Trauma Propagation in Social Networks

D, Sadish

7 March 2021

Trauma Propagation in Social Networks

Sadish[•]

March 7, 2021

I present evidence that rises in the severity of the Covid-19 pandemic in one location increased internet searches indicative of mental disorders in another location that is spatially separated but socially connected.

Keywords: Mental Health, Social Networks, Covid-19

[•](sadishdhakal@gmail.com) I have no source of funding or conflict of interest to declare concerning this paper. All errors remaining in the paper are my own.

Widespread mental illnesses accompany crises such as terrorist attacks [1, 2], natural disasters [3], and pandemics [4, 5]. Can the trauma of such crisis propagate through social connections that transcend physical and political borders?

I combine data on global social networks from a social media platform, on the volume of internet searches related to health symptoms in the US, and on global Covid-19 casualties. Using these data, I find evidence that rises in the severity of the Covid-19 pandemic in one location increased internet searches indicative of mental disorders in another location that is spatially separated but socially connected.

1 Data and Methods

I combine data from several sources.

Social Networks

Facebook released the first version of the Social Connectedness Index in 2018 [6]. The index is the ratio of the number of Facebook connections between residents of two locations and the number of possible connections between Facebook users in these locations, scaled to have a minimum value of 1 and maximum value of 1 billion [7]. It can be interpreted as the probability of there existing a social connection between individuals residing in two locations. In the US and Europe, the index is higher between two locations that share the same language, religion, history, and demographics such as age and education [6, 8].

In late 2020, Facebook released datasets in varied geographical resolutions [7]. I use two datasets: the first defines the index for pairs of US counties, and the second for US counties paired with countries. The index is not comparable across datasets as the scaling is unique to the dataset.

Given that over 36% of the world's population— 2.8 billion individuals— use Facebook every month [9], the data provide a measure of the global social network with high degree of representativeness.

Symptoms

Researchers have leveraged data from internet search engines and social media for

public health surveillance of infectious diseases [10, 11], influenza in particular. Google's data on searches have a high degree of representativeness because they reflect a large portion of internet searches. In 2020, searches on Google's platform accounted for 88% of the internet searches in the US and 90% worldwide [12].

In 2020, Google released data on internet searches related to over 400 health symptoms for six countries at varying geographic resolutions [13]. The data spanned over three years and were available at day or week levels of aggregation. To calculate the index, Google divide the total number of searches ascribed to a given symptom in a given geographic unit in a given time period by the total number of searches in that unit in that time period. The ratio was then scaled so that the highest value for that period in that geographical unit was 100. Given this scaling, the index can not be compared across geographical units or across time periods without adjustment. Searches by a single user on a given day were only counted once per symptom for up to three symptoms. Google added artificial noise to the data to preserve the privacy of individuals, and did not report data when there were too few searches for a particular symptom in a given location at a given time. This censoring means data are not missing at random. Missing data represent the lowest search volumes.

I use symptoms search data representing US counties in 2020. Given the high geographical resolution of US counties, weekly aggregation makes it more feasible to preserve privacy with minimal artificial noise and censorsing [13]. For this reason, data aggregated at the weekly level have fewer missing values. I use the data aggregated at the week level.

Traumatic Shocks

Researchers affiliated with the Covid-19 Open Data project [14] have curated data on Covid-19 cases and deaths from inter-governmental organizations [15, 16], government agencies, and other sources [17, 18] [1]. I use the number of Covid-19 cases and deaths as sources of psychological shocks.

Policy, Geography, and Weather

The Covid-19 Open Data project [14] has also compiled Oxford University's index of the stringency of government response to the Covid-19 pandemic [19], geographic

---

[1]full list of sources of casualty data [retrieved 2021-03-03]:
https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-epidemiology.md

coordinates, weather data [20], and population statistics [21], all of which I use [2].

1.1 Data Processing and Validation

Certain characteristics of the symptoms data require me to make a series of decisions to ensure that a sound empirical strategy can be applied to them.

