

Dynamic topic modeling of the COVID-19 Twitter narrative among U.S. governors and cabinet executives

Hao Sha and Mohammad Al Hasan and George Mohler P. Jeffrey Brantingham

Indiana University - Purdue University
Indianapolis

University of California
Los Angeles

Abstract

A combination of federal and state-level decision making has shaped the response to COVID-19 in the United States. In this paper we analyze the Twitter narratives around this decision making by applying a dynamic topic model to COVID-19 related tweets by U.S. Governors and Presidential cabinet members. We use a network Hawkes binomial topic model to track evolving sub-topics around risk, testing and treatment. We also construct influence networks amongst government officials using Granger causality inferred from the network Hawkes process.

Introduction

By mid-April 2020, the number of active COVID-19 cases has reached over 2 million and the number of deaths is over 140,000 world-wide. The United States has the largest share of confirmed cases (over 670,000) and confirmed deaths (over 27,000). Without a vaccine yet available, states throughout the U.S. are attempting to control transmission and reduce strain on the healthcare system through school and business closings, along with shelter-in-place orders. Careful planning and coordination is needed both to minimize risk from the disease, and to minimize the long-term economic impact.

In the U.S., a combination of federal and state-level decision making has shaped the country's response to COVID-19. The response is quickly evolving, making it difficult to understand how decision makers have influenced each other, and whom among the decision makers have emerged as leaders on different topics. To overcome this difficulty, we analyze the Twitter narrative of various decision makers through dynamic topic modeling. Specifically, we analyze a dataset of all COVID-19 related tweets by U.S. Governors, the President, and his cabinet members between January 1st 2020 and April 7th 2020. We use a Hawkes binomial topic model (HBTM) (Mohler et al. 2016) to track evolving sub-topics around risk, testing and vaccination/treatment. The model also allows for estimation of Granger causality (Xu, Farajtabar, and Zha 2016) that we use to construct influence networks amongst government officials.

Our work contributes to the growing body of literature on social media analytics and COVID-19. A summary of the most related work is as follows. In (Cinelli et al. 2020), general COVID-19 related topic diffusion across different social media platforms is analyzed. In (Yin et al. 2020), the authors study COVID-19 discussions on Chinese microblogs. Gender differences in COVID-19 related tweeting is investigated in (Thelwall and Thelwall 2020a) and in (Thelwall and Thelwall 2020b) the authors analyze consensus and dissent in attitudes towards COVID-19. Geolocated tweets are used to estimate mobility indices for tracking social distancing in (Xu, Dredze, and Broniatowski 2020).

Hawkes Binomial Topic Model

We analyze COVID-19 related tweets by U.S. governors and cabinet members using a network Hawkes binomial topic model¹ (HBTM) (Mohler et al. 2016) with intensity $\lambda_s(t, \vec{m})$ at node s in the network determined by,

$$\lambda_s(t, \vec{m}) = \mu_s(t) J_0(\vec{m} | p_0^s) + \sum_{t' > t_i} \theta_{ss_i} \omega_{ss_i} e^{-\omega_{ss_i}(t-t_i)} J_1(\vec{m}, \vec{m}_i | p_{off}^{ss_i}, p_{on}^{ss_i}). \quad (1)$$

A Hawkes process is a model for contagion in social media where the occurrence of a post increases the likelihood of more posts in the near future. In the HBTM, tweets are represented as bags of words following a Binomial distribution. When viewed as a branching process, the daughter event bag of words is generated by randomly turning on/off parent words through independent Bernoulli random variables.

In Equation 1 events at time t_i are associated with a mark $\vec{m}_i$, a vector of size W , the number of words in the overall dictionary across events. The binary variables indicate whether each word is present or absent in the event at time t_i . Spontaneous events occur according to a Poisson process with rate $\mu_s(t)$ at node s in the network (here a node is either a governor or cabinet member). Unlike in (Mohler et al. 2016), we let the spontaneous rate vary in time to reflect the exponential increase in overall COVID-19 related Twitter activity (for estimation we use a non-parametric histogram). The mark vector of spontaneous events is determined by,

$$J_0(\vec{m} | p_0^s) = p_0^s \sum_{j=1}^W m_j (1 - p_0^s)^{W - \sum_{j=1}^W m_j}, \quad (2)$$

