Bundling, *Right*: With Bundling

[Connecting a Clickable Map to a Separate Visualization]



We were inspired by this work to include a clickable visualization of the United States that showed data from the clicked state in a separate chart. We originally designed a grid-style map as featured here, but after usability testing decided to convert to a geographic map.

Data

We use the COVID-19-TweetIDs dataset (Chen et al, 2020). The full dataset comprises tweets fetched via Twitter's streaming API from January 21, 2020 to present day. At the date of this study, the dataset contains nearly 100 million tweets. The authors track specific keywords, hashtags, and influential accounts that were hand-selected by the authors to include all covid19-related conversations on Twitter. Sample keywords include *coronavirus*, *covid19*, *pandmeic*, *#quarantinelife*. Select accounts include WHO, CDCgov, CoronaVirusInfo. The full list of tracked keywords and accounts is available here.

We subset the dataset to include only English tweets and retweets posted in the United States in March of 2020. The filtered subset comprises 1,022,822 tweets and 5,959,502 retweets. We process each tweet to infer additional attributes. Refer to our detailed Methods section for details on how this filtering was performed. We further process the data to extract the overall sentiment for each tweet. Additionally, generate Botometer[2] likelihood scores for select user accounts in the dataset in order to evaluate the potential influence of social media bots on the #COVID19 conversation. Finally, we employ a community clustering algorithm to identify clusters of conversations around covid19 and track the volume of Twitter traffic devoted to each conversation over time. We describe each data processing step in detail in the Methods section of this report.

To support state-by-state comparison, we sourced official government state population projections from the United States Census Bureau.

[2] https://botometer.iuni.iu.edu/#!/

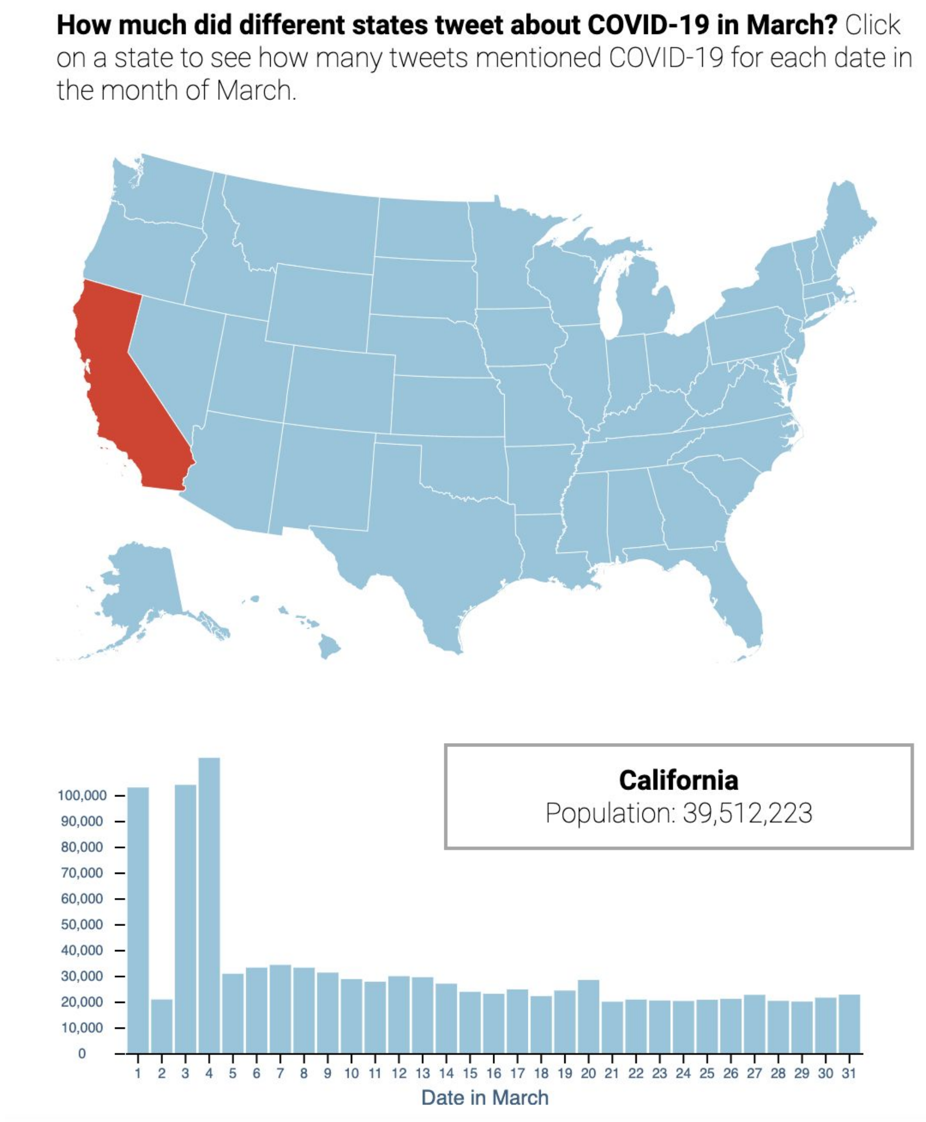
Visualization

We visualize the data in an interactive online dashboard, implemented in HTML and D3. We adopt a one-page layout where we sequentially walk the user through each part of the story. We make use of careful choice of color, highlighting, and font typesetting in combination with pointed messaging to the user to the user to guide the eye towards the key points we hope they take away from the visual experience.

Tweets by date

We begin with an interactive visualization of the number of tweets relating to COVID-19 sent in each state by date for the entire month of March. The visualization presents the user with a light blue map of the United States with state boundaries outlined in white. A red California is preselected and a bar graph below shows the number of tweets mentioning COVID-19 per day for the selected state. We preselected California rather than leave the map blank to hint to the user that they can interact with the individual states. Upon hovering over another state, the state turns red and the bar graph changes with an animation. The bars of the graph below grow from left to right to both hint at the passage of time from the start to the end of the month and to draw the user's eye to inform them that their mouse movement was connected to the change in graph.

To further bridge the connection between map and bar graph, we chose to outline the name of the state in a box above the bar graph. This serves to draw the user's eye and establish a clear visual connection between the bar graph and the state it represents, drawing upon the gestalt principle of closure. Additionally, we chose to show the population of the state in question rather than normalizing our data in order to more accurately represent where most covid-related tweets were coming from while subtly prompting the user to recognize that states with higher populations will likely see more tweets. We incorporated the bar graph's y-axis label (number of tweets) into the title rather than leave it on the side of the graph to make it extremely obvious what was being shown to the user. The x-axis of the bar graph represents the date in March.

