More Active Internet-Search on Google and Twitter Posting for COVID-19 Corresponds with Lower Infection Rate in the 50 U.S. States

Jiachen Sun
Sun Yat-sen University
Peter Gloor (pgloor@mit.edu)
MIT Center for Collective Intelligence

Research Article

Keywords: COVID-19, Google Trends, Twitter

Posted Date: July 9th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-40745/v1

Abstract

As the novel coronavirus disease 2019 (COVID-19) continues to rage worldwide, the United States has become the most affected country with more than 2.5 million total confirmed cases up to now (June 2, 2020). In this work, we investigate the predictive power of online social media and Internet search for the COVID-19 pandemic among 50 U.S. states. By collecting the state-level daily trends through both Twitter and Google Trends, we observe a high but state-different lag correlation with the number of daily confirmed cases. We further find that the predictive accuracy measured by the correlation coefficient is positively correlated to a state's demographic, air traffic volume and GDP development. Most importantly, we show that a state's early infection rate is negatively correlated with the lag to the previous peak in Internet search and tweeting about COVID-19, indicating that the earlier the collective awareness on Twitter/Google in a state, the lower is the infection rate.

Introduction

*"At every crucial moment, American officials were weeks or months behind the reality of the outbreak. Those delays likely cost tens of thousands of lives". NYT June 26, 2020 [*1]

Since the beginning of January 2020, the world has been turned upside down. Nothing is like it was before since the novel coronavirus disease was first reported in Wuhan, China, in December 2019 [2]. After initial blunders, China took energetic measures to combat the virus (e.g. the Wuhan shutdown) [3] while the Western world was still mostly complacent. Although epidemiologists have already warned at the end of January that COVID–19 would probably turn into a global crisis [4], politicians and the population in the US and Western Europe alike initially ignored the problem. The virus was seen as something far away, that like SARS and the avian flu would be active mostly in the high-density populations of Asia and then go away. And even when Italy was shaken with a virulent COVID–19 outbreak in February [5], which closed down the northern industrial heartland of Veneto, the US authorities were still mostly ignoring the problem [6]. Only when in mid-March New York started seeing soaring infection rates, did the population and the politicians start taking the disease seriously. This behavior is perfectly reflected in the Google search trend and the Twitter activity, motivating our research question: Is a state or political entity better capable of dealing with an infectious disease if the collective awareness is raised early on in the course of the disease? Does a population actively searching for information about COVID–19, and showing a robust dialog on Twitter about this topic deal more efficiently with the disease?

It has been illustrated that data from the online social media and Internet searches are correlated with several epidemics that have previously happened, such as seasonal influenza epidemics [7], Dengue [8], MERS [9] and H1N1 [10]. Regarding COVID–19, several works [11–14] have demonstrated significant correlation between the Internet search and the pandemic spreading among different countries. However, it still remains unclear what regional factors the Internet's predictive abilities may relate to, and whether they are useful surveilling the spread of the disease. If there is indeed predictive power in the Google search and tweeting behavior of a US polity such as a state or a city, it will give invaluable input to

policymakers, governments, and healthcare providers to better prepare and deal with potential future waves of the COVID–19 and other epidemics.

In this work, using the data from 50 states of the United States, we conduct a comparative study about the role of online social media and search trends in the COVID–19 epidemic. We show that the daily number of COVID–19 related tweets in Twitter exhibits a strong but *state-different* lag correlation with newly confirmed cases. The same can be observed on the Google Trends index using coronavirus-related search terms. These state-differences in predictive capabilities in terms of correlation strength and lag are closely related to a state's demographics, quantitively measured by a state's population size and density, air traffic volume and economic development. Further, our analysis on the state-level *early* COVID–19 incidents demonstrates a significantly negative correlation between the lag and the early infection rate, implying that an actively engaged population that searches for information and tweets about COVID–19 more ahead of the outbreak indicates a lower infection rate.

Results

- Predictive Power for Google Trend and Twitter

We focus on COVID–19 infections, Twitter tweets and Google search data for all 50 U.S. states, excluding Puerto Rico and the District of Columbia since some Internet data for these two regions are unavailable. Specifically, we collect the state-level daily COVID–19 confirmed cases from the New York Times. The number of COVID–19 related tweets in each individual state is extracted from an open COVID–19 Twitter chatter dataset [15]. We obtain the Google Trends index by using a combination of one of three keywords ('coronavirus', 'COVID' and 'COVID19') and a state's full name as an *integrated* search term (e.g. 'coronavirus Massachusetts', 'COVID California'), given that residents are usually more concerned about their local situation of the pandemic. More details of data used in this study are described in the Methods section.