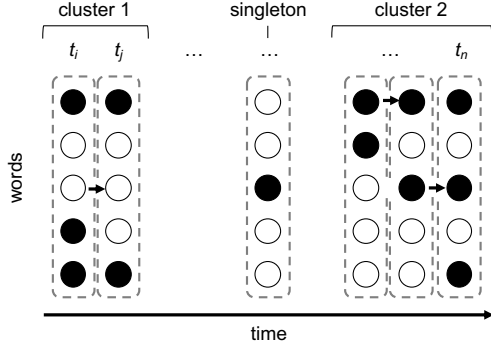


Figure 1: In the HTBM, spontaneous events occur with marks generated by a binomial random variable over the dictionary of keywords contained in the data set. Events then trigger offspring events whose marks are generated by switching parent event words off (white circle) with probability p_{off} and on (black circle) with probability p_{on} . Unique events are delineated with dashed lines. Clusters are groups of parent daughter events connected by triggering.

which is the product of W independent Bernoulli random variables with parameters p_0^s

The parameter $\theta_{ss'}$ determines the expected number of tweets by individual s triggered by a tweet by individual s' and can be viewed as a measure of influence. The expected waiting time between a parent-daughter event pair is given by $\omega_{ss'}^{-1}$. The mark of a daughter event is determined by two independent Bernoulli processes. Each word absent, or “turned off,” in the parent bag of words is added to the bag of words of the child event with probability $p_{on}^{ss'}$. Each word present in the parent bag of words is deleted with probability $p_{off}^{ss'}$. Thus J_1 is given by,

$$J_1(\vec{m}, \vec{m}_i | p_{off}^{ss'}, p_{on}^{ss'}) = p_{on}^{ss'} W_1^{\vec{m}, \vec{m}_i} (1 - p_{on}^{ss'}) W_2^{\vec{m}, \vec{m}_i} p_{off}^{ss'} W_3^{\vec{m}, \vec{m}_i} (1 - p_{off}^{ss'}) W_4^{\vec{m}, \vec{m}_i}, \quad (3)$$

where $W_1^{\vec{m}, \vec{m}_i}$ is the number of words present in the child vector and absent in the parent vector, $W_2^{\vec{m}, \vec{m}_i}$ is the number of words absent in both vectors, $W_3^{\vec{m}, \vec{m}_i}$ is the number of words in the parent vector absent in the child vector, and $W_4^{\vec{m}, \vec{m}_i}$ is the number of words present in both vectors.

After removing stop words we restrict the dictionary to the W most frequent words, on the order of several hundred most frequent words across tweets. The Model given by Eq. 1 can be viewed as a branching process and is estimated using Expectation-Maximization (EM) (Mohler et al. 2016). Using the EM algorithm for estimation has the added benefit that branching probabilities, estimates of the likelihood that tweet i was triggered by tweet j , are jointly estimated with the model:

$$q_{ij} = \frac{\theta_{s_i s_j} \omega_{s_i s_j} e^{-\omega_{s_i s_j} (t_i - t_j)} J_1(\vec{m}_i, \vec{m}_j | p_{off}^{s_i s_j}, p_{on}^{s_i s_j})}{\lambda(t_i, \vec{m}_i)}. \quad (4)$$

These branching probabilities can then be clustered to generate families of dynamic topics over time (Mohler et al. 2016).

Related work

We note that Hawkes branching point processes in general are a popular model for mimicking viral processes on social media. Previous studies have utilized temporal point processes to model Twitter (Zhao et al. 2015; Simma and Jordan 2012), Dirichlet Hawkes processes (Du et al. 2015; Xu and Zha 2017; Lai et al. 2014), joint models of information diffusion and evolving networks (Farajtabar et al. 2017), Hawkes topic modeling for detecting fake retweeters (Dutta et al. 2020), and Latent influencers are modeled in (Tan, Rao, and Neville 2018) using an Indian buffet Hawkes process. For a review of point process modeling of social media data see (Kim, Paini, and Jurdak 2020).

Compared to standard LDA-type Hawkes processes, the HBTM has the advantage that it jointly estimates a network that can be used to measure influence; additionally, HBTM automatically detects the number of clusters. The temporal aspect of HBTM-like dynamic topic models tend to improve topic coherence in relation to LDA (see Figure 2).