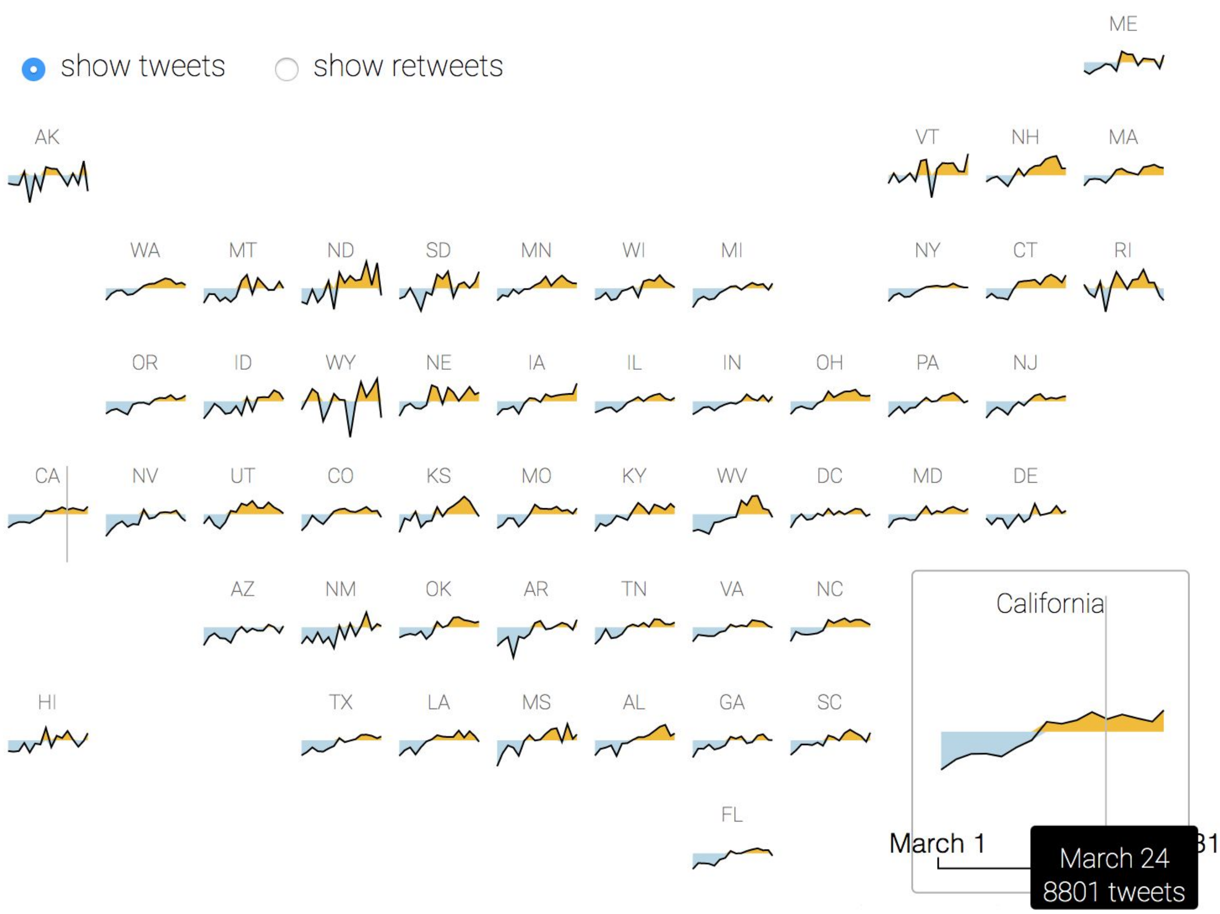


To create this visualization, we utilized tweet location data as well as state population data. Details can be found under the *Data* section.

Sentiment

Next, we introduce a small multiples plot of average sentiment on Twitter every day, split by different states. Refer to our Methods section for details on sentiment calculations.

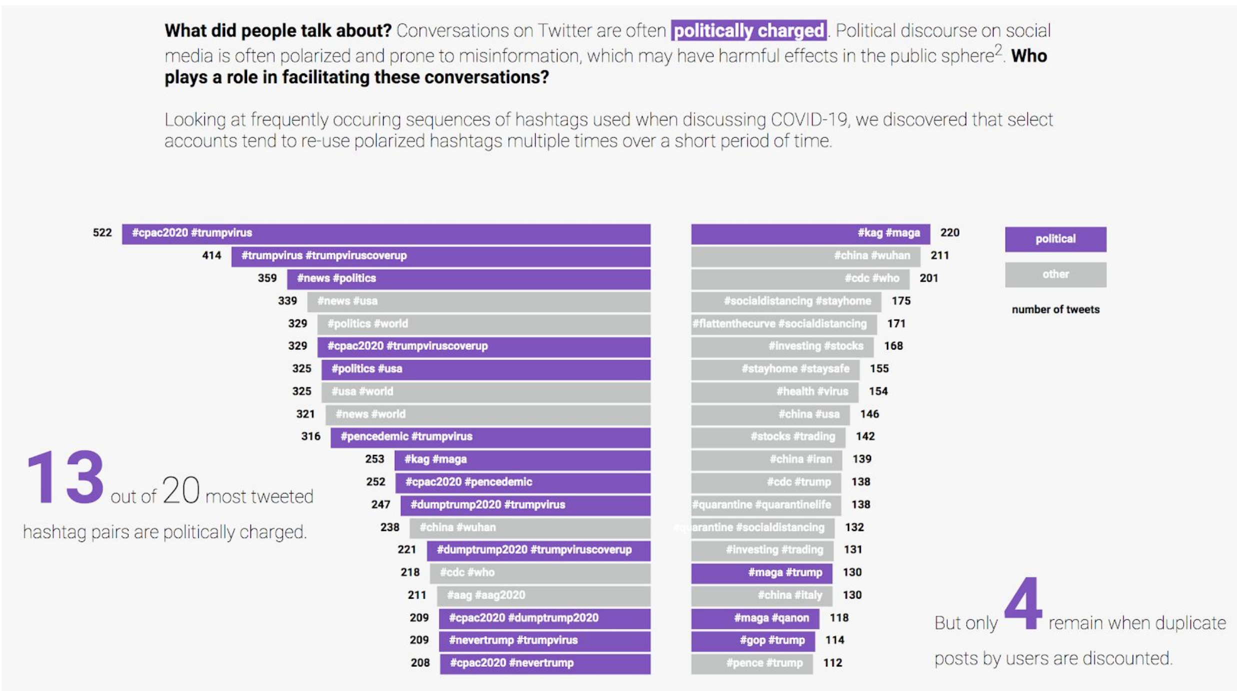
We chose the small multiples layout because it allows the user to quickly compare sentiment trends in different states. We use a familiar USA grid layout, which enables the user to quickly identify states of interest, the efficacy of which was confirmed during useability testing. We build interactivity into the visualization with a "detail" view in the lower right that updates when the user moves the mouse across the grid. In addition to showing a zoomed-in view of each state in the closeup, we also display the date and the number of posts tweeted that day for additional context. Finally, we prompt the user to switch between viewing trends calculated from *tweets* vs. *retweets* and observe an interesting change in the trends.