In Fig. 1, we illustrate a comparison among the daily confirmed cases, number of COVID–19 related tweets and the Google Trends indexes (with different search terms) in New York, Massachusetts, Iowa and California. One can observe that the overall graph patterns are different between states. We then investigate the relationship between the COVID–19 pandemic spreading and the Internet data in all 50 U.S. states. Fig. 2 shows the lagged Spearman correlation between the Internet data from Twitter and Google Trends and the reported COVID–19 cases for the selected 4 states. To quantify the predictive power of the tweeting behavior and the search activity for an individual state, we denote c^* as the highest correlation coefficient and l^* as the optimal lag achieving c^*. In principle, a larger c^* indicates a higher accuracy in predicting the state-specify pandemic. A larger l^* corresponds to an earlier peak of Internet searches and tweeting about COVID–19, indicating that residents start being active on the Internet earlier. We find that c^* and l^* are quite different among different states and between Google Trends and Twitter (see Fig. 2). For instance, for New York, c^* is merely 0.60 with l^* = 15 using the Twitter data but is up to 0.95 with l^* = 19 for tracking 'coronavirus New York' on Google Trends. For

California, the $c*$ of Twitter and of 'coronavirus California' on Google Trends are 0.67 and 0.81, respectively, while the $l*$ for both is above 30 days.

Fig.3 presents the distribution of $c*$ and $l*$ for Twitter and Google Trends for all 50 U.S. states. The average of $c*$ from Twitter is 0.64, while for Google Trend using the keyword 'COVID' $c*$ is nearly 0.70. These results imply that the tweeting activity and search interest indeed have the capability to predict the COVID–19 spreading. On the other hand, the average $l*$ on Twitter is about 26 days, revealing a smaller delay of the Twitter platform. Indeed, we find that $c*$ and $l*$ on Twitter are significantly correlated with p<0.001 (see the correlation coefficient between $c*$ and $l*$ in Supplementary Table 1), meaning that earlier collective tweeting may result in more accurate prediction. For Google Trends, the average $l*$ of the keyword 'coronavirus' (27.0) is somehow larger than both 'COVID' (21.3) and 'COVID–19'(24.5). An explanation could be that the majority of people searched by the word 'coronavirus' since the pandemic initially was reported under this name, while the names 'COVID' and 'COVID–19' were formally proposed by the World Health Organization at the end of February 2020.

- Correlation between $\boldsymbol{c}*$ and state conditions

We find that the wide difference of $c*$ among the 50 states is partially related to a state's economic and social conditions. Specifically, we consider population demographics, air traffic flow and the economic development level, which can be quantitively characterized by the following proxies. A state's *population size* as of 2019 is estimated by the U.S. Census Bureau, along with the *population density* measured by number of residents per square mile. The air traffic flow is measured by *enplanement (*i.e., the number of passengers boarding) in 2017 and 2018 (see details in the Methods Section). Besides, we collect each state's gross domestic product (*GDP)* as well as the *GDP per capita* as of 2019 4th quarter to measure economic output.

We calculate the Spearman correlation coefficient between these six variables and the $c*$of Twitter volume and Google Trends index, finding a significantly positive correlation, as shown in Table 1. In particular,

the more people, the higher population density, the higher air traffic and wealth a state has, the more accurate the Twitter and Google Trends predict the COVID–19 pandemic. This makes intuitive sense, as higher income is correlated with higher education, and higher geographic mobility leads to a higher information exchange, both raising early awareness of the pandemic. There is no significant correlation between $l*$ and the states' demographic variables.

	c^*			
	Twitter	Google Trend (coronavirus)	Google Trend (COVID)	Google Trend (COVID-19)
Population size (2019)	0.505***	0.210	0.340*	0.573***
Population density (2019)	0.374**	0.302*	0.414**	0.473***
Enplanements (2018)	0.303*	0.355*	0.416**	0.609***
Enplanements (2017)	0.301*	0.360*	0.421**	0.610***
GDP (2019 Q4)	0.535***	0.229	0.374**	0.599***
GDP per capita (2019 Q4)	0.244	0.379**	0.517***	0.432**