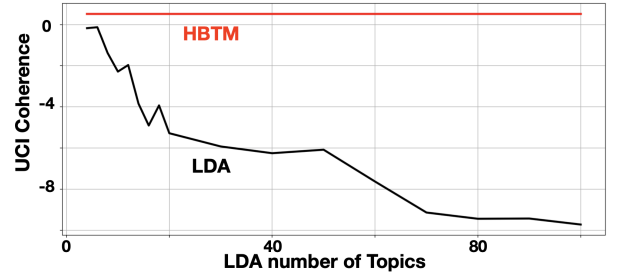


Figure 2: UCI coherence of HBTM vs. LDA when applied to COVID-19 related tweets by governors and cabinet members.

Data

We first collected the verified Twitter handles of all U.S. state governors, presidential cabinet members, and the president (a total of 73 politicians, see Fig. 5 for their handles). Next, we used the Twitter API to query all tweets by these users during the period of January 1, 2020 to April 7, 2020. We then performed a keyword expansion (Buntain, McGrath, and Behlendorf 2018; Mohler et al. 2016) to extract a list of keywords related to COVID-19. This method iteratively adds keywords to a query list whose frequencies in the set of matching tweets are significantly higher than in the general sample. We then scanned the corpus with the expanded keyword list, obtaining a set of 7881 COVID-19 related tweets by these politicians. These tweets were further sorted in time-ascending order and converted to a bag-of-word representation. The vocabulary was then restricted to the top 425 words according to frequency.

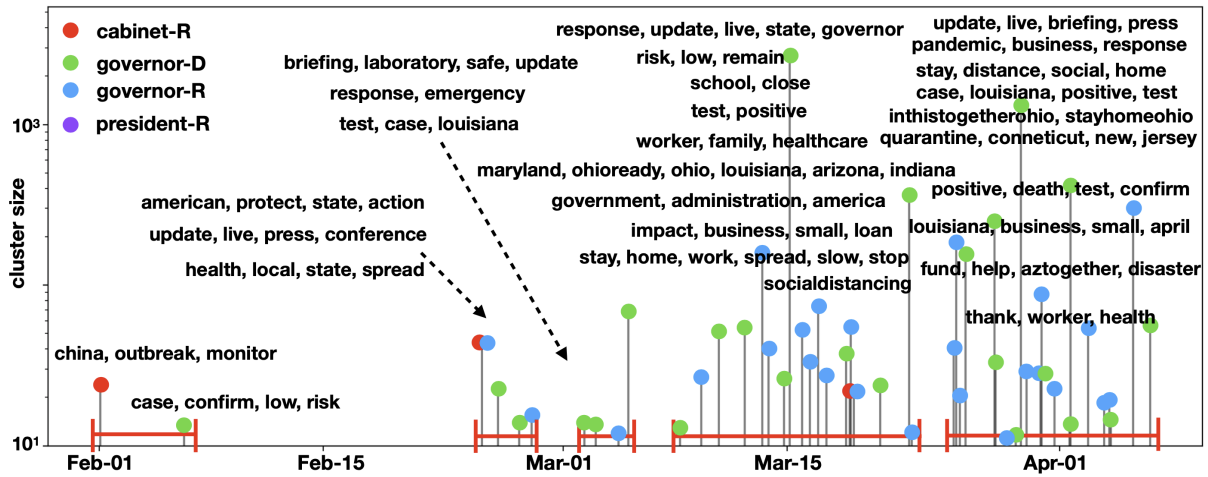


Figure 3: Topic timeline. Clusters with size greater than 10 are pinned. Keywords indicate the topic of the clusters. The marker color indicates the dominant component of the cluster.

Results

We cluster the data into space-time topics by sampling the branching probabilities q_{ij} in Equation 4. In particular, we assign tweets to the same group when a link between tweet i and tweet j is sampled. In Fig. 3, we show topic clusters over time consisting of more than 10 tweets. Each marker height represents the size of the cluster and the most frequent keywords per marker indicate the topics of the clusters.