Hashtag Clusters

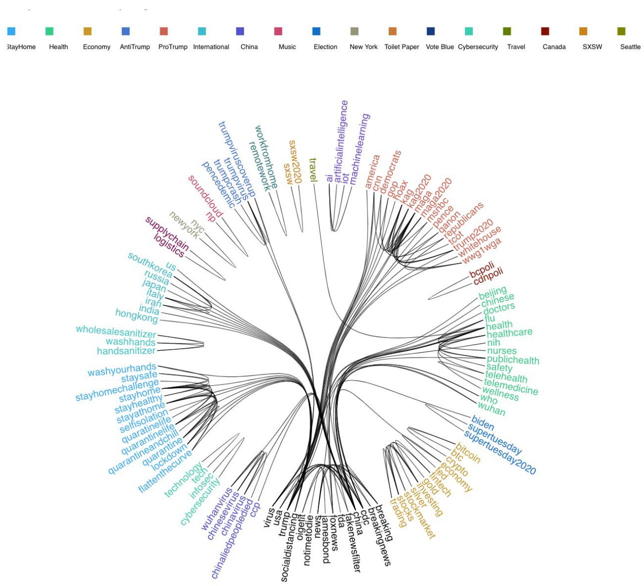
To understand the substance of the online conversations in more detail, we examined the top most used pairs of hashtags in the dataset. We visualized usage statistics of the hashtag pairs in a diverging bar chart layout. We prompt the user to observe the apparents political polarization in the top hashtags when ranked by raw usage counts vs. by counts of unique users that used the pair at least once.

Purple color highlighting is used to identify the politically charged hashtags and point them out to the user. We settled on purple since it is the mix of red and blue, two colors that are often used to encode political affiliation in the United States.

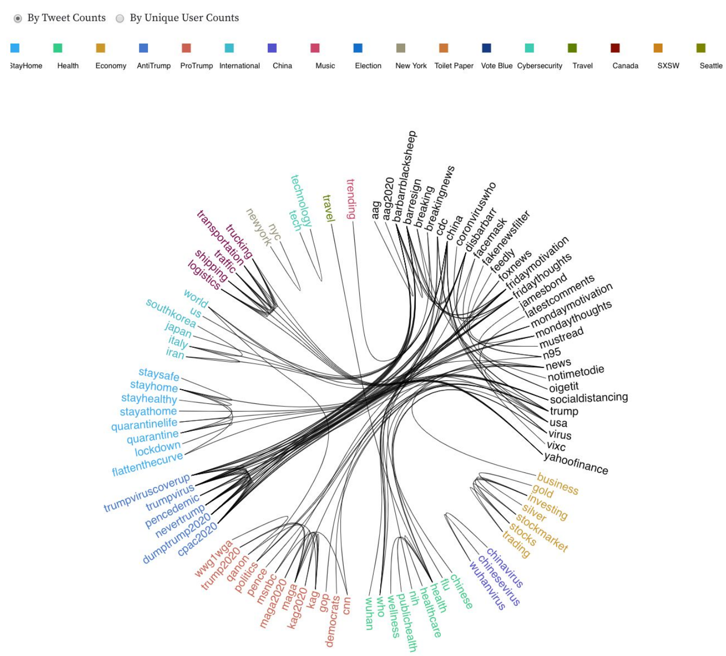


To begin to explore the topics that appeared through the use of hashtags, we display a hierarchical edge bundling graph that highlights the relationships between pairs of hashtags. Similar to the previous bar chart, users can toggle between two views: top hashtag pairs *by raw count* or top hashtag pairs *by user count*. The existence of an edge between two hashtags meant there was at least one user that used both hashtags in the same tweet (see Methods for detailed approach).

We used colors and proximity to encode the topic group to which a hashtag belonged. Topic names were manually assigned to groups of densely connected hashtags. When users hovered over a hashtag in the graph, that hashtag plus any connected hashtags scaled in size.

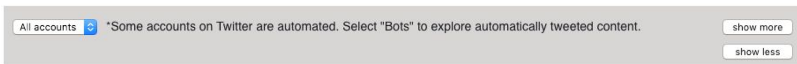


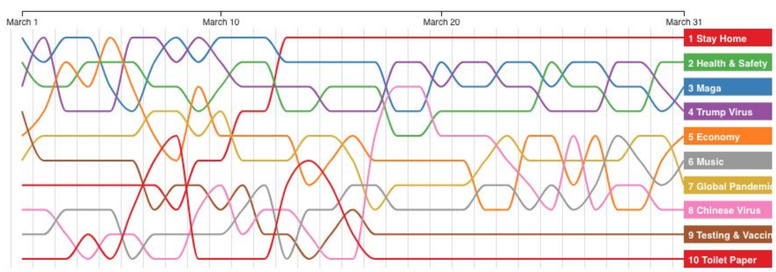
Above: By Unique User Counts
Below: By Raw Tweet Counts



Topic Timeline

To visualize how the conversation around COVID-19 evolved over time, we borrowed a [timeline design](#)[3] from Twitter. We think that this design does a good job of abstracting away from raw twitter counts and focusing on what topics users paid the most attention to over time. To infer topics from the data, we perform community clustering on a hashtag graph defined by hashtag usage statistics in the data (refer to Methods for details). We calculate the percentage of traffic that mention topic-specific hashtags every day, and rank the topics from most-talked about to least-talked about. These ranks then define the shape of the line for each topic in the plot.





This visualization has a few levels of interaction built in. See Appendix for detailed screenshots of each interaction. First, it responds to mouse movements over the timeline with responsive highlighting and a display of the specific date. The motivation behind this choice is twofold: (1) to facilitate an engaging experience for the user, and release us from crowding the x-axis with text.

Second, we highlight the topic-specific trendline when the user hovers over each topic. This allows the user to clearly visualize trends of individual topics that they are interested in. We also display the set of hashtags that make up each cluster above the plot, matching the hashtag color to the topic color for consistency.

Finally, we let the user opt into seeing more or less topics, as well as explore topics used by suspect bot accounts via the control panel at the top. We find interesting patterns in the types of hashtags that suspect bot accounts tend to post. Consistent with ongoing work, we find that bots tweet large amounts of politically charged right-leaning hashtags, though our conclusive power is limited by the small amount of accounts that we evaluated and identified as potential bots (see Methods).

[3] We used the linked page as a design template, but implemented all of the code ourselves in this observable notebook:
https://observablehq.com/@mashabelyi/covid19-discussion-of-coronavirus-on-twitter/2

Usability and Results

To test the efficacy of our visualizations, we conducted a segmented usability study consisting of multiple formative evaluations for each component of our visualization. We chose to evaluate each component individually prior to combining them with the goal of doing a deeper evaluation of each one. We also chose to segment our studies so as not to fatigue our users by forcing them to evaluate each of our four visualizations in depth.