Table 1. Correlation coefficient between c^* and states' variables in terms of population demographics, air traffic flow and the economic development level (N = 50). The significance level is denoted by stars in red: * p<0.05, ** p<0.01, *** p<0.001

- Correlation between early infected rate and $\boldsymbol{c*/l*}$

We further figure out the effect of an actively engaged population on the outbreak of the infection. Specifically, we focus on the *early* stage of the COVID−19 outbreak in the 50 U.S. states, a period when the government had not started yet to take serious control measures. The infection rate in this stage is a reasonable proxy to measure the extent to which a state's residents rely on their individual awareness to protect themselves again the pandemic. Quantitively, we define the *early* infection rate as the proportion of residents being infected in the earliest T days since the state-level first case was confirmed (see the distribution of the *early* infection rate among 50 states in the Supplementary Figure 2).

Having both the predictive capacity of Internet search and Twitter data and the early infection rate, we are able to find the relationship between the two. Surprisingly, we discover a strong negative correlation between l^* and the early infection rate, with T varying from 1 week to 3 weeks, as shown in Table 2. This relationship indicates that the earlier people start tweeting and searching, the lower is the infection rate. In other words, the earlier the collective awareness on Twitter and on Google search, the less people get infected when the virus outbreaks. Moreover, we also find a significantly negative correlation between c^* and the infection rate using the Twitter data and Google Trends for the terms 'COVID' and 'COVID- 19' on

selected $T's$ (see Table 2), implying that the more predictive pro-active Internet-search behavior is, the lower the initial infected rate.

Table 2. Correlation coefficient between early infection rate for different T (number of days) and l^* and c^* from Internet data (N = 50). Similar to Table 1, the red stars represent the significance level.

		Early infection rate		
		T=7	T=14	T=21
l^*	Twitter	-0.371**	-0.405**	-0.416**
	Google Trend (coronavirus)	-0.471***	-0.517***	-0.500***
	Google Trend (COVID)	-0.473***	-0.517***	-0.505***
	Google Trend (COVID-19)	-0.445**	-0.516***	-0.522***
c^*	Twitter	-0.566***	-0.593***	-0.476***
	Google Trend (coronavirus)	-0.139	-0.08	-0.100
	Google Trend (COVID)	-0.371**	-0.374**	-0.167
	Google Trend (COVID-19)	-0.543***	-0.510***	-0.270

Conclusion

In conclusion, this study showed that there is a high but state-different correlation between the results of Google search and tweeting about COVID−19 related keywords and the number of confirmed COVID−19 cases among 50 U.S. states. These significant correlations occur as early as 27 days before confirmation of the infections, indicating the usefulness of Internet search and online social media tracking to surveil the pandemic's outbreak locally. We further found that the differences in predictive power between these states are closely related to a state's demographics characterized by population size and density, air traffic and economic development. Most importantly, we discovered that if there is an actively tweeting population which leads a vibrant dialog on Twitter about COVID−19, the early infection rate will be lower. Similarly, the more ahead of the outbreak a population starts googling for COVID−19 information, the lower the early infection rate.

Methods

C - Cases in U.S. We collect the COVID–19 confirmed cases from the New York Times (https://www.nytimes.com/), based on reports from state and local health agencies. 50 U.S. states' daily number of cases are used in this study. For each state, the study period is from the date of the first confirmed case in this state to June 2, 2020.

CTwitter ata. The COVID–19 tweets on Twitter are acquired from an open COVID–19 Twitter chatter dataset [15], which is a collection of the identifiers of tweets specifically using coronavirus-related keywords (coronavirus, 2019nCoV, COVD19, CoronavirusPandemic, CoronaOutbreak, etc.), starting from January 27, 2020. After hydrating the full JSON objects from these tweets' identifiers, we extract the daily number of tweets in the U.S. at state level according to a tweet's location. Specifically, we first identify all geo-located tweets (i.e., tweet associated with a geographic place), only retaining tweets with a location in the US. Then we assign a tweet to a state using its specific location, such as city and town (see the heatmap of the number of available geo-located tweets in 50 U.S. states in the Supplementary Figure 1).