The clusters show roughly four phases in time, with a significant gap between the first phase and the rest. In the first phase (early February), the federal government (most frequent handle @SecAzar, Alex Azar, Sec. of Health) informed the public of the **outbreak** in **China** and claimed to closely **monitor** the situation. Also in this phase several state governors (most frequent handle @NYGovCuomo, Andrew Cuomo, Gov. of New York) started reporting **confirmed cases**, but stated that the **risk** was **low**, as the number of cases was limited.

The second and the third phases (early March) appeared almost a month later. From the keywords in these two phases, we can see that the government started to take **action** to **protect** the **American** citizens (possibly overseas in the regions of the outbreak). We can also see that **live updates** and **press conferences** were given to **brief** the public. Keywords like **spread** and **emergency** indicate that the outbreak was getting worse in the U.S. Meanwhile, the keyword **test** was mentioned frequently alongside **laboratory**, as limitations in U.S. testing was driving some of the narrative.

The fourth phase starts around mid-March, when clusters became larger and denser. In this phase, **live updates** were held by many governors on a regular basis (the highest peak in Fig. 3). We also see the separation between the federal and state governments, as the clusters divided into **government, administration, america** and the various states (**maryland, ohio, louisiana, arizona, indiana**). The Louisiana governor John Bel Edwards (@LouisianaGov) and the Ohio governor Mike DeWine (@GovMikeDeWine) were among the most active on Twitter sending information to the people in their respective states.

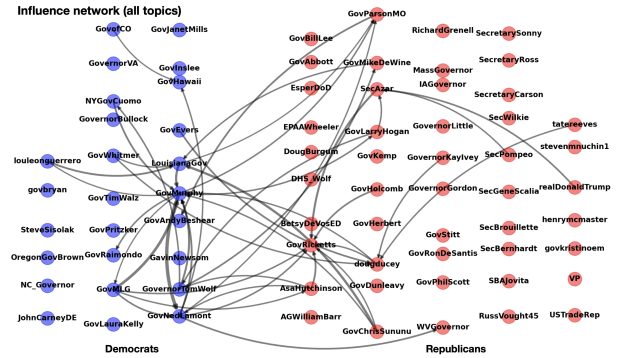


Figure 4: Granger causality (Xu, Farajtabar, and Zha 2016) influence network. Democrats (blue), Republicans (red). Weights of the edges of the directed graph correspond to the fraction of events estimated to be triggered across the edge. Edges with weights less than 10 are removed.

The topic of **risk** appears in this phase, and the message is that risk **remains low**. New topics also emerged on social distancing policies such as **school close**, **stay home**, and **work (from) home**. During the third phase the government began addressing problems like **healthcare** for **workers** and **families**, and **loan(s)** for **small businesses** due to the **impact** of the pandemic. The slogan **social distancing** was widely adopted in this phase.

In the most recent phase, a cluster with frequent words **live update**, **press conference**, and **briefing** is the largest, alongside a narrative around the number of **tested**, **confirmed positive** and **death cases** in different states. The Louisiana and Ohio governors continued to be the most active. Also **small businesses** remained a concern during this phase and the keyword **disaster** indicates the negative impact of COVID-19. Meanwhile, **quarantine** and **stay home** were encouraged and reiterated on Twitter. The sacrifices of **health workers** were acknowledged (**thank**).