Usability Methods

We took a common approach to each usability study for consistency and interpretability. We began each test with a brief introduction and description about our project. At this stage we also collected demographic data about the participants.

Pre and Post test Questions. We came up with pre-and post-study questions for each visualization to help us assess what, if anything, the participants learn from viewing the visualization.

First Impressions. In the next stage of evaluation, we shared our screen and showed the user the visualization for a brief period of time. We asked the user to talk through the first thing they saw, as well as what they understood the visualization to represent. All studies included a first impression question:

- Take a few seconds and look at this visualization. What is the first thing you notice? Describe what you believe it's showing you.

Functional, task-specific questions. In the next step, we tested a set of functional questions pertaining to the specific goal of each visualization. The goal of these questions was to evaluate how well the user is able to interpret the information presented in the visualization. We asked the users these questions and recorded their response. At this stage we also asked the user to rate the visualization on a 1-5 likert scale for *interpretability* and *engagement*. All studies included the following likert questions:

- On a scale from 1-5, 5 being very easy, 1 being very confusing, rate how understandable this visualization is.
- On a scale from 1-5, 5 being very engaging, 1 being not engaging, rate how engaging this visualization is.

Interactive time. Next we shared the link to an observable notebook and asked the user to share their screen with us while they interact with the visualization. We let the users freely interact with the visualization. We asked the user to talk out loud and note if they encounter any confusion. In addition to some more specific questions that varied by visualization, all studies asked the user to respond to the following questions at this stage:

- What are some strengths and weaknesses of this visualization?
- If there was anything you could change about this visualization, what would it be?
- What would make this more interesting/engaging ?

Wrap it up. We ended the study by asking the user the same pre-test question that we asked in the beginning of the study to see if their response changed. We then thanked the user and noted down any additional thoughts, questions, and recommendations they had for us.

Test Measures

We employ pre and post test question accuracy and switch percentage to measure the extent to which users absorb new information from the visualization. Additionally, we report results of the likert-style questions, which gauge the aesthetic interpretability, the level of engagement the user felt with the visualization, and the usability of the interactions.

Visualization 1: Tweets by state by date

Qualitative Results. We were told that the title should be adjusted to make it more clear what exactly the visualization was showing, so that anyone could easily understand what was being plotted. Additionally, we were told that because no state was highlighted to begin with, users were confused as to why California's data was shown and not the entire country's. Other tips included that it may help bridge the connection between the map and the bar graph if the state in question were highlighted in both the map and the title of the bar graph.

The most jarring comment we received was that all of our users stated that they preferred a real map to our simplified grid. This was a shock, as we expected the grid to be more interpretable and easier to click. Lastly, one of our users mentioned that the data communicated its message effectively despite the lack of normalization and suggested we keep it that way.

Quantitative Results. For our pre-test/post-test questionnaire, both subjects failed to answer the question correctly with the knowledge they possessed prior to exploring our visualization. However, at the conclusion of each usability test, both participants were able to answer the question with no hesitation.

Question	Pre-test Accuracy	Post-test Accuracy
What do you think the overall trend in the number of tweets that mention covid-19 looks like for the entire month of March, from the 1st to the 31st?	0.00	1.00

Below are the results for our aesthetic and functional evaluations where we asked three separate likert-style questions to evaluate different aspects of our visualization. Each question was rated on a scale from 1 to 5, 1 being the lowest score and 5 the highest.

Question	Participant 1	Participant 2	Average Score
Interpretability	3.5	4	3.75
Level of Engagement	2	3	2.5
Usability	4.5	5	4.75

The consensus seemed to be that the visualization was interpretable but not overly so – participants commented that it was difficult at first to make the connection between the map and the bar chart. As for engagement, both commented that it was not very engaging, noting that the grid layout of the map made it visually boring. Finally, both participants found the visualization very usable, stating that it was very simple and straightforward.

Changes Made. From the aforementioned usability study results, we adjusted a few small features as suggested by users to enhance the visual connection between the state map and the bar chart. We then converted the title to a question followed by a short description to ensure the user understands exactly what the visualization is illustrating. Additionally, one of our participants believed that normalization was not necessary and that the message was more effectively communicated by keeping the raw data, so we chose not to normalize. Instead, we chose to add the population of each state above the bar graph to provide some additional explanation of tweet distribution.

The largest change we made was converting the map feature of our visualization from a grid to a geographic map of the United States. We had originally anticipated that this conversion would render the NorthEastern region of the map difficult to click, but upon implementation we found that this issue was negligible and that the new map format increased user engagement.

Visualization 2: Sentiment on Twitter

Qualitative Results. Overall, both participants enjoyed the viewing and interacting with the visualization. Both enjoyed the ability to switch between “tweets” and ‘“retweets” views, as well as the detail interaction on hover over the states. One of the participants raised several issues of clarification pertaining to how exactly the sentiment scores were calculated. At the core of the confusion was the desire to know exactly what swayed the sentiment one way or another on each day. The participant questioned whether one overall sentiment score was a descriptive enough measure of the rich conversation happening on twitter, or if the score should be further broken down into categories.

Quantitative Results. Both participants changed their response to the pre-test question after having interacted with the visualization, which indicates that new knowledge was acquired during the interaction.

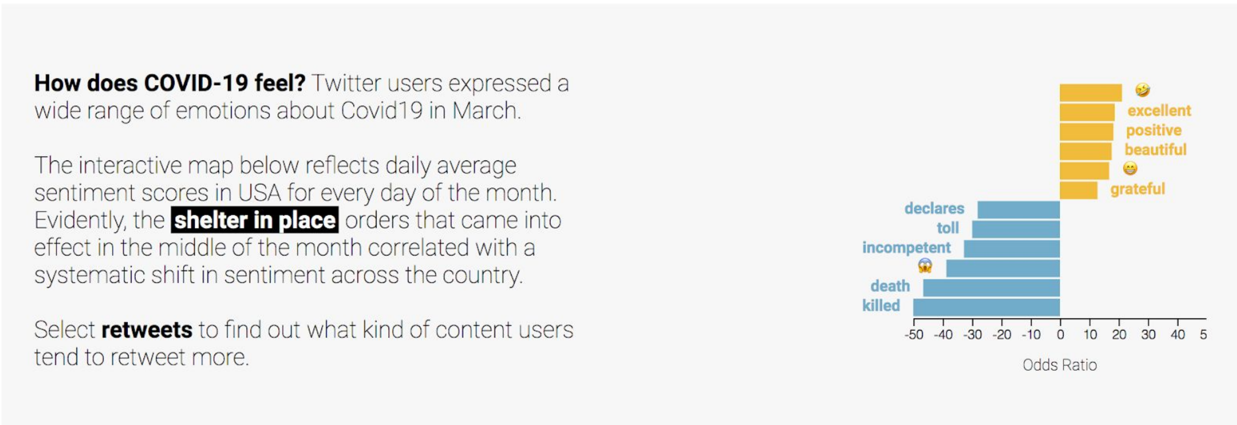
Question	Switch Percentage
What emotions do you think people from the United States expressed on Twitter in the month of March? Do you think the overall sentiment changed at all during the month of March? How?	1.0

Results of likert-style questions pertaining to aesthetics:

Likert Ratings	Participant 1	Participant 2	Average
On a scale from 1-5, 5 being very clear, 1 being very confusing, rate how would you rate this visualization?	3	5	4
On a scale from 1-5, 5 being very engaging, 1 being not engaging, how would you rate this visualization?	4	5	4.5

Changes Made.
Following the advice of our participants to introduce more context into the visualization, we plan to add an additional panel to the chart that displays the most distinctively positive and negative words used on Twitter on a given day. The panel will display a diverging bar chart, with the top half displaying bars for the top 5 most positive words in yellow (matching the yellow in the main graph), and the bottom half dedicated to the top 5 most negative words. We plan for the content in the panel to update as the user moves the mouse across the page, mimicking the closeup panel.