Google Trends and Keywords. As the most used search engine in U.S., Google Trends (https://www.google.com/trends) provides an excellent proxy for Internet-search trends. The Google Trends index measures the search activity of a term compared to the most actively searched keyword for a selected region. In this work, we use the pytrends API to track three respective keywords, 'coronavirus', 'COVID' and 'COVID19' on Google Trends among 50 U.S. states. For each individual state, we use a combination of a coronavirus-related keyword and the state's full name as an integrated search term throughout this paper. The time parameter is set to two weeks earlier than the COVID–19 outbreak date in each state.

Correlation analysis. The Spearman correlation is employed in this study using Python's SciPy function. Specifically, we conduct lagged correlation analyses to assess the temporal relationships between Internet data and COVID–19 pandemic. For each state, we right-shift the daily Internet data from Twitter and Google Trends (with different search terms) by a variable lag and calculate the Spearman correlation to the daily reported COVID–19 cases. The maximum lag is set to 40 days. Spearman correlation is also used to examine the correlation between the $c*$ and the state's variables, and between the $c*/l*$ and the early infection rate, at significance levels from *p<0.05 to ***p<0.001.

Proxy of air traffic flow. Using the Air Carrier Activity Information System database (*https://www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/passenger/collection/)*,, we obtain the enplanement data at every commercial service airport in U.S. for 2017 and 2018. As a proxy of a state's air traffic flow we calculate the sum of the enplanements of all airports located in a state.

Declarations

Code availability

Sourcecodesandallprocesseddatausedinthisstudyisavailableat
https://www.researchgate.net/profile/Jiachen_Sun2/publications

Author Contribution

P. G. conceived the project. J. S. designed the experiments and analyzed the results. J. S. and P. G. wrote the manuscript.

Additional information

Competing interests. The authors declare no competing interests.

References

1. Watkins, , Holder, J., Glans J., Cai, W., Carey, B. and White J., How the virus won. *New York Times*. https://www.nytimes.com/interactive/2020/us/coronavirus-spread.html
2. Zhu, , Zhang, D., Wang, W., i, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., u, R. and Niu, P., 2020. A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*.
3. Tian, H., iu, , i, Y., Wu, C.H., Chen, B., Kraemer, M.U., i, B., Cai, J., Xu, B., Yang, Q. and Wang, B., 2020. An investigation of transmission control measures during the first 50 days of the COVID-19 epidemic in China. *Science*, 368(6491), pp.638-642.
4. i, , Pei, S., Chen, B., Song, Y., Zhang, T., Yang, W. and Shaman, J., 2020. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV-2). *Science*, 368(6490), pp.489-493.
5. Remuzzi, A. and Remuzzi, G., 2020. COVID-19 and Italy: what next?. *The Lancet*.
6. ipton, E., Sanger, D., Haberman, M., Shear, D.M., Mazzetti, M. and Branes, E.J., He could have seen what was coming: behind Trump's failure on the virus. *New York Times*. https://www.nytimes.com/2020/04/11/us/politics/coronavirus-trump-response.html
7. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, ., Smolinski, M.S. and Brilliant, ., Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), pp.1012-1014.
8. de Almeida Marques-Toledo, , Degener, C.M., Vinhal, ., Coelho, G., Meira, W., Codeço, C.T. and Teixeira, M.M., 2017. Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. *PLoS neglected tropical diseases*, 11(7), p.e0005729.
9. Shin, Y., Seo, D.W., An, J., Kwak, H., Kim, S.H., Gwack, J. and Jo, M.W., 2016. High correlation of Middle East respiratory syndrome spread with Google search and Twitter trends in Korea. *Scientific reports*, 6, p.32920.
10. Wilson, , Mason, K., Tobias, M., Peacey, M., Huang, Q.S. and Baker, M., 2009. Interpreting "Google Flu Trends" data for pandemic H1N1 influenza: the New Zealand experience. *Eurosurveillance*, 14(44),

p.19386.