In Figure 4, we show inferred influence among governors

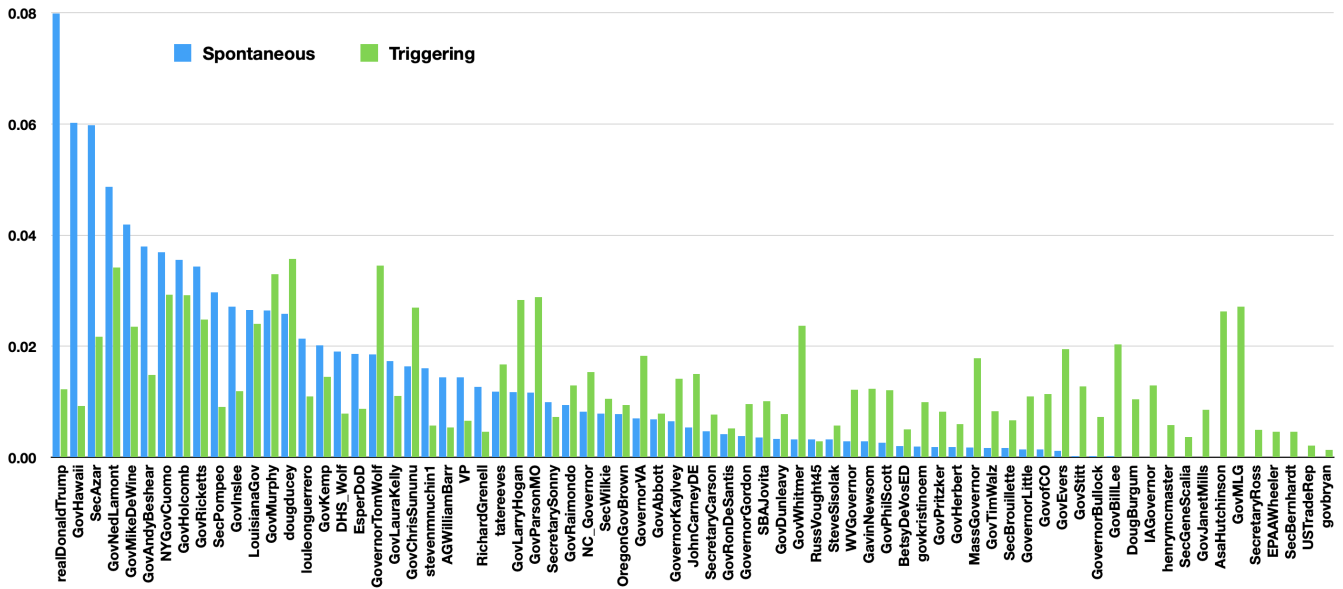


Figure 5: Spontaneous vs. triggering effects of politicians on Twitter. Vertical axis: base intensities (spontaneous) and effective influences (triggering) are normalized over politicians; horizontal axis: Twitter handles of politicians. To save space, vertical axis is truncated at 0.08, rendering President Trump’s spontaneous rate off the chart (~ 0.16).

and cabinet members by plotting a network where each edge weight from $i \rightarrow j$ is determined by the total estimated number of tweets triggered at node j by tweets from node i . The network shows influence across party lines, with Democrat governors **GovNedLamont**, **GovernorTomWolf**, **GovMurphy** and **LouisianaGov** highly connected with Republican governors **GovRicketts**, **GovLarryHogan** and **GovParsonMO**. We caution that this network captures Granger causality (Xu, Farajtabar, and Zha 2016), and does not control for confounding effects. In Figure 5, we plot the estimated baseline rate of spontaneous tweets per governor and cabinet member, along with each individuals estimated influence (average number of subsequent tweets in the network directly triggered by a Tweet). Here we observe that President Trump has the highest rate of spontaneous tweets, followed by the Governor of Hawaii and Secretary Azar. Governors Ducey, Wolf and Lamont are the largest estimated influencers.

Risk, treatment and testing sub-topics

In addition to applying the HBTM to all COVID-19 related tweets, we also apply the model separately to three sub-categories. We first apply HBTM to tweets containing the word “risk”. A sequence of clusters are illustrated in the top row of Fig. 6. The emergence of this sub-category coincides with the start of the second phase of the general timeline, and it appears that the **CDC** was among the first to mention how **serious** the risk was and asked for **immediate** actions. However, the subsequent clusters in early March indicate that both state and federal governments (Republicans and Democrats) were telling the public that the risk **remains low**. Also in this period, we observe calls for **washing hands** to **reduce** risk, and that **seniors** were identified to be

the most vulnerable. After March 15, the narrative changes and the **high** risk to the general **population** is acknowledged. Keywords like **age** and **adult** indicate the high risk across **age** groups, even for young **adults**. The word **high** frequently co-occurs with **test** and **quarantine**; due to the **high** risk of transmission, state governments increased **testing** and enforced **quarantine(s)**. Overall, from left to right, the sequence of clusters show a clear trend in the narrative from low risk in late February to high risk in April.