We think that this update would provide additional insight into exactly what made the average sentiment more positive or more negative on a given day. At the time of writing, we unfortunately have not had time to implement this update. However, we improved the messaging that precedes the visualization to include a static bar chart with the top 5 most positive and negative tokens found in the full dataset.



Visualization 3: Hashtag Clusters

Qualitative Results. Overall participants found this graph to be engaging and informative. However, they suggested this visualization could be missing a clear structure. Some suggestions included a succinct text explanation or a legend for the color groups. Additional feedback pointed at the potential of presenting a meaningful context for the most used hashtag pairs, for example incorporating sample tweets.

Quantitative Results. Comparing the results to the same question asked pre-test and post-test, we believed this visualization achieved task success. Both participants gained some new information and were able to recall afterwards.

Question	Pre-test Accuracy	Post-test Accuracy
Within the Covid-19 discussion, can you think of pairs of topics/hashtags that are often used together? (This is excluding hashtags like #covid-19, #coronavirus, #pandemic, etc)	0.0	1.0

Likert Ratings	Participant 1	Participant 2	Average

On a scale from 1-5, 5 being very clear, 1 being very confusing, rate how would you rate this visualization?	4	3.5	3.75
On a scale from 1-5, 5 being very engaging, 1 being not engaging, how would you rate this visualization?	3	5	4

Changes Made. We assigned topics, based on the aforementioned clustering method, to the hashtags and added a legend for the coloring. Additionally, we created a second graph using hashtag pairs by raw count to compare against the current one. Users can now compare the two views and see the difference in topics themselves.

Visualization 4: Topic Timeline

Qualitative Results. It took both participants some time to discover and understand the hover-on-topic interaction. Both discovered it about 30 seconds into the interactive portion of the usability test. However, both had an “aha” moment when they finally discovered and really enjoyed using it for the remaining duration of the test. Both mentioned that they would have liked unlocking even more insights on hover, such as the number of tweets per topic per day, or the relative changes in tweetage per topic per day. Both also expressed interest in understanding what external events may have influenced the changes in trends every day.

Overall, the participants enjoyed exploring this visualization. However, they also expressed a few points of confusion through qualitative feedback. Most notably, it was not completely clear what the “select bots” option implied. Also, both participants expressed desire to see more detail about each topic, such as the number of tweets corresponding to the peaks and valleys of the line plots.

Quantitative Results. We implemented 2 pre-test questions for this visualization. We report mixed results for the two. Both participants changed their response to one of the pre-test questions after having interacted with the visualization, which indicates that new knowledge was acquired during the interaction. Both did not understand the role of bots on social media prior to the interaction, which did not cha

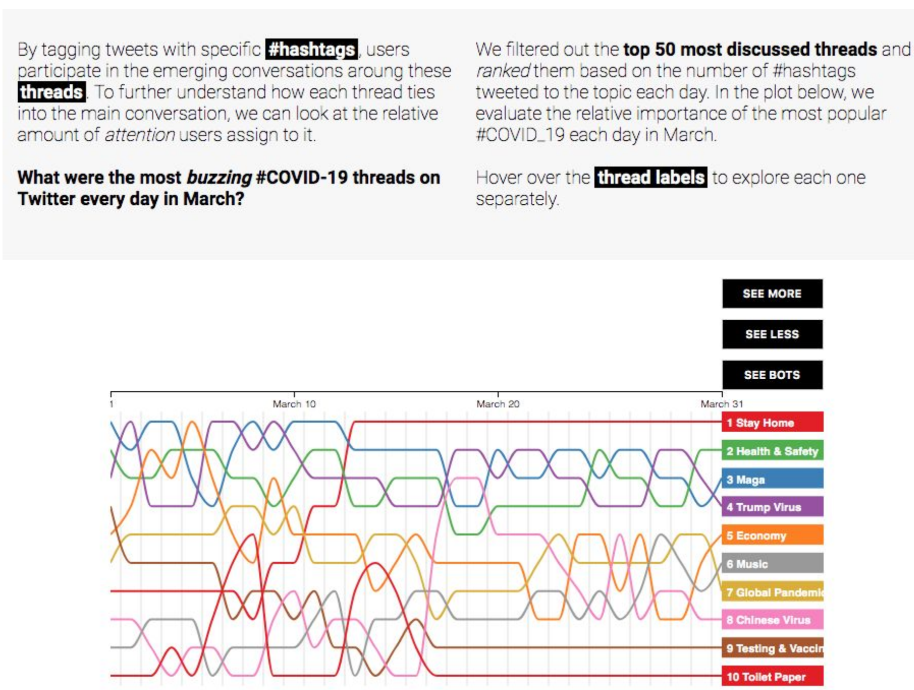
Question	Switch percentage
What do you think were the top 3 most discussed topics around Coronavirus on Twitter in March? Do you think the top 3 most discussed topics remained the same every day in march? Or they changed over time?	1.0
Are you aware of the role Bots (automated accounts) play on social media?	0

Likert ratings. Participants 1 and 2 found the visualization to be both clear and engaging.

Likert Ratings	Participant 1	Participant 2	Average
On a scale from 1-5, 5 being very clear, 1 being very confusing, rate how would you rate this visualization?	5	4	4.5
On a scale from 1-5, 5 being very engaging, 1 being not engaging, how would you rate this visualization?	4	5	4.5

Changes Made

Though participants rated the visualization as mostly clear and engaging, this usability test exposed quite a few points of potential improvements. To make the visualization more intuitive, we have improved the messaging that viewers read at the top of the chart. We also updated the controls interface to make it less clunky and fit in with the rest of the design a bit more. We also hid any information about Bots and only left a “SEE BOTS” button. When a user hovers over the button, we provide additional information about what the button does. If they choose to click it, we show data from bot-generated accounts. We hope that with this change, there are less instructions to comprehend at first glance, and that users are not forced to interact with the Bot feature (which is at best experimental right now) but may choose to do so.