11. Effenberger, M., Kronbichler, A., Shin, J.I., Mayer, G., Tilg, and Perco, P., 2020. Association of the COVID-19 pandemic with internet search volumes: a google trendstm analysis. *International Journal of Infectious Diseases.*
12. i, C., Chen, .J., Chen, X., Zhang, M., Pang, C.P. and Chen, H., 2020. Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. *Eurosurveillance*, 25(10), p.2000199.
13. in, H., iu, C.H. and Chiu, Y.C., 2020. Google searches for the keywords of “wash hands” predict the speed of national spread of COVID-19 outbreak among 21 countries. *Brain, Behavior, and Immunity*.
14. Walker, , Hopkins, C. and Surda, P., 2020, April. The use of google trends to investigate the loss of smell related searches during COVID-19 outbreak. *In International Forum of Allergy & Rhinology*.
15. Banda, Juan M., Tekumalla, Ramya, Wang, Guanyu, Yu, Jingyuan, iu, Tuo, Ding, Yuning, Artemova, Katya, Tutubalina, Elena and Chowell, Gerardo. 2020. A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration. doi: 10.5281/zenodo.3723939. http://www.panacealab.org/covid19/

Figures

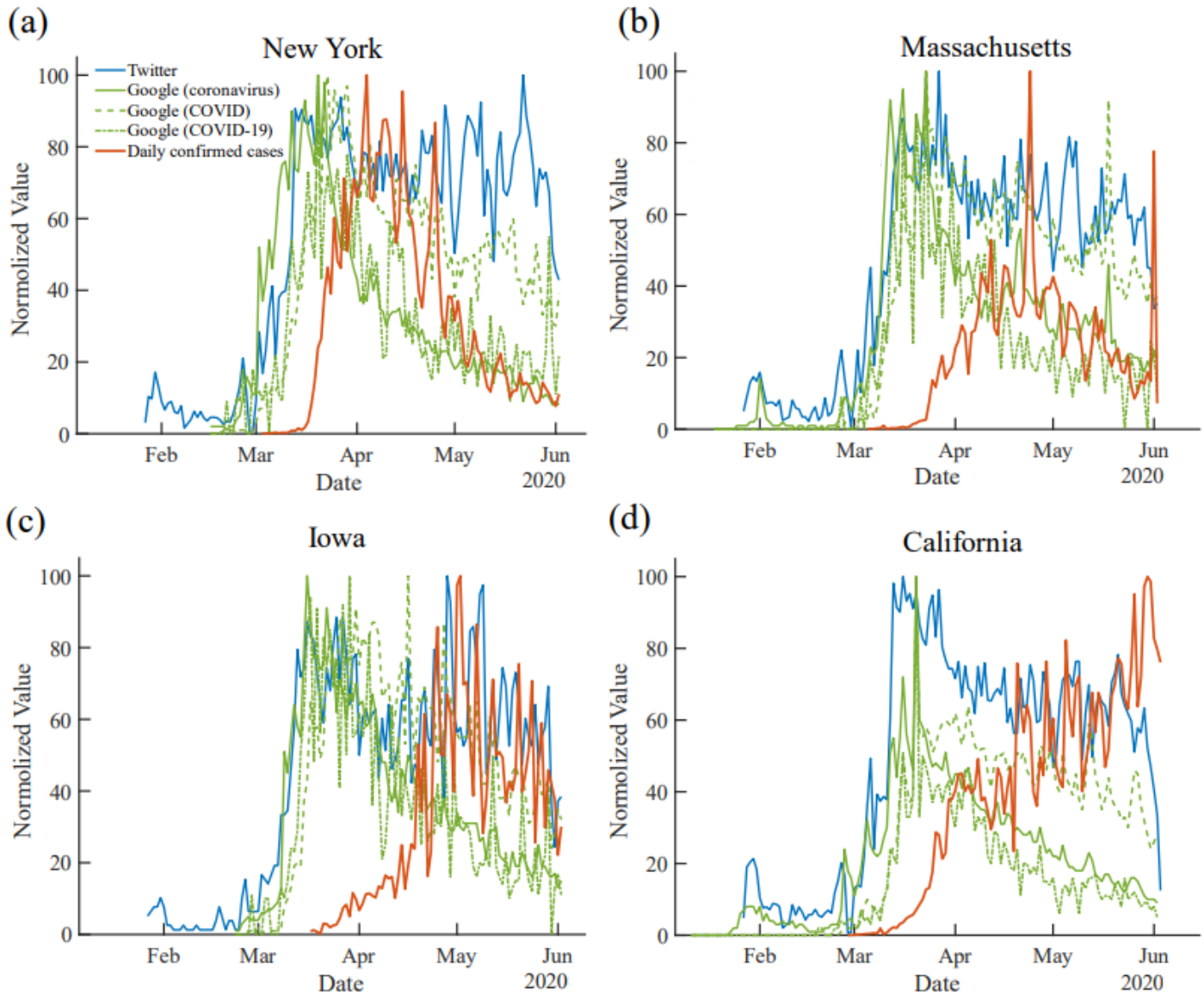

Figure 1

Number of COVID-19 related tweets, Google Trends index using different COVID-19 keywords (integrated with the state's full name) and daily infected number in 4 states. The values of each curve are normalized to [0, 100] for comparison.