Next, we apply HBTM to tweets containing the words “vaccine” and “treatment”. The resulting clusters are illustrated in the middle row of Fig. 6. In mid-March, keywords **launch**, **trial**, **clinicaltrial**, **phase**, and **candidate** indicate that vaccine **candidates** were identified and entered the **clinical trial phase**. We can also see the National Institute of Health (**NIH**) **partner** with the pharmaceutical industry in developing the vaccine. Later in March, we start to see clusters where state governors (mainly Democrats) commented on the lack of **resources**, **equipment**, **ventilators**, and **hospital beds**. We also see cabinet members (specifically Sec. of Health @SecAzar) giving updates about vaccine development (**genetic sequence** and **clinical trial**). Another narrative is around an agreement (**agree**) with insurance companies to **ease** the **burden** of the pandemic on their **customers**. Additionally, we see the request to **create global researcher team** in developing a vaccine. In general, the clusters here suggest that the search for a vaccine has been a collective effort that crosses political parties and national boundaries.

In the bottom row of Fig. 6, we show clusters found by applying HBTM after filtering the dataset on the keyword “test”. In early March, we see that **new test kits** were **available**. Tweets mention (**negative**) test results of some individuals by the Democrat governors and cabinet members. Con-

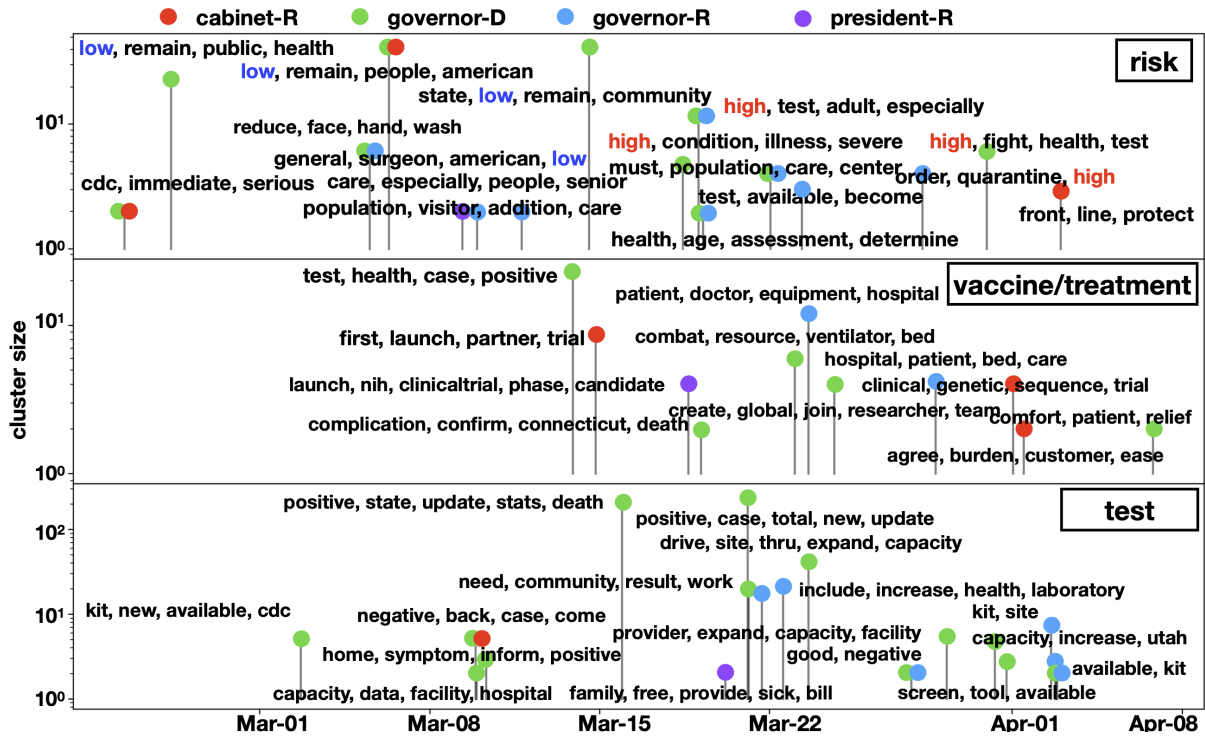


Figure 6: Timeline of sub-topics on risk, treatment and testing. Clusters with size at least 2 are pinned. Keywords indicate the topic of the clusters. The marker color indicates the dominant component of the cluster.