Choice of Outcome Symptoms

With over 400 symptoms to choose from, a researcher has undue latitude to retroactively justify associations between shocks and a set of symptoms. A researcher who is given a random variable and asked to determine if it has a statistically significant association with 400 random variables will find 20 statistically significant associations if they are willing to accept a 5% probability of falsely rejecting the hypothesis of no association.

Exacerbating the problem, a wide array of symptoms accompany mental illnesses. Headaches, dizziness, nausea, tinnitus, perspiration, tremors, as well as gastrointestinal symptoms such as constipation, stomach ache, and diarrhea are all associated with stress and mental disorders [22, 23]. The presence or absence of an effect on these symptoms would be inclusive with regards to an effect on mental health.

I must identify a set of symptoms to serve as outcomes and interpret the effects of traumatic shocks on them collectively.

Censored Data

My empirical strategy involves fixed effects with high dimensionality (several thousand fixed effects) as well as clustering of standard errors along two dimensions. Some of the most reliable software available for such analysis use least squares approaches. However, estimates from least squares regressions are sensitive to extreme values. The problem is exacerbated because the lowest values of symptoms are censored.

I must take measures to nullify the influence of censorship on my analysis.

Choice of Comparator Symptoms

---

[2]full list of sources for geography and population data [retrieved 2021-03-03]: https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-demographics.md

Despite challenges, the large number of symptoms come with the opportunity to attempt falsification tests. If the structure of social networks were at play, the severity of the Covid-19 pandemic would affect Coid-19 symptoms in ways consistent with the functioning of social networks. Using the same data on social connectedness as used here, researchers have found that Covid-19 was more likely to spread between locations with stronger connections [24]. There is also more travel between locations which share a high social connectedness index [8]. Albeit imperfect, the comparison with Covid-19, which transmits in close physical proximity, can be informative.

I must identify a set of Covid-19 symptoms to use as comparators.

Data Processing

To overcome these challenges in research design, I use the following procedure:

I discard all symptoms with data missing for more than 33% of observations. I am left with about 20% of variables. With fewer variables, I have less latitude in the selection of outcomes. Since missing data represent the lowest, censored values, the threat of invalidity due to censorship is curtailed in analyzing the remaining variables. I also restrict all regression samples to counties for which no more than 33% of the outcome symptoms have missing values.

I discard the highest 0.2% of observations to ensure that they do not have undue influence on the mean, particularly given that the lowest values are censored. This process further ensures that least squares regression remains a valid approach to estimation.

Discarding observations at the extrema of the distribution makes the results of my analysis less generalizable. However, the cost of loss in generalizibility is minimal given that the data are still representative of a large proportion of the population.

The magnitude of the effects are difficult to interpret on their own. I standardize each symptom to have mean zero and standard deviation one. Standardizing allows us to compare variations in one symptom with that in another, and also to compare the effects of traumatic shocks on them.

I then select mental disorders defined as such by the American Psychiatric Association in the Diagnostic and Statistical Manual of Mental Disorders [22]. These disorders are: alcoholism, attention deficit hyperactive disorder (ADHD), major depressive disorder (MDD), panic attacks, and insomnia. I exclude erectile dysfunction

as the disorder is specific to males. Suicidal ideation and self-harm are excluded as a result of the earlier procedure because data on these symptoms are missing for over a third of observations. I include anxiety, depression, psychosis, and sleep disorder which map to a broader range of disorders in the manual. Though the World Health Organization's Classification of Mental and Behavioral Disorders uses a slightly different taxonomy, the selected symptoms align with its definitions of mental disorders as well [23].

Next, I shortlist the following Covid-19 symptoms for comparison with mental disorders: common cold, cough, fever, infection, pneumonia, and sorethroat.

Validation

Figure 1 shows that the symptoms of Covid-19 are strongly correlated with each other. Mental disorders are also correlated, albite to a lesser degree. Correlations between mental disorders and Covid-19 symptoms are minimal.