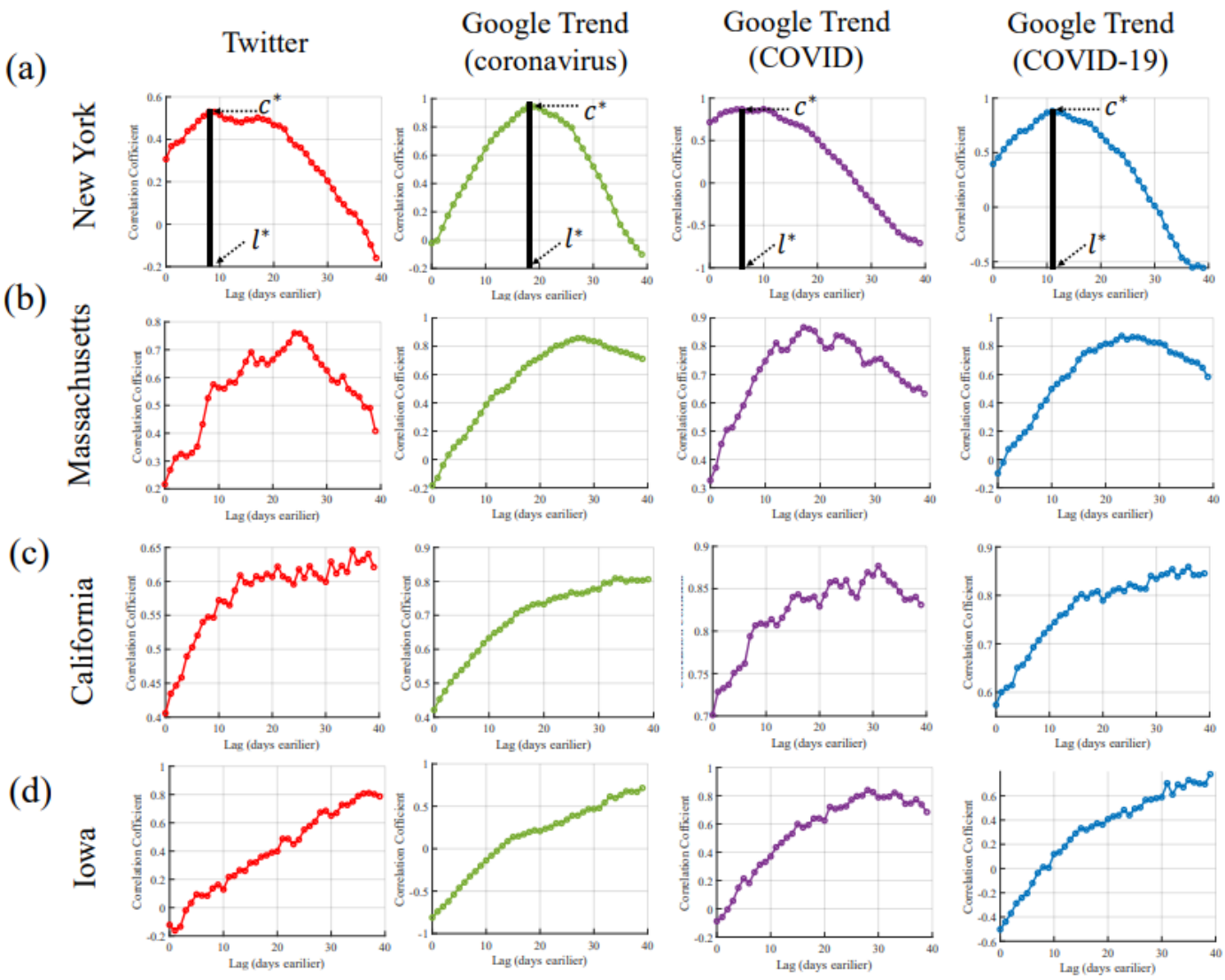

Figure 2

Illustration of lagged correlation between new confirmed COVID-19 infections and data from Google Trends and Twitter in selected 4 states.