Table 1: Officials ranked by in-degree (most influenced) and out-degree (most influential) in influence networks.

Topic	In-degree	Out-degree
all	GovMurphy, GovRicketts, LouisianaGov	GovNedLamont, GovMurphy, GovMLG
risk	GovMikeDeWine, NYGovCuomo, GovMLG	GovMikeDeWine, GovPritzker, SecAzar
treatment	SecAzar, GovNedLamont, GovofCO	GovofCO, GovChrisSununu, GovNedLamont
test	GovNedLamont, GovMikeDeWine, LouisianaGov	NYGovCuomo, GovHerbert, GovKemp

cern about the **capacity** of testing **facilities** and **hospitals** is also discussed in early March. In mid-March, testing is expanded to the **community**, followed by requests for **expanding facility capacity** and **increasing laboratories**. During this period, state governors (especially Democrats, the two highest green markers in Fig. 6) start updating test results (in particular number of **positive cases**) and providing **stats** in their press conferences. The HBTM model identifies a cluster in which **drive thru site** is suggested as a way to **expand** testing **capacity**. In early April, we observe that the narrative has shifted away from a lack of testing resources; keywords indicate that **screen tools**, test **kits**, and test **sites** are available, and the testing **capacity** has **increased**.

In Figure 7, we plot Granger causality influence networks for the risk, treatment and testing sub-topics. Again we see connections crossing party lines. In the case of testing, the network is characterized by a dense set of connections between a select set of governors. The risk and treatment networks are characterized by more active nodes with fewer connections. In Table 1 we also list the most influential officials by sub-topic along with those officials most influenced.

Conclusion

We analyzed the COVID-19 Twitter narrative among U.S. governors and presidential cabinet members using a Hawkes binomial topic model. We observed several narratives between January 1st and early April 2020, including a shift in the assessment of risk from low to high, discussion of a lack of testing resources which later subsided, and sub-topics around the impact of COVID-19 on businesses, efforts to create treatments and a vaccine, and calls for social distancing and staying at home. We also constructed influence networks amongst government officials using Granger causality inferred from the network Hawkes process. President Trump stands out for spontaneity, yet appears to have little influence with respect to network cross-excitation. Polarization is not obvious in the Granger influence networks; we observe a high level of cross party event triggering and influence seems more geographically clustered and related to state size.

We see several potential directions for future work. Here

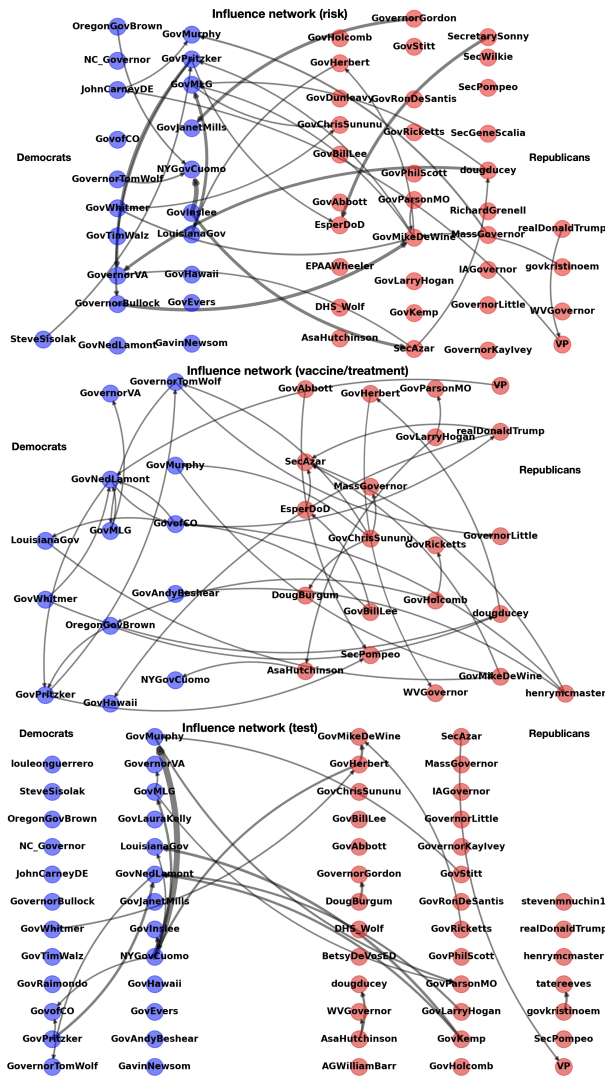


Figure 7: Granger causality influence network for “risk” (top), “treatment” (middle) and “test” (bottom) sub-topics.