I perform Principal Component Analysis on all of the selected symptoms. The first two principal components explain 63% of the variation, and the first three explain 69% . Loadings on the first two principal components in figure 2 distinguish between mental health disorders and Covid-19 symptoms. Loadings on the third principal component distinguish alcoholism, insomnia, and sleep disorders from other mental disorders.
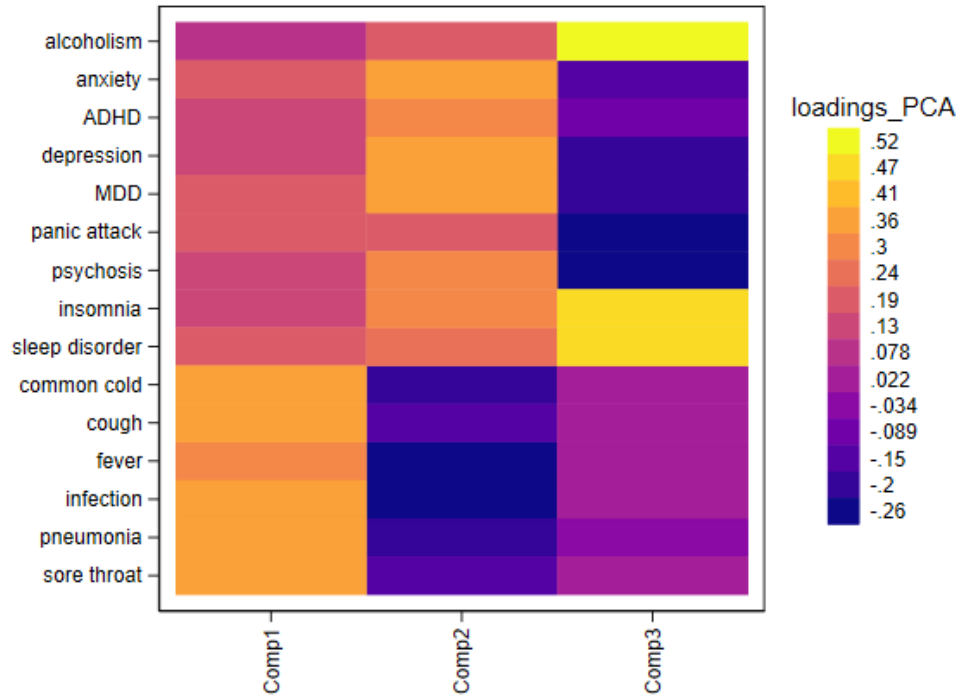
Figure 1: Correlation Between Symptoms
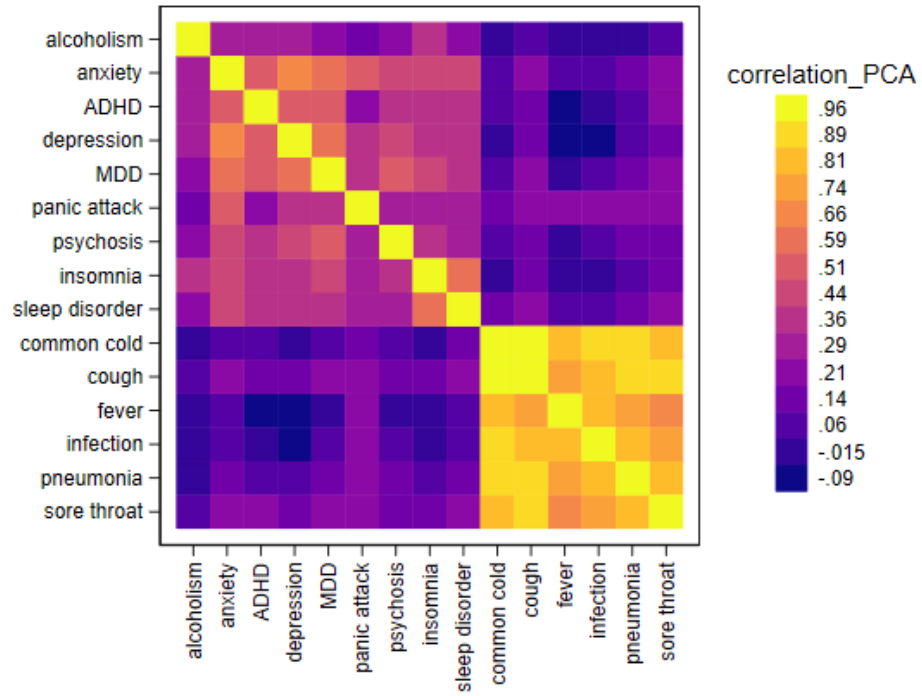
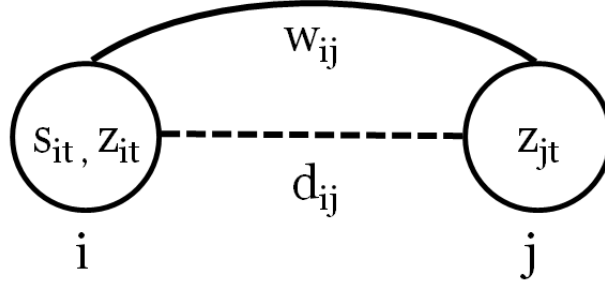The figure shows correlations between symptoms.