Updated interface

Unfortunately, we have not had time to implement additional changes at the time of writing this report, but we outline a few changes that we hoped to implement below:

In response to the desire for more information about each topic, we plan to display the total percentage of the conversation that was devoted to each topic on hover. We opt to display the amount of traffic in terms of percentage rather than raw counts because, as we mention in the Data and Methods sections, our dataset only reflects ~1% of the total Twitter traffic. Hence, it is more informative to display percentages, which can still be compared from topic to topic. The value for each topic would show up on the left side of the plot, next to the beginning of the trendline for the topic. It would be color-coded to match the topic label color.

Second, we would like to add a more clear description of how the topics were generated and how the ranking was performed. Also it would be great to emphasize that the topics in view are only a subset of the *top N* topics that our algorithm discovered, and not the whole conversation. Along these lines, it may make sense to label the top line as “the most discussed topic” on March 31, to further disambiguate what the rankings mean.

Responding directly to user feedback, we plan to allow the user to freeze the highlighted view of one or more topics on click, enabling them to compare topics to each other more easily without the background clutter of other topics getting in the way.

Finally, we intended for this timeline plot and the hierarchical plot to sync on topic colors as well as the hashtags and hashtag links within each topic. We were not able to get this working in time for submission, so we present them as two separate plots on the final visualization page.

Methods

Location by State
We limit our analysis to tweets and retweets that were posted in the USA. This is a potential limitation of our study, since online conversations likely cross international boundaries. We focus on the US subset for this study, and defer analysis of the full dataset for followup work.

There are 4 ways in which a tweet is marked with location in the API response: *geolocation coordinates*, *user location*, *place name* and *place coordinates*. Due to a recent change in the way Twitter collects data, *geolocation* for most tweets in our dataset was not available. We use a three-fold process to extract state-level location tags from the other three attributes for as many tweets as possible:

1. For tweets that have latitude and longitude coordinates we reverse geocode these coordinates. If the returned address is in the USA, we label the tweet with the returned state label. (~0.1% of all tweets)
2. When users opt into location-tagging their post, they generally select from a list of candidate Places. ~3% of all tweets in our dataset are assigned a such place with a corresponding bounding box for the lat+lng coordinates. We reverse geocode the center of this bounding box and keep tweets for which the geocoder returns a USA address[4].
3. User accounts of ~70% of the tweets in our dataset have a location associated with them. The location is user-specified, which means it does not always stand for a real geographical region. Perhaps for privacy reasons, many users choose made-up locations such as "La-La land", "The Moon", "Hogwarts".

We sequentially search the user specified location string for occurrences of (1) state abbreviations (CA, AL, MA, etc), (2) full state names, (3) USA cities, and (4) various spellings of "the united states of America" (e.g. usa, us, america). If there is a match at any step, we convert the matching string to a 2-letter state abbreviation and label the tweet with that location.

We find that using this method with a list of other countries included as step (5), we are able to resolve 51% of the user_location strings in the dataset into a valid location. The unresolved 50% of the strings fall into one of two categories: (1) names of foreign cities that we did not include in our processing step, and (2) made up locations such as *SkyDome*, *wherever threads are written, here, heaven.*

[4] A potential flaw in this approach is that if a post is tagged with a “USA” place tag, then the center of the encompassing bounding box might end up in the middle of the US, which would wrongly lead us to believe that the Tweet originated in Kentucky. However, it is worth noting that we don’t observe an unusual inflation in post volume from any of the central states in our final analysis.

Bot Detection
We collect Botometer (Davis et al., 2016) scores for 18000 accounts in our dataset. The API is rate limited, which prevents us from collecting scores for all user accounts. This greatly limits the amount of analysis we can do with the Botometer scores, but nevertheless yields some interesting insights.

Botometer scores each user account with the likelihood of that account being an automated social media bot. The score is calculated based on user activity, following, and other user-specific attributes. Since all our tweets are in English, we use the english overall bot-likelihood score that the Botometer API returns.

We select accounts for processing based on the following criteria:

(1) accounts that tweeted novel cotent (not retweets) >20 times in March (3614 accounts),

(2) top 15,000 most retweeted accounts. All of the 15,000 most retweeted accounts were retweeted >30 times in March.

We select highly active user accounts under the assumption that bots might be tweeting more often than regular users. However, we recently learned that this assumption was lacking in domain knowledge. In an analysis of the same dataset that we use for our work, Ferrara (2020) finds that accounts with highest botometer scores actually tweet less than other accounts. Presumably, most of the bot activity originates from a large number of accounts, each one tweeting only a few times in March.

We consider a user to be a potential bot if it receives a probability score >0.5 from Botometer. This is a reasonable score based on Ferrara's analysis (Ferrara 2020, Table 4). In his evaluation of 1,056,124 accounts, the top 15th percentile of the bot score distribution had an average bot probability 0.6.

Sentiment
We calculate a sentiment score [-1,1] for each tweet in the dataset. We use the Vader[5] sentiment analysis tool (Hutto and Gilbert, 2014) that is optimized for sentiment detection on social media. Vader differs from other sentiment extraction tools in that it includes emojis and word commonly used internet slang (e..g *nah*, *meh*) in its sentiment lexicon. We then calculate the average sentiment score for each day and each state in our dataset.

We assess the statistical significance of each state point with a t-test comparing the distribution of tweet sentiment scores for a given day and state to a randomly sampled set of values between -1 and 1 from a normal distribution with mean=0, std=0.3 (motivated by the fact very polarized tweets with sentiment scores -1 and 1 are rare). We find that applying a smoothing window of 2 days to the data results in statistically significant averages for most of our data points.

[5] https://github.com/cjhutto/vaderSentiment

We also identify the most distinctive positive and negative unigrams in the dataset with a chi-square test of independence. For each token, we evaluate the hypothesis that the usage pattern of that token in positive and negative tweets is indistinguishable.

Topic Clusters

We use a graph clustering approach to infer sub-topics of conversations observed in the data. We denote each hashtag present in the dataset as a node in a graph. Two nodes are connected if the two hashtags appear together in at least one tweet in our database. The edges are weighted by the number of unique users who use the two hashtags in the same tweet at least once. We use the unique user count rather than unique tweets count to avoid being fooled by bots who produce large numbers of tweets with identical hashtag inventory.

For interpretability, we remove all hashtags that were tracked during data collection. These include hashtags like *#coronavirus, #covid19* that are shared across all posts in the dataset. We also remove common misspelings of the same hashtags (*#covd19, #coronavirius)*. Finally, we remove noisy, uninformative tags like *#tuesdayvibes* by excluding all tags that begin with the day of the week.

We apply a label propagation community detection algorithm (Raghavan et al 2007) to identify clusters in the hashtag graph. We use the networkx[6] python implementation. Since label propagation is non-deterministic, we execute the algorithm 20 times. After each execution, we identify the most frequent hashtag in the largest 50 communities that the algorithm outputs. We also keep track of cluster membership of each other node in the graph at every iteration. We find that, with the exception of a few tags, the same set of 50 hashtags define the top 50 clusters in each of the 20 iterations. We select the top 50 to be "cluster leaders", and form final clusters by matching each hashtag node with the "cluster leader" that it was most commonly associated with.