we limited the analysis to only COVID-19 related tweets among U.S. government officials. The HBTM can be used to explore the COVID-19 narrative among the general population and may highlight issues around trust in institutions, adherence to social distancing, and economic impacts. Furthermore, analyzing non-COVID related tweets by government officials prior to the pandemic and constructing an evolving influence network may provide insights into how bi-partisan cooperation changes during national emergencies.

References

- Buntain, C.; McGrath, E.; and Behlendorf, B. 2018. Sampling social media: Supporting information retrieval from microblog data resellers with text, network, and spatial analysis. In *Proc. of the Hawaii Intl. Conf. on System Sciences*.
- Cinelli, M.; Quattrociocchi, W.; Galeazzi, A.; Valensise, C. M.; Brugnoli, E.; Schmidt, A. L.; Zola, P.; Zollo, F.; and Scala, A. 2020. The covid-19 social media infodemic. *arXiv preprint arXiv:2003.05004*.
- Du, N.; Farajtabar, M.; Ahmed, A.; Smola, A. J.; and Song, L. 2015. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 219–228. ACM.
- Dutta, H. S.; Dutta, V. R.; Adhikary, A.; and Chakraborty, T. 2020. Hawkeseye: Detecting fake retweeters using hawkes process and topic modeling. *IEEE Transactions on Information Forensics and Security*.
- Farajtabar, M.; Wang, Y.; Gomez-Rodriguez, M.; Li, S.; Zha, H.; and Song, L. 2017. Coevolve: A joint point process model for information diffusion and network evolution. *The Journal of Machine Learning Research* 18(1):1305–1353.
- Kim, M.; Paini, D.; and Jurdak, R. 2020. Real-world diffusion dynamics based on point process approaches: A review. *Artificial Intelligence Review* 53(1):321–350.
- Lai, E.; Moyer, D.; Yuan, B.; Fox, E.; Hunter, B.; Bertozzi, A. L.; and Brantingham, J. 2014. Topic time series analysis of microblogs. Technical report, DTIC Document.
- Mohler, G.; Buntain, C.; McGrath, E.; and LaFree, G. 2016. Hawkes binomial topic model with applications to coupled conflict-twitter data. *DOI: 10.13140/RG.2.2.13638.83527*.
- Simma, A., and Jordan, M. I. 2012. Modeling events with cascades of poisson processes. *arXiv preprint arXiv:1203.3516*.
- Tan, X.; Rao, V.; and Neville, J. 2018. The indian buffet hawkes process to model evolving latent influences. In *UAI*, 795–804.
- Thelwall, M., and Thelwall, S. 2020a. Covid-19 tweeting in english: Gender differences. *arXiv preprint arXiv:2003.11090*.
- Thelwall, M., and Thelwall, S. 2020b. Retweeting for covid-19: Consensus building, information sharing, dissent, and lockdown life. *arXiv preprint arXiv:2004.02793*.
- Xu, H., and Zha, H. 2017. A dirichlet mixture model of hawkes processes for event sequence clustering. In *Advances in Neural Info. Processing Systems*, 1354–1363.
- Xu, P.; Dredze, M.; and Broniatowski, D. A. 2020. The twitter social mobility index: Measuring social distancing practices from geolocated tweets. *arXiv preprint arXiv:2004.02397*.
- Xu, H.; Farajtabar, M.; and Zha, H. 2016. Learning granger causality for hawkes processes. In *International Conference on Machine Learning*, 1717–1726.
- Yin, F.; Lv, J.; Zhang, X.; Xia, X.; and Wu, J. 2020. Covid-19 information propagation dynamics in the chinese sina-microblog. *Math. Biosciences and Eng.* 17(3):2676.
- Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. Seismic: A self-exciting point process model for predicting tweet popularity. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1513–1522. ACM.