Figure 2: Principal Component Loadings of Symptoms

The figure shows the loadings for the first three principal components.

1.2 Model

Figure 3: Model of a Social Network

The figure depicts two nodes as circles. Variables pertaining to each node are written inside their respective circles. The nodes have a social link (solid line) and a spatial link (dotted line), each with its own weight.

I model the propagation of a traumatic shock in a network as follows and estimate it using least squares regression:

$$s_{it} = \alpha + \psi z_{it} + \phi \sum_j w_{ij} z_{jt} + \delta \sum_j d_{ij} z_{jt} + \boldsymbol{x}\boldsymbol{\beta} + \iota_i + \tau_t + \epsilon_{it}$$

where,

i and j represent nodes in the network. i is a US county that is the destination of the shock. The shock originates in j, which is either a US county or a country.

t represents time in weeks.

s_{it} are the symptom observed at time t for nodes i.

z_{it} and z_{jt} are the shocks observed at time t for nodes i and j respectively. I define

shocks as the natural log of weekly Covid-19 casualties— either cases or deaths—
per million inhabitants.

w_{ij} is the weight of the edges ij in the social network. I define it as the natural
log of the social connectedness index for the edges ij. $w_{ij} = 0$ for all $i = j$. The
construction of the social connectedness index uses a different scaling for pairs of US
counties than for pairs of US counties and countries. Weights in the two networks
are not comparable.

d_{ij} is the weight of the edges ij in the spatial network. I define it as the natural log
of the distance between i and j. $d_{ij} = 0$ for all $i = j$.

$\boldsymbol{x}$ is a $1 \times m$ vector of m characteristics of node i that vary over time.

ι_i is the unobserved error fixed for node i.

τ_t is the unobserved error fixed for time t.

ϵ_{it} is the unobserved error for node i at time t.

α, ψ, ϕ, and δ are parameters to be estimated; as are the elements of the $m \times 1$
vector $\boldsymbol{\beta}$.

The parameter ψ is the effect of a shock that occurs locally at i. ϕ represent the rate
of shock propagation in the social network and is the parameter of interest. Omitting
local shocks z_{it} from the model would not bias the estimates of ϕ if shocks correlated
only with z_{it} and not with $\sum_j w_{ij} z_{jt}$. Given that I use Covid-19 casualties as shocks,
omitting z_{it} could bias the estimates. Researchers have shown that Covid-19 was
more likely to spread between locations with stronger social ties as measured by the
same social connectedness index used here [24].

I include $\sum_j d_{ij} z_{jt}$ in the model and estimate δ, the rate of shock propagation in
the spatial network. Political borders are artificially defined and do not necessarily
coincide with natural geographical boundaries. So, nodes closer to each other can
have correlated symptoms. Also, w_{ij} is inversely corerlated with d_{ij} [6]. Omitting
$\sum_j d_{ij} z_{jt}$ from the model would bias estimates of ϕ, conflating propagation in the
spatial network with that in the social.