For example, in the 20 iterations, *#supertuesday* formed its own cluster 16 times, and was part of the #*maga* cluster 4 times. In the final assignment, it forms its own cluster.

For the Topic Timeline plot, we filter out the top 50 largest communities by hashtag usage and manually assign topic names to the formed clusters[7]. We then calculate the number of times each hashtag is used in our dataset, factoring by date, and use these values to rank each topic by usage every day.

Hashtag Pairs

In the creation of the network graph, we took the dataset of all COVID-19 related tweets with hashtags and removed all synonyms of #coronavirus because they were disportionately higher than all other hashtags (given we only looked at tweets about the current pandemic).Then we further filtered down the dataset by removing entries that shared both the same userID and the exact list of hashtags. In other words, if a user @AllAmericanGirl used the same set of hashtags

[6] https://networkx.github.io/

[7] See list of the 50 communities and their hashtags here.

['aag', 'aag2020'] 200 times, we would only count it once. Then we tallied every pair of hashtags that appeared in the same tweet and returned the most common pairs of hashtags.

Limitations:

There was one caveat: say a tweet used 10 hashtags, then we would get (10-1)! pairs of hashtags. This meant we may be counting tweets with heavy use of hashtags more heavily. We thought this problem was partially mitigated by us only using the pair count to select the top 300 pairs and not using it to encode the edge width or opacity.

Another limitation here was some of the hashtags weren't included in the clusters and therefore were grouped as "Other" even though they conceivably were part of the existing cluster. If we had more time, we would have made sure they were also part of the clusters.

Limitations

We are aware of the sampling limitation of the Twitter search API which returns roughly 1% of the full twitter stream, which may introduce some biases in the data. Studies of the effect of this subsampling on downstream analysis have brought to attention its potential degrading effects on modeling and interpretability (Wu et al, 2020).

Additionally, we observe a dramatic peak in post frequency in all the states on March 3 and March 4. This is potentially caused by the API spuriously returning a larger percent of the data on those two days.

Author Contributions

Task	Contributors
Data collection - hydrate Twitter dataset	Masha
Data processing - filter for english Tweets, process location, parse datetime, add sentiment labels, hashtag clustering	Masha
Lit review	Masha, Joanne
Identifying research questions and hypotheses	Masha, Cameron, Joanne
Initial design sketches	Joanne, Cameron
EDA, viz prototypes (D3 assignment)	Cameron, Joanne, Masha
User Testing Guidelines	Cameron
Tweets by date Viz - design + implementation + usability test	Cameron
Topic Clusters Viz - design + implementation + usability test	Joanne
Sentiment Grid Viz - design + implementation + usability test	Masha
Topics Timeline Viz - design + implementation + usability test	Masha
Diverging Bar Chart	Joanne, Masha
Final writeup	Cameron, Joanne, Masha
Final website design	Masha, Cameron, Joanne
Final website implementation	Masha

References

T. Alshaabi, J. R. Minot, M. V. Arnold, J. L. Adams, D. R. Dewhurst, A. J. Reagan, R. Muhamad, C. M. Danforth, and P. S. Dodds. **How the world's collective attention is being paid to a pandemic: COVID-19 related 1-gram time series for 24 languages on Twitter.** *Pre-print*. 2020. [Project Link]

Chew C, Eysenbach G. **Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak.** Sampson M, editor. PLoS One.

Emilio Ferrara. **#COVID-19 on Twitter: Bots, Conspiracies, and Social Media Activism**. *Pre-print.* 2020.

Fu K-W, Liang H, Saroha N, Tse ZTH, Ip P, Fung IC-H. **How people react to Zika virus outbreaks on Twitter? A computational content analysis**. Am J Infect Control. 2016.

Lee JY, Sundar SS. **To Tweet or to Retweet? That Is the Question for Health Professionals on Twitter.** Health Commun. 2013

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. **Fake news on Twitter during the 2016 U.S. presidential election**. Science, 2019

Raghavan, Usha Nandini, Réka Albert, and Soundar Kumara. **Near Linear Time Algorithm to Detect Community Structures in Large-Scale Networks.** Physical Review E 76.3 (2007)

S Shahsavari et al. **Conspiracy in the Time of Corona: Automatic Detection of COVID-19 Conspiracy Theories in Social Media and the News**. *Pre-print.* 2020.

Signorini A, Segre AM, Polgreen PM. **The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic**. Galvani AP, editor. PLoS One. 2011

Laura Spinney. **In Congo, fighting a virus and a groundswell of fake news**. Science, 2019

O. Varol, E. Ferrara, C. Davis, F. Menczer, A. Flammini, "**Online human-bot interactions: Detection, estimation, and characterization**" in Proceedings of the 11th International AAAI Conference on Web and Social Media (AAAI, 2017), pp. 280–289. [Botomete]

Wu, Siqi & Rizoiu, Marian-Andrei & Xie, Lexing. **Variation across Scales: Measurement Fidelity under Twitter Data Sampling**. 2020

Xinnng Gui et al. **Understanding the Patterns of Health Information Dissemination on Social Media during the Zika Outbreak**. AMIA Annu Symp Proc. 2018.

Yuxin Zhao, Hulian xu. **Chinese Public attention to COVID-19 Epidemic: Based on Social Media.** *Pre-print*. 2020.

Appendix

Data Sources

- COVID-19-TweetIDs dataset

Code

- Our Github repo with data processing code
- Our final project Observable notebook
- Our tweet by date visualization *individual* Observable notebook
- Our covid discussion tracker *individual* Observable notebook
- Our hashtag pairing visualization *individual* Observable notebook
- Our hashtag bar graph visualization *individual* Observable notebook
- Our emotional response to covid visualization *individual* Observable notebook
- Website code: https://github.com/mashabelyi/covid-twitter