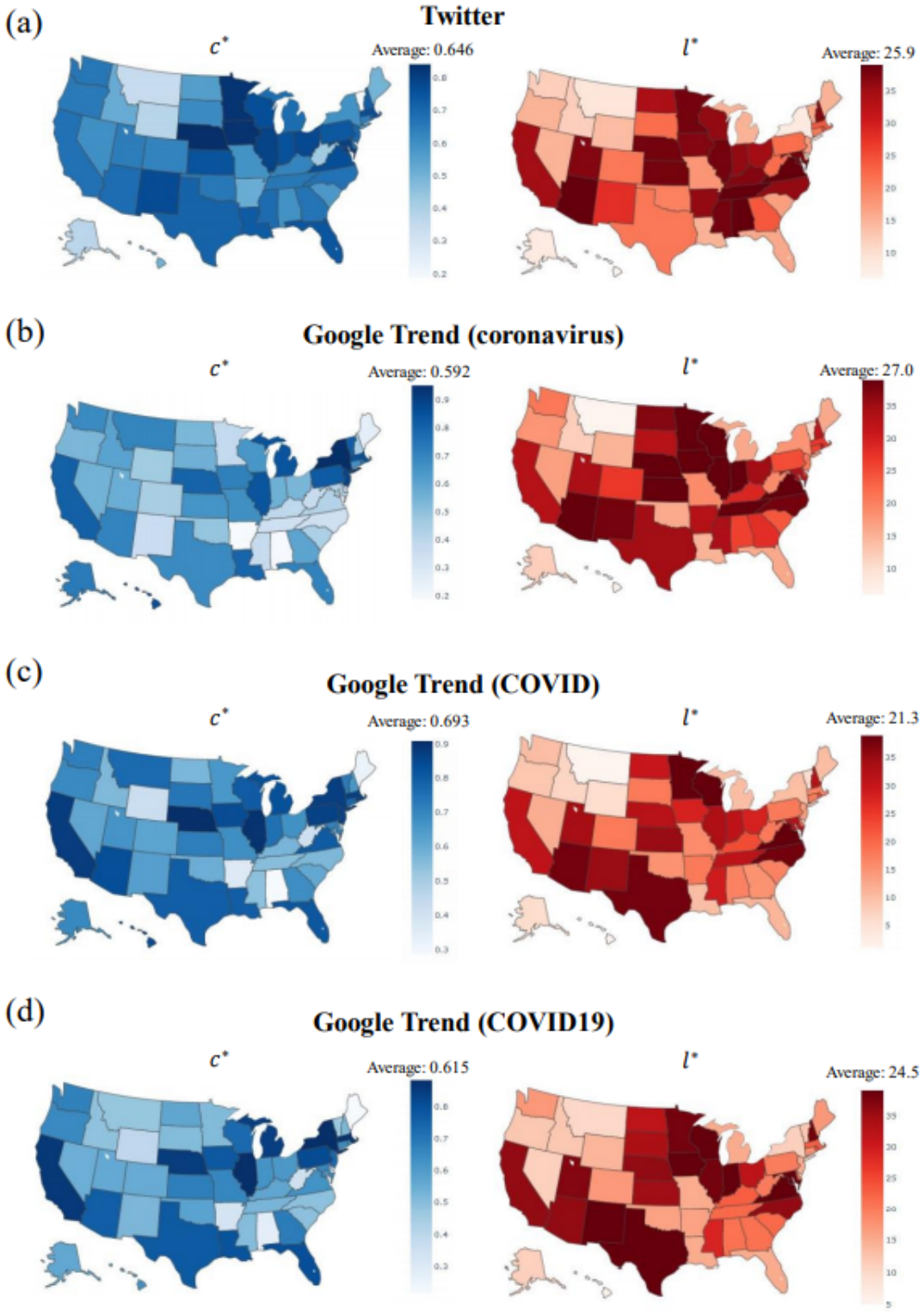

Figure 3

Distribution of $c*$ and $l*$ over 50 states for (a) Twitter and (b-d) Google Trends with different keywords.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- COVID19PredictionSI.pdf