The vector $\boldsymbol{x}$ includes the average local temperature at i in time t, which is correlated
with seasonal symptoms. It also includes the stringency of government policy at i

in time t standardized to have mean zero and standard deviation one. It includes either the natural log of weekly Covid-19 case per million inhabitants, or the natural log of weekly Covid-19 deaths per million inhabitants, whichever is not already used as z_{it} in a given regression. When not included to account for confounding, these variables are in the model to improve the precision of estimates.

Since the index of symptom searches use a unique scaling for each geographical unit, they cannot be compared without adjustment. With destination node fixed effects ι_i included, we can interpret changes in symptoms as deviations from its mean within i. These fixed effects also account for unobserved characteristics of node i such as geography and demographic composition. Consider i's degree of social and spatial isolation. Low values of $\sum_j w_{ij}$ mean higher social isolation and high values of $\sum_j d_{ij}$ mean higher spatial isolation. Isolated counties experience shocks in their network with a diminished intensity. In other words, $\sum_j w_{ij}$ is correlated with $\sum_j w_{ij}z_{jt}$ and $\sum_j d_{ij}$ with $\sum_j d_{ij}z_{jt}$. The degree of social and spatial isolation may correlate with symptoms of mental disorders and Covid-19, too. If not accounted for, we would conflate the effects of isolation with that of $\sum_j w_{ij}z_{jt}$. As $\sum_j w_{ij}$ and $\sum_j d_{ij}$ are constant within i, destination node fixed effects ι_i account for social and spatial isolation.

Weekly Covid-19 casualties rose and fell in waves globally. At any given time, casualties in two different locations were likely correlated. So, symptoms s_{it} may correlate with $\sum_j z_{jt}$, which is the severity of the Covid-19 pandemic in i's entire network at a given time t. As $\sum_j z_{jt}$ correlates with $\sum_j w_{ij}z_{jt}$ (and $\sum_j d_{ij}z_{jt}$), failing to account for $\sum_j z_{jt}$ would lead us to conflate its effects with that of $\sum_j w_{ij}z_{jt}$. $\sum_j z_{jt}$ is constant within each week t. So, accounting for time fixed effects τ_t ensures that the rate of shock propagation in the social network ϕ is not biased by the global intensity of the pandemic or other unobserved global events.

The model readily appears in the literature of peer effects and social networks [25–27], but deviates in that it excludes the term $\sum_j w_{ij}s_{jt}$. The parameter of this term embodies how i's symptoms are associated with that of its network, and may not have a causal interpretation. Such indirect, peer effects are not always identifiable [25]. Since this identification problem first received attention, researchers have recognized several circumstances under which identification of peer effects is possible [27].

My objective is to estimate the *total effect* [28] of $\sum_j w_{ij}z_{jt}$ on s_{it}, not to separate the *direct effect* from the *indirect effect* mediated by $\sum_j w_{ij}s_{jt}$. The model I have proposed is agnostic to the composition of effects. The total effect is identified when

11

the peer effect $\sum_j w_{ij} s_{jt}$ is omitted from the model [28]. Introducing the term in the model is akin to the fallacy of conditioning on a post-treatment variable [29] or on intermediate outcomes [30].

I limit the scope of the model to contemporaneous propagation of shocks. I do not model the diffusion of shocks over time. In the Appendix, I show how the model has roots in a simpler linear model with interactions.

2 Results

Figures 4 presents estimates of the rate of shock propagation ϕ in the local and global social networks. The figure shows that an increase in casualties in the local social network led to higher internet searches indicative of mental disorders except for insomnia and sleep disorders. These shocks also led to increases in search for certain Covid-19 symptoms. Shocks in the global social network also led to increases in searches indicative of mental disorders, including insomnia and sleep disorders. These shocks had no effect on most symptoms of Covid-19, and negative effects on some.

These results provide evidence that the traumatic shocks of Covid-19 casualties propagated in both local and global social networks. Principal Component Analysis earlier had identified insomnia and sleep disorders as distinct from other mental disorders. The contrast in the effect of local and global shocks on these variables tells us that shocks in each propagate differently. The differential effects of local and global shocks on Covid-19 symptoms are also consistent with how Covid-19 spreads. International travel and mobility is much more restricted than local travel. The structure of the local social network is likely more predictive of physical contact than the global one.

Figure 4: Rates of Propagation of Traumatic Shocks in Social Networks
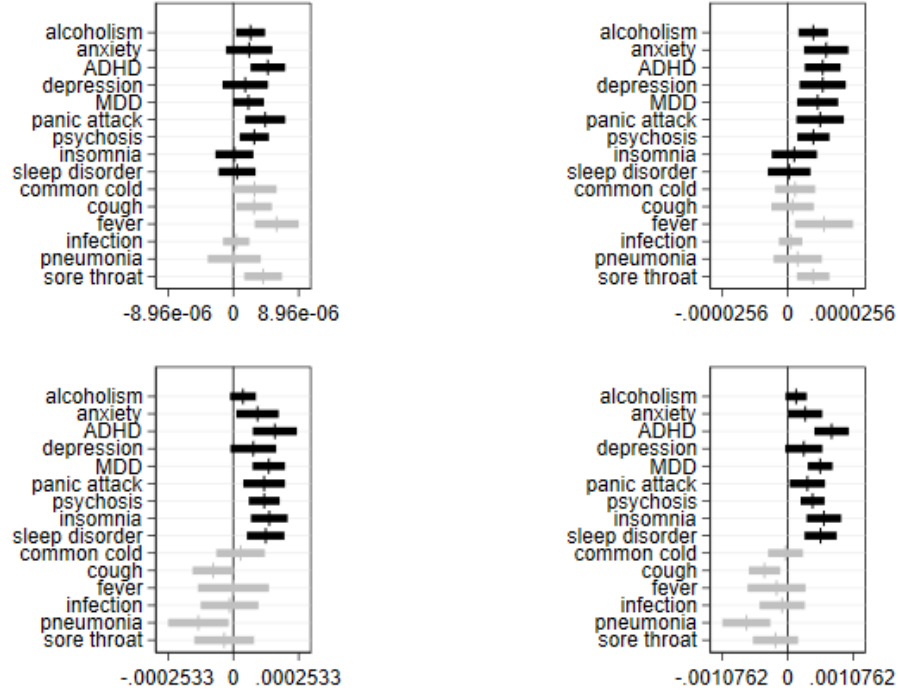
The figure shows coefficients and 95% confidence intervals for estimated rates of shock propagation ϕ in the local and global social networks. The y-axis lists outcomes. Each point on the y-axis represents one regression. Estimates and confidence intervals are to be read against the x-axis. Estimates for mental disorders are represented with black and comparator symptoms with grey. The **top** panels depicts the local social network and the **bottom** panels depicts the global social network. The **left** panels use the natural log of cases per million residents as shocks z_{it} and the **right** panels use log of deaths per million residents. All regressions include the two measures of casualty as independent variables, either as z_{it} or as elements of $\boldsymbol{x}$. All regressions also include two other independent variables: destination node i's local average temperature at time t, and the standardized intensity of government policy at i in time t. Sample for each regression includes only destination nodes (US counties) i which have at most 33% of missing values for the dependent variable; each regression might use a different sample. Standard errors are clustered at the destination node i and time t levels.

References

1. Holman, E. A., Garfin, D. R. & Silver, R. C. Media's Role in Broadcasting Acute Stress Following the Boston Marathon Bombings. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 93–98 (2013).
2. Thompson, R. R., Jones, N. M., Holman, E. A. & Silver, R. C. Media Exposure to Mass Violence Events Can Fuel a Cycle of Distress. *Science Advances* **5,** 1–6 (2019).
3. Thompson, R. R., Holman, E. A. & Silver, R. C. Media Coverage, Forecasted Posttraumatic Stress Symptoms, and Psychological Responses Before and After an Approaching Hurricane. *JAMA Network Open* **2,** 1–12 (2019).
4. Holman, E. A., Thompson, R. R., Garfin, D. R. & Silver, R. C. The Unfolding COVID-19 Pandemic: A Probability-Based, Nationally Representative Study of Mental Health in the United States. *Science Advances* **6,** 1–7 (2020).
5. Thompson, R. R., Garfin, D. R., Holman, E. A. & Silver, R. C. Distress, Worry, and Functioning Following a Global Health Crisis: A National Study of Americans' Responses to Ebola. *Clinical Psychological Science* **5,** 513–521 (2017).
6. Bailey, M., Cao, R., Kuchler, T., Stroebel, J. & Wong, A. Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives* **32,** 259–280 (2018).
7. Facebook. *Facebook Social Connectedness Index* `https://data.humdata.org/dataset/social-connectedness-index` (2021).
8. Bailey, M. *et al.* The Determinants of Social Connectedness in Europe (2020).
9. Facebook. *Facebook Reports Fourth Quarter and Full Year 2020 Results* `https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx` (2021).
10. Ginsberg, J. *et al.* Detecting Influenza Epidemics Using Search Engine Query Data. *Nature* **457,** 1012–1015 (2009).
11. Aiello, A. E., Renson, A. & Zivich, P. N. Social Media– and Internet-Based Disease Surveillance for Public Health. *Annual Review of Public Health* **41,** 101–118 (2020).
12. Statcounter. *Search Engine Market Share United States Of America* `https://gs.statcounter.com/search-engine-market-share/` (2021).
13. Google. *COVID-19 Symptoms Search Trends* `https://pair-code.github.io/covid19_symptom_dataset/?country=GB` (2021).
14. Wahltinez, O. *et al. COVID-19 Open-Data* `https://goo.gle/covid-19-open-data` (2021).

15. European Centre for Disease Prevention and Control. *COVID-19 datasets* `https://www.ecdc.europa.eu/en/covid-19/data` (2021).

16. World Health Organization. *WHO Coronavirus Disease (COVID-19) Dashboard* `https://covid19.who.int/` (2021).

17. Roser, M., Ritchie, H., Ortiz-Ospina, E. & Hasell, J. *Coronavirus Pandemic (COVID-19)* `https://ourworldindata.org/coronavirus` (2021).

18. The New York Times. *Coronavirus (Covid-19) Data in the United States* `https://github.com/nytimes/covid-19-data` (2021).

19. Hale, T. *et al. Covid-19 Government Response Tracker* `https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker` (2021).

20. National Centers for Environmental Information. *Data Access* `https://www.ncdc.noaa.gov/data-access` (2021).

21. https://datacommons.org/datasets. *Data Sources* `https://datacommons.org/datasets` (2021).

22. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders* 2013.

23. World Health Organization. *The ICD-10 Classification of Mental and Behavioural Disorders* 1990.

24. Kuchler, T., Russel, D. & Stroebel, J. The Geographic Spread of COVID-19 Correlates with the Structure of Social Networks as Measured by Facebook (2020).

25. Manski, C. F. Identification of Endogenous Social Effects: The Reflection Problem. *Review of Economic Studies* **60,** 531–542 (1993).

26. Jackson, M. O. Social and Economic Networks. *Princeton University Press* (2008).

27. Bramoullé, Y., Djebbari, H. & Fortin, B. Identification of Peer Effects through Social Networks. *Journal of Econometrics* **150,** 41–55 (2009).

28. Pearl, J. Causality: Models, Reasoning, and Inference. *Cambridge University Press* (2009).

29. Montgomery, J. M., Nyhan, B. & Torres, M. How Conditioning on Posttreatment Variables Can Ruin Your Experiment and What to Do about It. *American Journal of Political Science,* 1–16 (2018).

30. Schisterman, E. F., Cole, S. R. & Platt, R. W. Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. *Epidemiology* **20,** 488–495 (2009).

Appendix

Deriving a Model of Shock Propagation in a Social Network

Suppose there is a node i. I observe the symptom of i at time t as s_{it}. At time t, the node experiences a shock z_{it}. I model the effect of the shock on s_{it} as:

$$s_{it} = \alpha + \psi z_{it} + \boldsymbol{x}\boldsymbol{\beta} + \epsilon_{it}$$

where $\boldsymbol{x}$ is a $1 \times m$ vector of m characteristics of node i. The vector $\boldsymbol{\beta}$ has m elements, each a parameter for an element of $\boldsymbol{x}$.

Suppose there is a node j that has a social connection with i. The strength of the social connection is given by w_{ij}. A shock at j could propagate to i. So I update the above equation to:

$$s_{it} = \alpha + \psi z_{it} + \phi w_{ij} z_{jt} + \mu w_{ij} + \nu z_{jt} + \boldsymbol{x}\boldsymbol{\beta} + \epsilon_{it}$$

where ϕ is the rate at which the shock propagates from j to i while μ and ν represent independent effect w_{ij} and z_{jt} might respectively have on s_{it}. I am interested in obtaining an unbiased estimate of ϕ. However, the spatial distance between i and j might be correlated with the w_{ij} as well as with s_{it}, biasing my estimate of ϕ. So, I revise the model as:

$$s_{it} = \alpha + \psi z_{it} + \phi w_{ij} z_{jt} + \delta d_{ij} z_{jt} + \mu w_{ij} + \gamma d_{ij} + \nu z_{jt} + \boldsymbol{x}\boldsymbol{\beta} + \epsilon_{it}$$

where d_{ij} is the spatial distance between i and j, and γ represents the effect distance might have on s_{ij}.

Suppose now that i is in a network with more than two nodes indexed by j. We can revise the model as:

$$s_{it} = \alpha + \psi z_{it} + \sum_j \phi_j w_{ij} z_{jt} + \sum_j \delta_j d_{ij} z_{jt} + \sum_j \mu_j w_{ij} + \sum_j \gamma_j d_{ij} + \sum_j \nu_j z_{jt} + \boldsymbol{x}\boldsymbol{\beta} + \epsilon_{it}$$

For each j, the model contains five parameters ϕ_j, δ_j, μ_j, ν_j, and γ_j. Instead, we impose that these parameters are the same regardless of j because all information specific to j that is relevant for s_{it} are encoded in the weights w_{ij} and d_{ij}. Then, we can write these parameters as being independent of j (by dropping the subscript j):

$$s_{it} = \alpha + \psi z_{it} + \sum_j \phi w_{ij} z_{jt} + \sum_j \delta d_{ij} z_{jt} + \sum_j \mu w_{ij} + \sum_j \gamma d_{ij} + \sum_j \nu z_{jt} + \boldsymbol{x}\boldsymbol{\beta} + \epsilon_{it}$$

I factorize these parameters and rewrite the model as:

$$s_{it} = \alpha + \psi z_{it} + \phi \sum_j w_{ij} z_{jt} + \delta \sum_j d_{ij} z_{jt} + \mu \sum_j w_{ij} + \gamma \sum_j d_{ij} + \nu \sum_j z_{jt} + \boldsymbol{x}\boldsymbol{\beta} + \epsilon_{it}$$

Note that $\mu \sum_j w_{ij} + \gamma \sum_j d_{ij}$ is constant within i while $\nu \sum_j z_{jt}$ is constant within t. I replace these terms with node fixed effects ι_i and time fixed effects τ_t to arrive at a more unrestricted model:

$$s_{it} = \alpha + \psi z_{it} + \phi \sum_j w_{ij} z_{jt} + \delta \sum_j d_{ij} z_{jt} + \iota_i + \tau_t + \boldsymbol{x}\boldsymbol{\beta} + \epsilon_{it}$$