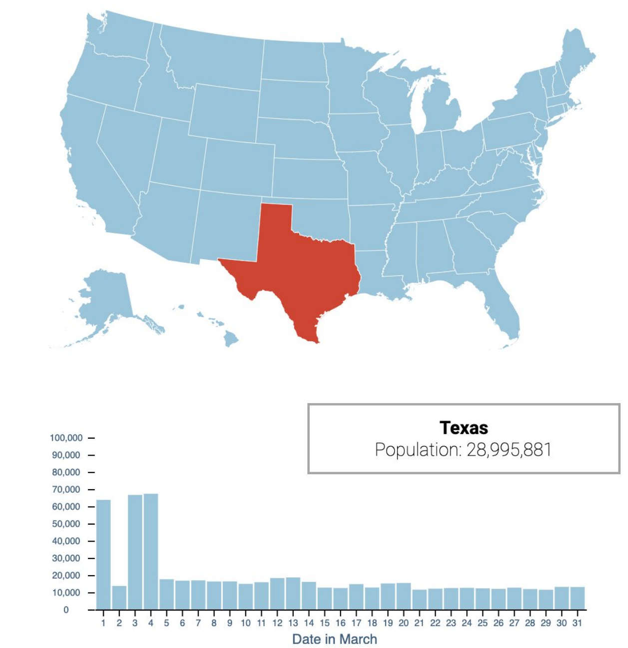
Tutorials

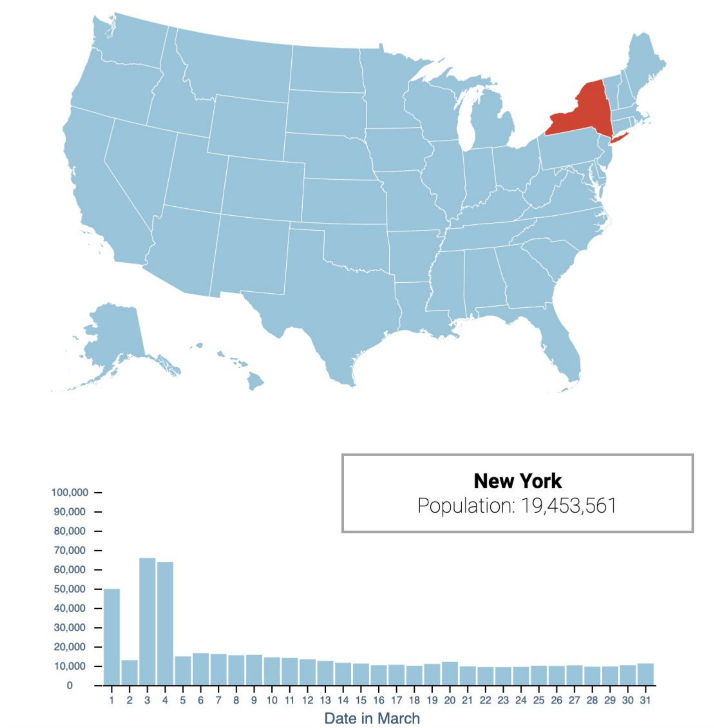
Hierarchical Edge Bundling in D3

Bar Graph in D3

Additional screenshots

Tweets by Date:





Topics:

How are different hashtags used together in Covid-19 discussions?

From analyzing the pairs of hashtags often used together on Twitter, we found that the most commonly used hashtags tend to be **politically charged**.

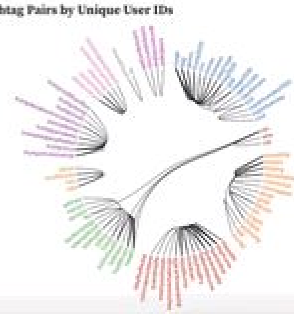
	Hashtag 1	Hashtag 2	Count
0	#CPAC2020	#TrumpVirus	522
1	#TrumpVirus	#TrumpVirusCoverup	408
2	#news	#politics	350
3	#CPAC2020	#TrumpVirusCoverup	329
4	#politics	#world	327
5	#USA	#news	325
6	#news	#world	317
7	#USA	#world	317
8	#Pencedemic	#TrumpVirus	316
9	#USA	#politics	315

However, **only a handful of Twitter accounts were using** these most trended hashtags. See which Twitter accounts used these hashtags **more than 100 times** below:

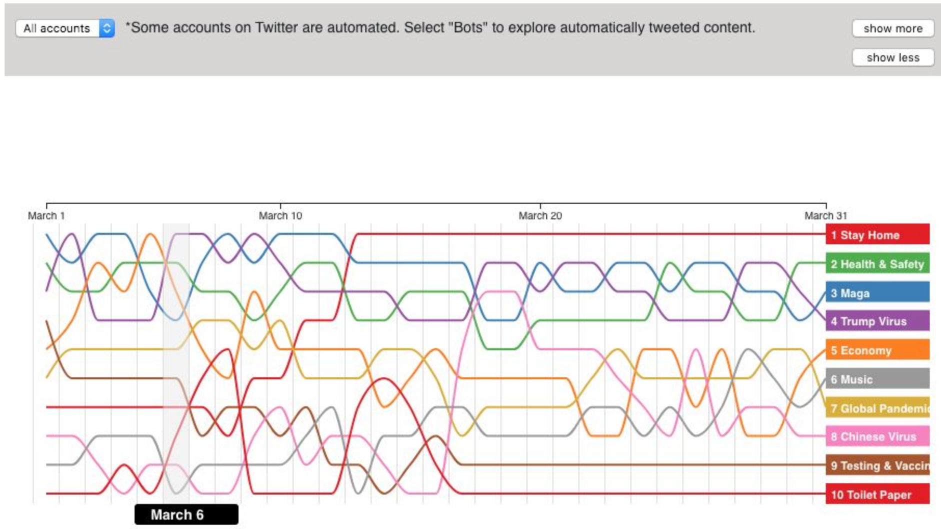
Choose a pair of top hashtags

Grouping hashtags by unique user IDs, we found the hashtags to be less politically charged and more broadly about health, the economy, the quanrantine life, etc. Hashtags are also most commonly related to those in the same topic.

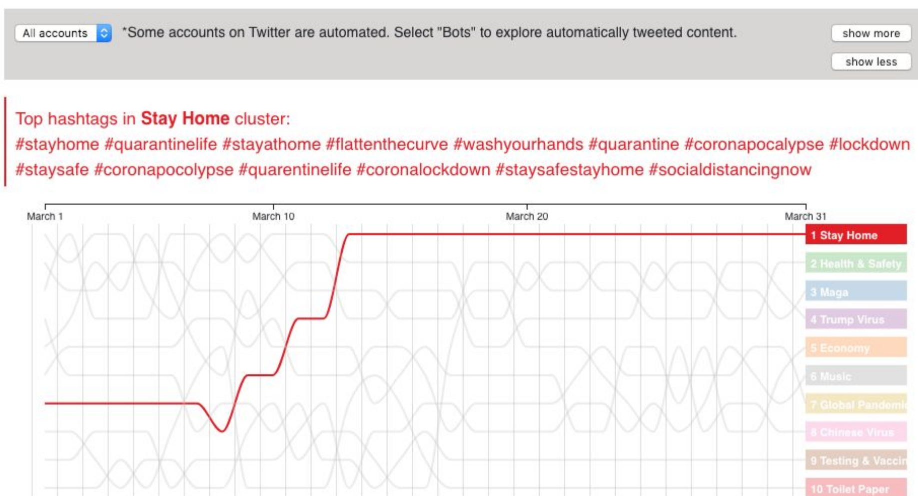
Top 100 Hashtag Pairs by Unique User IDs



Topic Timeline



Feedback on hover over the timeline.



Feedback on hover over a topic (Stay at Home) label

Design sketches: