The geographic spread of COVID-19 correlates with the structure of social networks as measured by Facebook*

Theresa Kuchler[†] Dominic Russel[‡] Johannes Stroebel[§]

We use aggregated data from Facebook to show that COVID-19 is more likely to spread between regions with stronger social network connections. Areas with more social ties to two early COVID-19 "hotspots" (Westchester County, NY, in the U.S. and Lodi province in Italy) generally had more confirmed COVID-19 cases by the end of March. These relationships hold after controlling for geographic distance to the hotspots as well as the population density and demographics of the regions. As the pandemic progressed in the U.S., a county's social proximity to recent COVID-19 cases and deaths predicts future outbreaks over and above physical proximity and demographics. In part due to its broad coverage, social connectedness data provides additional predictive power to measures based on smartphone location or online search data. These results suggest that data from online social networks can be useful to epidemiologists and others hoping to forecast the spread of communicable diseases such as COVID-19.

To forecast the geographic spread of communicable diseases such as COVID-19, it is valuable to know which individuals are likely to physically interact (Piontti et al., 2018). In particular, since social ties shape patterns of physical interaction, observing the strength of social connections between cities and regions can be useful for determining a locality's risk of future disease outbreaks. Yet, the geographic structure of social networks is usually difficult to measure on a national or global scale. In this paper, we overcome this challenge by using aggregated data from Facebook to measure social connections between regions. We then show that these connectedness measures can help forcecast the geographic spread of communicable diseases such as COVID-19.

*Date: December 11, 2020. Public versions of the social connectedness data used in this paper, as well as similar data for a wide range of other geographies, are accessible at https://data.humdata.org/dataset/social-connectedness-index. The full replication code is available at https://github.com/social-connectedness-index/example-scripts. The authors have a research consulting relationship with Facebook. Since this project only uses data that is available to the broader research community, nobody at Facebook reviewed the contents of this paper.

[†]New York University, Stern School of Business. Email: tkuchler@stern.nyu.edu
[‡]New York University, Stern School of Business. Email: drussel@stern.nyu.edu
[§]New York University, Stern School of Business. Email: johannes.stroebel@nyu.edu (Corresponding)

We construct a measure of the social connectedness between U.S. counties and between Italian provinces. This *Social Connectedness Index* captures the probability that Facebook users in a pair of these regions are Facebook friends with each other (Bailey et al., 2018b). We hypothesize that regions connected through many friendship links are likely to have more physical interactions between their residents, providing opportunities for the spread of communicable diseases. Indeed, our measure has been shown to be predictive of travel patterns across Europe (Bailey et al., 2020d) and within urban areas (Bailey et al., 2020a), suggesting it contains important information about real-world interactions. Most directly, Coven et al. (2020) use our *Social Connectedness Index* to show that counties with higher levels of social connectedness to New York City were more likely to be destinations for those fleeing the city during the pandemic, providing direct evidence for our propopsed mechanism.

After introducing our *Social Connectedness Index*, we show that regions with stronger social ties to early COVID-19 "hotspots" — Westchester County, NY, in the U.S., and Lodi province in Italy — had more documented COVID-19 cases per resident as of March 30, 2020. These relationships are robust to controlling for the geographic distance to these early hotspots, as well as demographic characteristics of the regions. Social connectedness to Westchester has more predictive power for forecasting county-level COVID-19 cases than social connectedness to any other county outside the New York-Newark CSA. These case studies provide initial evidence that social connectedness might serve as a valuable predictive measure in addition to physical distance and other inputs to current epidemiological models.

We then exploit the changing geography of the pandemic in the U.S. to conduct a more systematic in-sample analysis. We construct regional measures of COVID-19 exposure through social connections ("social proximity to cases") and physical distance ("physical proximity to cases"). We find that changes in a county's social proximity to cases in one time period are strongly correlated with the county's subsequent growth in own local cases. Even after controlling for physical proximity to cases and other regional demographics, a doubling in social proximity to cases in one two-week period corresponds to a 24.9% increase in own cases in the next two-week period. These results are unlikely to be explained by differential testing between regions, as an increase in social proximity to deaths in one period also corresponds to an increase in actual deaths in the next period.

To mimic a real-world epidemiological use case, we also conduct a simple out-of-sample prediction exercise. We find that models that include our measure of social proximity to cases are better able to predict a region's future case growth than alternative models that rely only on geographic distance and other demographics. We also compare the predictive value of social proximity to cases to measures from Google searches related to COVID-19 symptoms and the smartphone-based Location Exposure Index (LEX) introduced by

Couture et al. (2020). In counties with both LEX and Google search data, social proximity to cases provides only small additional predictive value — perhaps not surprisingly, given that the real-world movement of people captured by the LEX is precisely the mechanism we conjecture explains the predictive power of social proximity to cases. However, when using the best available model to make predictions for *all* U.S. counties (for many of which no LEX data or Google search information is available), models that include social proximity to cases sizably improve accuracy. This highlights one important advantage of social connectedness data: its broad coverage and global availability.

Our use of the *Social Connectedness Index* to forecast COVID-19 spread adds to an active body of research that studies how aspects of social media and internet-usage patterns can be used for tracking and preventing disease (for an overview, see Aiello et al., 2020). One strand of this literature uses the content of individuals' internet searches or social media posts; most famously, Google Flu Trends used search queries related to influenza for early outbreak detection (Ginsberg et al., 2009). Other researchers have also used content from Twitter posts (Rodríguez-Martínez and Garzón-Alfonso, 2018; Jahanbin and Rahmanian, 2020), Facebook likes (Gittelman et al., 2015), Wikipedia searches (Generous et al., 2014), and Instagram posts (Correia et al., 2016) to predict public health outcomes. A second strand of research, which has received much attention during the COVID-19 pandemic, uses geolocation data to track individuals' movement patterns. These data have been used to explore the determinants and effects of social distancing behavior (for an overview, see Giuliano and Rasul, 2020), as well as forecast disease spread (e.g., Jia et al., 2020; Bengtsson et al., 2015; Wesolowski et al., 2012, 2015; Peixoto et al., 2020). A third strand of that work uses crowdsourced information, including surveys, to monitor disease symptoms and detect outbreaks (see Facebook Symptom Survey; Smolinski et al., 2015; Paolotti et al., 2014).

In comparison to this literature, our stable network-based measure is less likely to suffer from changes in internet behavior or seasonality, both of which have hampered Google Flu Trends (Olson et al., 2013). In addition, our measures do not require individuals to have experienced symptoms, which potentially allows us to identify at-risk localities before disease transmission.[1] Finally, because our measures are based only on aggregated connections (instead of individual movement), they are easily accessible to researchers and consistently available for a large number of granular geographies around the world. For example, the *Social Connectedness Index* is available at the NUTS3 level in Europe, the GADM2 level in the Indian Subcontinent and Canada, and the GADM1 level throughout much of the rest

[1]However, this suggests that our data might partner well with these measures. For example, if one can detect an early outbreak using surveys, they could then predict (and potentially prevent) the next outbreak using information on social connectedness.

of the world.[2] The index not only measures connections *within* countries, but also *between* countries, which may be otherwise challenging with mobility data from different cellphone providers (and important for tracking the international spread of communicable diseases).

More generally, our results add to a literature that has applied aspects of network theory to build spatial epidemiological models (for overviews, see Keeling and Eames, 2005; Keeling and Rohani, 2011; Danon et al., 2011). Works in this literature move beyond the basic assumption that individuals within a population are "fully mixed," or equally likely to interact; instead, they better represent the dynamics of real-world connections (e.g., Newman, 2002; Klovdahl, 1985; Klovdahl et al., 1994; Mossong et al., 2008; Yang et al., 2020). While some of these studies parameterize models with information on local networks, we are unaware of any that introduce a measure with comparably high levels of coverage and granularity. Our hope is that our unique measure of social connectedness can help parameterize future epidemiological work. In addition, we hope that the *Social Connectedness Index* can advance the literature on the determinants and effects of urban and regional social networks (see Bailey et al., 2020a; Kim et al., 2017; Büchel and Ehrlich, 2020; Mossay and Picard, 2011; Brueckner and Largey, 2008; Glaeser et al., 1992).

It is important to note that our objective in this paper is not to incorporate social connectedness into a state-of-the-art epidemiological model. Instead, we provide a unique measure to assess regions' outbreak risk, answering the call of Avery et al. (2020), among others, who highlight an "urgent need" for "creative and entrepreneurial methods" of interpreting and sharing data to model coronavirus spread. To that end, the data used in this paper, as well as similar data for a number of other geographies, are available at https://data.humdata.org/dataset/social-connectedness-index. We encourage interested researchers to use them.

1 Data Description

To measure the intensity of social connectedness between locations, we use a de-identified and aggregated snapshot of all active Facebook users and their friendship networks from March 2020.[3] As of the end of 2019, Facebook had nearly 2.5 billion monthly active users around the world: 248 million in the U.S. and Canada, 394 million in Europe, 1.04 billion in Asia-Pacific, and 817 million in the rest of the world (Facebook, 2020). The data therefore

[2]Interested researchers may also access U.S. ZCTA-level data by emailing sci_data@fb.com.

[3]We use the data from March 2020, since this this allows our analyses to correspond most closely to a real-time forecasting exercise. The publicly available *Social Connectedness Index* data is based on a snapshot from August 2020. Since the SCI is extremely stable over time, results do not change across the various data sets.

has extremely wide coverage, and provides a unique opportunity to map the geographic structure of social networks around the world. Locations are assigned to users based on their information and activity on Facebook, including their public profile information, and device and connection information. Establishing a connection on Facebook requires the consent of both individuals, and there is an upper limit of 5,000 on the number of connections a person can have. As a result, Facebook connections are generally more likely to be between real-world acquaintances than links on many other social networking platforms.

Our measure of the social connectedness between two locations i and j is the *Social Connectedness Index (SCI)* introduced by Bailey et al. (2018b):

$$Social\ Connectedness_{i,j} = \frac{FB\ Connections_{i,j}}{FB\ Users_i * FB\ Users_j}. \tag{1}$$

Here, $FB\ Connections_{i,j}$ is the total number of Facebook friendship links between Facebook users living in location i and Facebook users living in location j. $FB\ Users_i$ and $FB\ Users_j$ are the number of active users in each location. $Social\ Connectedness_{i,j}$ thus measures the relative probability of a Facebook friendship link between a given Facebook user in location i and a given Facebook user in location j: if this measure is twice as large, a given Facebook user in region i is twice as likely to be friends with a given Facebook user in region j.

In previous work, we have shown that this measure predicts a large number of important economic and social interactions. For example, social connectedness as measured through Facebook friendship links is strongly related to patterns of sub-national and international trade (Bailey et al., 2020b), patent citations (Bailey et al., 2018b), and investment decisions (Kuchler et al., 2020).[4] More generally, we have found that information on individuals' Facebook friendship links can help understand their product adoption decisions and their housing and mortgage choices (Bailey et al., 2018a, 2019a,b).

Data on COVID-19 cases in the U.S. by county come from Johns Hopkins University Center for Systems Science and Engineering. Data on COVID-19 cases for each Italian province come from the Italian Dipartimeno della Protezione Civile. Because differential testing across regions may introduce bias in case-based results, we will also use information on COVID-19 related deaths from each source in Section 3.

[4] A growing body of research also uses this measure to study issues related to COVID-19. For example, Bailey et al. (2020c) use the *SCI*, along with individual-level data from Facebook, to show that social network exposure to COVID-19 cases shapes individuals' social distancing behaviors. Holtz et al. (2020) use *SCI* data to document spillover effects in state-level COVID-19 health policies. Charoenwong et al. (2020) and Makridis and Wang (2020) work with the *SCI* to study other behavioral effects of exposure to COVID-19 through friends.

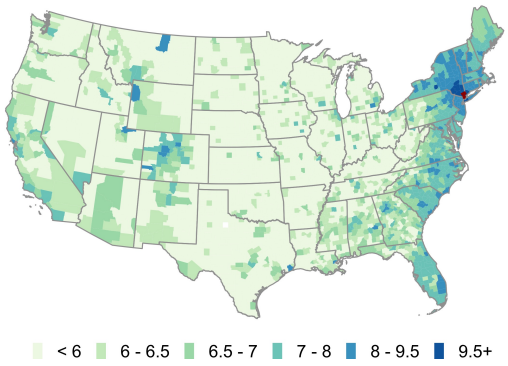
2 Early Hotspot Analysis

In this section, we explore how the domestic spread of confirmed COVID-19 cases is related to the social connectedness to two early COVID-19 "hotspots": Westchester County, NY, in the U.S., and Lodi Province in Italy. Westchester County includes New Rochelle, a community that had the first major confirmed COVID-19 outbreak in the eastern United States (Chappell, 2020). By March 20th, the county had over 9,300 cases, second only to nearby New York City. Additionally, a number of articles reported that wealthy residents from Westchester and the New York area had fled to other parts of the U.S. (Tully and Stowe, 2020), providing a vector that could potentially spread the disease. Indeed, geneticists and epidemiologists later found that travel from New York seeded much of the first wave of U.S. COVID-19 outbreaks (Carey and Glanz, 2020). Social connections to Westchester may thus provide particularly important information for tracking early COVID-19 spread, especially given that Coven et al. (2020) find that social connectedness to New York City predicted travel patterns from the city early in the pandemic. Lodi is an Italian province of around 230,000 inhabitants in the heavily impacted region of Lombardy. It contains Codogno, where the earliest cases of COVID-19 in Italy were detected, and was at the center of Italy's outbreak (Horowitz et al., 2020).

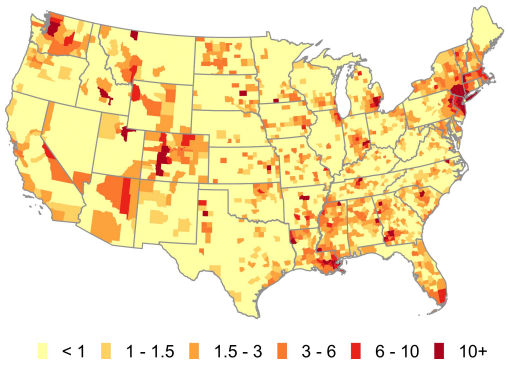
Panel (a) of Figure 1 shows a heatmap of the social connectedness of Westchester County, NY, to other U.S. counties; darker colors correspond to stronger social ties. Panel (b) shows the distribution of COVID-19 cases per 10,000 residents across U.S. counties on March 30, 2020, with darker colors corresponding to higher COVID-19 prevalence. These maps show a number of similarities. Perhaps most notably, coastal regions and urban centers appear to have both high levels of connectedness to Westchester and larger numbers of COVID-19 cases per resident. But a number of more subtle patterns also emerge. Both measures are high in the communities along the Florida coast (in particular along the southeastern coast, near Miami), in western and central Colorado (in particular in areas with ski resorts), and in the upper Northeast. These areas are all popular vacation destinations and second home locations for many well-heeled residents of Westchester. Indeed, the governors of Florida and Rhode Island publicly lamented the number of New York area residents fleeing to their states and spreading COVID-19 (Mower, 2020; Carlisle, 2020). By contrast, many areas that are geographically closer but less socially connected to Westchester, such as counties in western Pennsylvania and West Virginia, had fewer confirmed COVID-19 cases on March 30. There are also a number of patterns of COVID-19 prevalence that connectedness to Westchester alone cannot explain. Areas around King County, WA (Seattle), for example, have relatively low connectedness to Westchester, but were an independent early hotspot of COVID-19.

Figure 1: Social Network Distributions from Westchester and COVID-19 Cases in the U.S.

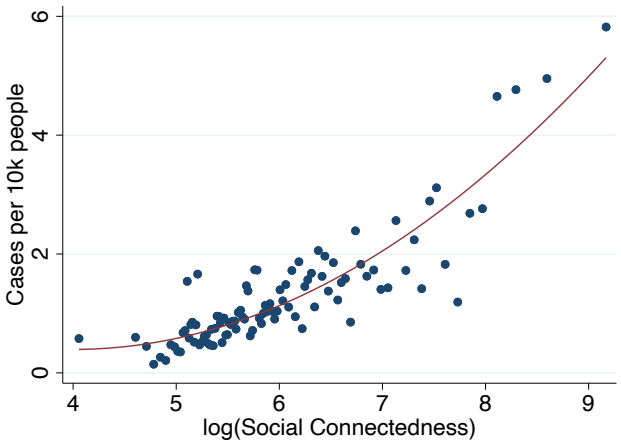
(a) Log of SCI to Westchester County, NY



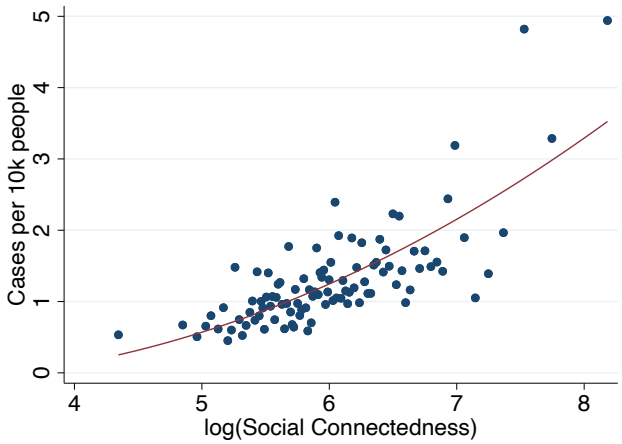
(b) COVID-19 Cases per 10k Residents by County



(c) Westchester binscatter without controls



(d) Westchester binscatter with controls



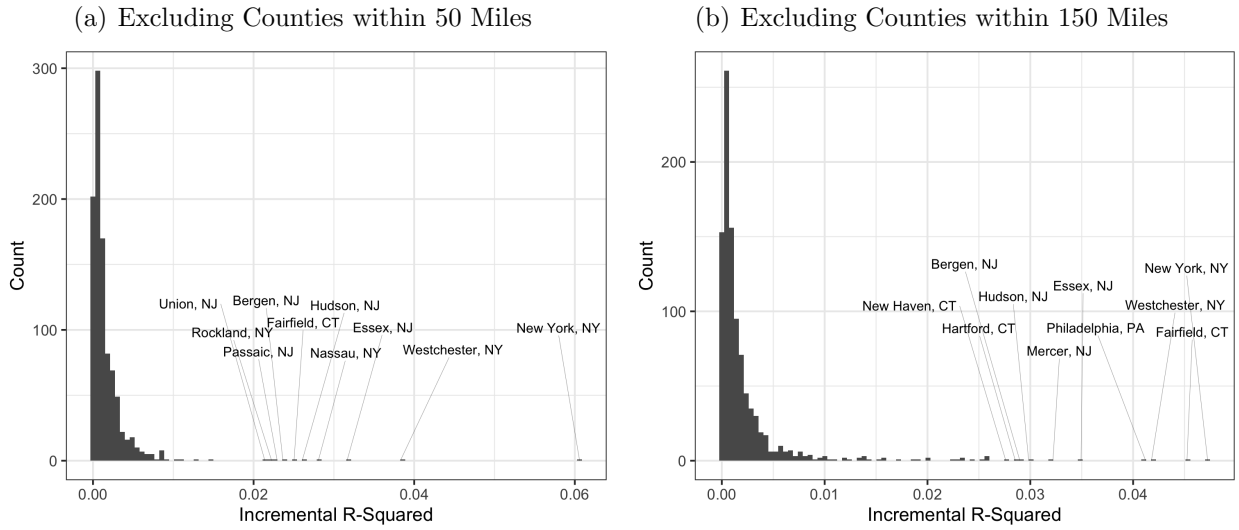
Note: Panel (a) shows the social connectedness to Westchester for U.S. counties. Panel (b) shows the number of confirmed COVID-19 cases per 10,000 residents by U.S. county on March 30, 2020. Panels (c) and (d) show binscatter plots with counties more than 50 miles from Westchester as the unit of observation. To generate the plot in Panel (c), we group $log(SCI)$ into 100 equal-sized bins and plot the average against the corresponding average case density. Panel (d) is constructed in a similar manner. However, we first regress $log(SCI)$ and cases per 10,000 residents on a set of control variables and plot the residualized values on each axis. Red lines show quadratic fit regressions. The controls for Panel (d) are 100 dummies for the percentile of the county's geographic distance to Westchester; population density; median household income; and dummies for the six National Center for Health Statistics Urban-Rural county classifications.

The two bottom panels of Figure 1 explore the relationship between COVID-19 prevalence and social ties to Westchester more formally. Panel (c) shows a binscatter plot of social connectedness to Westchester County and the number of COVID-19 cases per 10,000 residents. We exclude those counties within 50 miles of Westchester County: while those areas have strong social links to Westchester, they are also close enough geographically such that their populations might interact physically with Westchester residents even in the absence of social links (e.g., in supermarkets and houses of worship). There is a strong positive relationship between social ties to Westchester and COVID-19 prevalence. Quantitatively, a doubling of a county's social connectedness to Westchester is associated with an increase of about 0.88 COVID-19 cases per 10,000 residents. The R-squared of this relationship is 0.093, suggesting that, in a statistical sense, 9.3% of the cross-county variation in COVID-19 cases can be explained by counties' social connectedness to Westchester.

One concern with interpreting these initial correlations is that they might be primarily picking up other factors that affect the spread of COVID-19, and that are correlated with social connectedness to Westchester. Specifically, even after dropping counties within 50 miles of Westchester, the correlations might be primarily picking up geographic distance to Westchester (which is related to the number of friendship links to Westchester). As a result, including social connectedness might not improve predictive power for models that already control for some of these other variables. In Panel (d), we therefore present a binscatter plot of the relationship between social connectedness to Westchester County and COVID-19 cases that controls for a number of these possible confounding variables (in addition to excluding nearby counties). Most importantly, we non-parametrically control for the geographic distance between each county and Westchester County by including 100 dummies for percentiles of that distance. We also control for median income, population density, and a classification of how urban/rural a county is. Even conditional on these other factors, Panel (d) shows a strong positive relationship between COVID-19 cases as of March 30, 2020 and social connectedness to Westchester County. With these controls, a doubling of a county's social connectedness to Westchester is associated with an increase of about 0.80 COVID-19 cases per 10,000 residents. The total R-squared of the statistical relationship is 0.190, while the incremental R-squared from controlling for social connectedness to Westchester is 0.037.

Another potential concern stems from the fact that the underlying social network and the site of the initial hotspot are nonrandom. This may confound our interpretation if, for example, counties with ties to Westchester were also destinations for European travelers seeding the virus in the United States. To contextualize the effect of connections to Westchester in particular, we next run "placebo" regressions, identical to the one shown for Westchester in panel (d) of Figure 1, for every U.S. county with a population over 50,000.

Figure 2: Incremental R^2 from Adding Connections to Individual U.S. Counties



Note: Panels show results from regressions to predict COVID-19 cases per 10k people by county on March 30, 2020. The incremental R^2 is the increase in R^2 from adding $log(SCI)$ and $log(SCI)^2$ to a particular U.S. county, over and above a set of baseline control variables: 100 dummies for percentiles of distance to the county under investigation; population density; median household income; and dummies for the six National Center of Health Statistics Urban-Rural county classifications. The graphs show the distributions over the incremental R^2s for adding social connectedness to each county with a population over 50,000 in turn. Each regression in panels (a) and (b) excludes counties within 50 and 150 miles of the county of interest, respectively. In each panel the 10 largest incremental R^2 are labeled.

Figure 2 shows the incremental R-squared from adding social connectedness in each of these regressions. Panel (a) excludes counties within 50 miles of the chosen county, as in Figure 1. Westchester's 0.037 incremental R-squared is second only to New York City, and each top 10 county is in the New York-Newark Combined Statistical Area (CSA).[5] That connections to each of these counties matters so strongly suggests that, although Westchester contained the earliest discovered COVID-19 outbreak in the eastern U.S., community spread may have already been present in many neighboring counties. Panel (b) shows results for regressions excluding counties within 150 miles of the chosen county. Doing so will exclude New York City and Westchester cases from every regression for a New York-Newark CSA county.[6] Counties within the CSA remain as 9 of the top 10, including the top 3: New York City, Fairfield, and Westchester.[7] These findings highlight that social connections to other counties that may have similar demographics to Westchester, but that did not have an early COVID-19 outbreak, do not help with forecasting COVID-19 spread. In turn, this

[5]Although New York City contains multiple counties, early NYC COVID-19 data was not disaggregated. As such, we combine the counties in this section. NYC is disaggregated in all analyses in Section 3.

[6]The maximum distance from any county in the CSA to New York City and Westchester is 91 miles and 104 miles, respectively.

[7]Anecdotally, Williamson and Hussey (2020) linked Fairfield to COVID-19 spread early in the pandemic.

suggests our previous results are not due to omitted variables whereby counties with links to Westchester are more susceptible to COVID-19 outbreaks for some other reason.

It is important to highlight that the purpose of this exercise is to demonstrate the *predictive power* of social connectedness measured via online social networks for COVID-19 prevalence. The control variables highlight that the *Social Connectedness Index* has such predictive power over and above a number of variables on which data is already easily available, and that may partially proxy for social connections in models of communicable disease spread. We will benchmark this predictive power against other measures, such as smartphone location pings and Google searches for COVID-19 symptoms, in Section 3.

Figure 3 explores the analogous relationships for Lodi province in Italy.[8] The provinces with highest COVID-19 case densities and connectedness to Lodi are in the surrounding Lombardy region, as well as the nearby Piemonte and Veneto regions. There are also relatively high levels of both connectedness to Lodi and COVID-19 cases in Rimini, a popular tourist destination along the Adriatic sea. A number of provinces in southern Italy send workers and students to the industrial Lombardy region, and therefore have strong social ties to that region. While some of these areas have seen a number of COVID-19 cases, they are not disproportionally larger, perhaps reflecting the efforts of Italian authorities to restrict the movement of individuals (Kington, 2020). Panels (c) and (d) repeat the binscatter exercises from Figure 1 (there are fewer data points in Figure 3 than there are in Figure 1, since there are fewer Italian provinces than U.S. counties). We exclude provinces within 50 kilometers of Lodi. In Panel (d) we control for geographic distance using 20 dummies for quantiles of distance from each province to Lodi, as well as GDP per inhabitant and population density. As before, we find that the *Social Connectedness Index* appears to have predictive power above these other measures that might commonly be used to proxy for social interactions. Quantitatively, the estimates from Panel (d) suggest that a doubling of the SCI corresponds to an increase of 16.6 COVID-19 cases per 10,000 residents. The incremental R-squared of including social connectedness to Lodi over the other control variables is 0.057.[9]

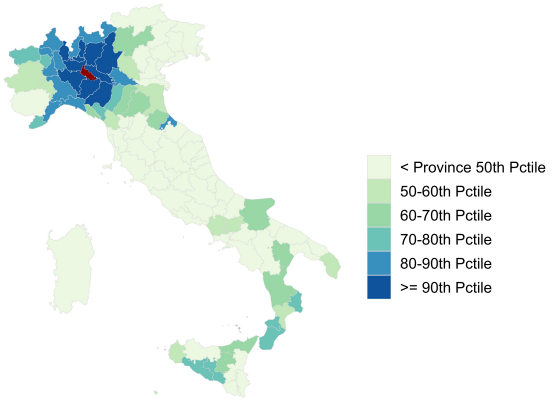
Taken together these case studies illustrate the potential usefulness of our measure of

[8]Because Italian provinces on the island of Sardinia do not align with European NUTS3 regions (the level at which we measure social connectedness), we include Sardinia as a single observation in our analysis.

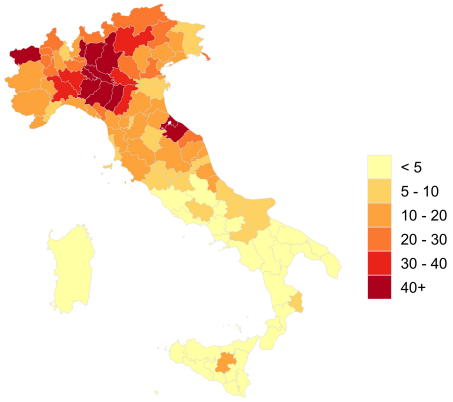
[9]In Appendix A, we conduct an additional exercise to mimic a potential real-world use case in which U.S. public health officials might have seeked to predict disease spread from the initial Westchester outbreak. Since, by March 10, there was not yet enough documented domestic COVID-19 spread to parameterize a forcecasting model based on U.S. data, these officials might have looked to Italy to understand how social connections to early hotspots translate into subsequent case growth. To replicate such an analysis, we train a model using Italian provinces and their conntectedness to Lodi to predict Italian COVID-19 cases as of March 10. We then use that model to forecast U.S. COVID-19 cases as of March 30 based on counties' social connectedness to Westchester as the initial hotspot. We find that including social connectedness as a model input in this exercise improves out-of-sample predictions of COVID-19 cases across U.S. counties.

Figure 3: Social Network Distributions of Lodi and COVID-19 Cases in Italy

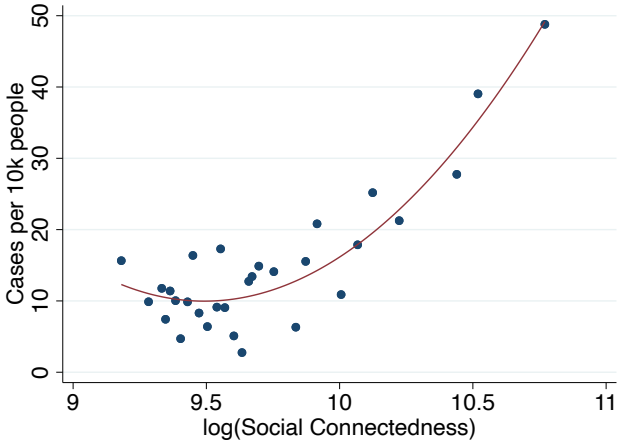
(a) Percentile of SCI to Lodi Province, Italy



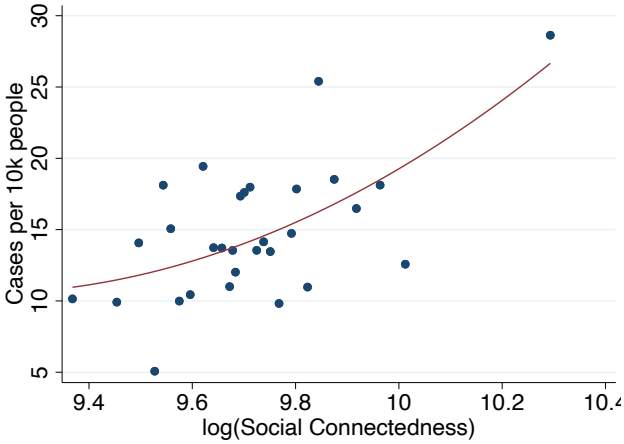
(b) COVID-19 Cases per 10k Residents by Province



(c) Lodi binscatter without controls



(d) Lodi binscatter with controls



Note: Panel (a) shows the social connectedness to Lodi for Italian provinces. Panel (b) shows the number of confirmed COVID-19 cases by Italian province on March 30, 2020. Panels (c) and (d) show binscatter plots with provinces more than 50 kilometers from Lodi as the unit of observation. To generate the plot in Panel (c) we group $log(SCI)$ into 30 equal-sized bins and plot the average against the corresponding average case density. Panel (d) is constructed in a similar manner. However, we first regress $log(SCI)$ and cases per 10,000 residents on a set of control variables and plot the residualized values on each axis. Red lines show quadratic fit regressions. The controls for Panel (d) are 20 dummies for quantiles of the province's geographic distance to Lodi; GDP per inhabitant; and population density.

social connectedness for predicting disease spread. In the next section, we will use a time series of case growth from March through November, as well as additional predictive measures from smartphone locations and Google searches, to explore this potential in more detail.

3 Time Series Analysis

In this section, we exploit the changing geography of the pandemic in the U.S. to more systematically investigate the predictive value of the *Social Connectedness Index* for forecasting the spread of COVID-19. We construct two primary time-varying metrics: "Social Proximity to Cases", a county-level measure of exposure to COVID-19 cases through social networks, and "Physical Proximity to Cases", a county-level measure of exposure through physical proximity. While the two measures will be related (because individuals generally have stronger social ties to those who are geographically nearby, as documented in Bailey et al., 2018b), the examples in the previous section illustrate that some geographically distant places — such as Westchester and the east coast of Florida — can have strong social ties. To benchmark the predictive power of social connectedness, we also construct measures using data from smartphone locations and Google symptom searches.

Key Variable Construction. We construct our measure of social proximity to cases as:

$$Social\ Proximity\ to\ Cases_{i,t} = \sum_j Cases\ Per\ 10k_{j,t} * \frac{Social\ Connectedness_{i,j}}{\sum_h Social\ Connectedness_{i,h}}. \quad (2)$$

$Cases\ Per\ 10k_{j,t}$ is the number of confirmed COVID-19 cases per 10,000 residents in county j as of time t. The sums j and h are over all counties. Analogously, we construct a measure of a county's physical proximity to cases as:

$$Physical\ Proximity\ to\ Cases_{i,t} = \sum_j Cases\ Per\ 10k_{j,t} * \frac{1}{1 + Distance_{i,j}}. \quad (3)$$

Here, $Distance_{i,j}$ is the physical distance between counties i and j measured in miles. We create a further related exposure measure using smartphone location data. Specifically, Couture et al. (2020) create a Location Exposure Index (LEX) that measures, among smartphones that pinged in a given county i today, the share that pinged in each county j at least once during the previous 14 days. We use these matricies to construct:

$$LEX\ Proximity\ to\ Cases_{i,t} = \sum_j Cases\ Per\ 10k_{j,t} * \frac{LEX_{i,j,t}}{\sum_h LEX_{i,h,t}}. \quad (4)$$

We also use data from Google LLC on searches related to COVID-19 symptoms. The data include a county by week normalized (within county) probability that a user will make a symptom-related search. For each county and two-week period, we define the change in searches related to a symptom as the percent change in this probability between the second week of the period and the second week of the previous period. We use searches related to fever, cough, and fatigue. We provide additional details in Appendix D. Finally, to explore whether it is the specific bilateral patterns of connectedness or a county's overall level of connectedness that is most relevant for predicting the spread of COVID-19, we include controls for the share of a county's Facebook connections that are within 50 and 150 miles.[10]

Empirical Specification. We first study the relationship between observed case growth and "lagged" (i.e., in past time periods) growth in our measures. We hypothesize that if social connectedness is an important predictor of the path of COVID-19 spread, a lagged measure of social proximity to new cases will have a positive relationship with new case counts in the next period. For each county i and time period t, our baseline specification is:

$$\begin{aligned} log(\Delta\ Cases\ per\ 10k+1)_{i,t} &= \beta_1 * log(\Delta Cases\ per\ 10k+1)_{i,t-1} \\ &+ \beta_2 * log(\Delta Cases\ per\ 10k+1)_{i,t-2} \\ &+ \beta_3 * log(\Delta Social\ Proximity\ to\ Cases)_{i,t-1} \\ &+ \beta_4 * log(\Delta Social\ Proximity\ to\ Cases)_{i,t-2} \\ &+ \beta_5 * Share\ Friends\ within\ 50mi_i \\ &+ \beta_6 * Share\ Friends\ within\ 150mi_i \\ &+ \beta_7 * log(\Delta Physical\ Proximity\ to\ Cases)_{i,t-1} \\ &+ \beta_8 * log(\Delta Physical\ Proximity\ to\ Cases)_{i,t-2} \\ &+ X_{i,t} + \epsilon_{i,t} \end{aligned} \tag{5}$$

Here, t is defined as one of the eight two-week time periods between March 30 and November 2, 2020. For each time period t, prior two-week periods are denoted $t-2$ and $t-1$ (for example, March 3 - 16 and March 16 - 30 for the first period starting March 30). We always include two lags of own case growth, and explore the effects of lagged changes of social and physical proximity to cases. In some specifications we will add controls for $log(\Delta LEX\ Proximity\ to\ Cases+1)_{i,t}$, lagged by one and two time periods, and for the percent change in Google searches related to fever, cough, and fatigue for this period or

[10] With the underlying assumption that Facebook usage rates (as a share of the true population) are roughly equal across counties, we construct these controls using the *Social Connectedness Index* as $Share\ Within\ K\ MI_i = [SCI_{i,j} * Pop_j * 1(Distance_{i,j} < K)]/[\sum_h SCI_{i,h} * Pop_h]$, for county i.

lagged by one period. $X_{i,t}$ are a set of time-specific fixed effects, including percentiles of population density and median household income. In our strictest specification, we also add time $\times$ state fixed effects. To rule out differential testing across regions driving our results, we also conduct a similar exercise replacing COVID-19 cases with COVID-19-related deaths. For these analyses, we use four-week time periods, beginning with April 28 - May 25, with our exposure measures lagged by four and eight weeks.

Regression Analysis. Panel A of Table 1 shows that past growth in social proximity to COVID-19 cases in one period has a strong positive relationship with actual growth in cases in the subsequent period. In columns 1 and 2, we document this relationship without controlling for physical distance to cases (column 2 adds state $\times$ time fixed effects to the specification in column 1, to control for time-varying state-level differences in public health measures). In contrast, columns 3 and 4 show that there is no systematic relationship between the share of a county's friends that are within 50 and 150 miles and COVID-19 cases. This suggests that it is the specific bilateral patterns of social connections that correlate with disease spread, not simply that counties with more "open" networks experience worse outbreaks in every period. Columns 5 and 6 show that physical proximity to cases is also strongly correlated with subsequent case growth, a relationship which may confound the one in columns 1 and 2. To address this, columns 7 and 8 include both the physical proximity and social proximity measures. While the coefficient on social proximity to cases declines somewhat — suggesting some of the relationship is due to physical proximity — the relationship remains highly statistically and economically significant. In our strictest specification, which includes state $\times$ period fixed effects, a doubling of social proximity to cases in one period corresponds to a 24.9% increase in cases per capita in the next period.

Panel B of Table 1 presents the same specifications, using COVID-19 deaths (instead of COVID-19 cases) as the dependent variable. The relationships are very similar, suggesting that our results are not driven by differential testing across counties that might have been correlated with social proximity to cases.

In Appendix B we conduct two additional regression exercises. First, we run regression 5 separately for each time period, allowing us to study how the relationship between social connections and new COVID-19 cases changes over the course of the pandemic. Table A2 shows that, in every two-week period from March 30 to November 2, a one-period lagged measure of social proximity to cases was a statistically significant predictor of actual case growth. In Table A3, we add additional measures from smartphone locations and symptom searches to our regression framework. We find that changes in Google symptom searches — both in the current period and in the previous period — and lagged LEX proximity to

Table 1: COVID-19 Case Growth and Prior Proximity to Cases

Panel A	log(Change in Cases per 10k Residents + 1)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
2 Week Lag: log(Change in Social Proximity to Cases + 1)	0.589*** (0.041)	0.415*** (0.036)					0.414*** (0.041)	0.321*** (0.037)
4 Week Lag: log(Change in Social Proximity to Cases + 1)	-0.124*** (0.037)	-0.080** (0.032)					-0.002 (0.036)	0.010 (0.032)
Share of Friends within 50 Miles			0.096 (0.106)	0.031 (0.086)			0.050 (0.100)	0.076 (0.082)
Share of Friends within 150 Miles			0.018 (0.123)	0.214* (0.113)			-0.256** (0.124)	0.143 (0.109)
2 Week Lag: log(Change in Physical Proximity to Cases + 1)					1.432*** (0.129)	1.754*** (0.184)	1.244*** (0.118)	1.388*** (0.176)
4 Week Lag: log(Change in Physical Proximity to Cases + 1)					-1.208*** (0.131)	-1.433*** (0.196)	-1.037*** (0.121)	-1.225*** (0.187)
2 Week Lag: log(Change in Cases per 10k Residents + 1)	0.317*** (0.022)	0.316*** (0.018)	0.646*** (0.012)	0.526*** (0.011)	0.604*** (0.011)	0.514*** (0.010)	0.372*** (0.022)	0.351*** (0.019)
4 Week Lag: log(Change in Cases per 10k Residents + 1)	0.113*** (0.019)	0.092*** (0.016)	0.077*** (0.009)	0.063*** (0.008)	0.097*** (0.009)	0.072*** (0.008)	0.071*** (0.019)	0.056*** (0.017)
Time x Pop. Density FEs	Y	Y	Y	Y	Y	Y	Y	Y
Time x Median Household Income FEs	Y	Y	Y	Y	Y	Y	Y	Y
Time x State FEs		Y		Y		Y		Y
Sample Mean	2.177	2.177	2.177	2.177	2.177	2.177	2.177	2.177
R-Squared	0.717	0.755	0.706	0.752	0.718	0.754	0.725	0.757
N	47,040	47,025	47,040	47,025	47,040	47,025	47,040	47,025

Panel B	log(Change in Deaths per 10k Residents + 1)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
4 Week Lag: log(Change in Social Proximity to Deaths + 1)	0.471*** (0.058)	0.240*** (0.049)					0.273*** (0.049)	0.141*** (0.046)
8 Week Lag: log(Change in Social Proximity to Deaths + 1)	-0.018 (0.054)	-0.057 (0.041)					0.187*** (0.052)	0.084* (0.043)
Share of Friends within 50 Miles			0.109 (0.076)	0.149** (0.070)			0.060 (0.066)	0.156** (0.066)
Share of Friends within 150 Miles			0.040 (0.083)	0.129 (0.078)			-0.014 (0.081)	0.116 (0.074)
4 Week Lag: log(Change in Physical Proximity to Deaths + 1)					0.738*** (0.069)	0.899*** (0.125)	0.691*** (0.067)	0.802*** (0.124)
8 Week Lag: log(Change in Physical Proximity to Deaths + 1)					-0.657*** (0.077)	-0.828*** (0.136)	-0.699*** (0.078)	-0.865*** (0.142)
4 Week Lag: log(Change in Deaths per 10k Residents + 1)	0.163*** (0.032)	0.230*** (0.027)	0.467*** (0.021)	0.366*** (0.018)	0.425*** (0.018)	0.361*** (0.016)	0.247*** (0.027)	0.276*** (0.026)
8 Week Lag: log(Change in Deaths per 10k Residents + 1)	0.016 (0.033)	0.052** (0.022)	0.025 (0.019)	0.019 (0.018)	0.063*** (0.018)	0.032* (0.018)	-0.064** (0.032)	-0.019 (0.023)
Time x Pop. Density FEs	Y	Y	Y	Y	Y	Y	Y	Y
Time x Median Household Income FEs	Y	Y	Y	Y	Y	Y	Y	Y
Time x State FEs		Y		Y		Y		Y
Sample Mean	0.375	0.375	0.375	0.375	0.375	0.375	0.375	0.375
R-Squared	0.374	0.455	0.360	0.454	0.384	0.459	0.392	0.461
N	21,952	21,945	21,952	21,945	21,952	21,945	21,952	21,945

Note: Table shows results from regression 5. In Panel A, each observation is a county × two-week period (between March 30 and November 2, 2020). The dependent variable is log of one plus the number of new COVID-19 cases per 10,000 residents. In Panel B, each observation is a county × four-week period (between April 28 and November 2, 2020). The dependent variable is log of one plus the number of new COVID-19 deaths per 10,000 residents. Columns 1 and 2 include log of growth in social proximity to cases (deaths) lagged by one and two periods (two and four weeks in Panel A, four and eight weeks in Panel B). Columns 5 and 6 include analogous measures of physical proximity to cases (deaths). Columns 3 and 4 also control for the share of a county's Facebook connections that are within 50 and 150 miles. Columns 7 and 8 include all measures. All columns include controls for one and two period lagged changes in cases (deaths), as well as time-specific fixed effects for percentiles of county population density and median household income. Columns 2, 4, 6, and 8 include additional time × state fixed effects. Standard errors are clustered at the time × state level. Significance levels: *(p<0.10), **(p<0.05), ***(p<0.01).

cases are strongly correlated with present case growth. However, even in the presence of each of these other predictors, changes in the social proximity to cases remains a significant predictor of subsequent case growth in sample. We next benchmark the predictive power of social connectedness to these measures using an out-of-sample prediction exercise.

Out-of-Sample Prediction Analysis. Building on our previous results, we next conduct a simple out-of-sample prediction exercise. During a pandemic, local policymakers might want to determine their localities' risks for an outbreak in real time to inform public health measures. With this case in mind, we build a series of simple models that use available data at time t to predict case growth in counties at time $t + 1$. We test the added predictive value of social proximity to cases by building separate models that include and exclude this measure, as well as other possible predictors. Because we do not use the "test" data to train the models, a reduction in prediction error would be reflective of a true improvement in real-world predictions of COVID-19 cases that could have been achieved from using social connectedness data (as opposed to the increase in in-sample R^2 in our previous analyses).

Table 2 shows the results of this prediction exercise. The results in all columns are generated using a random forest, an ensemble prediction algorithm commonly used in data science applications. The algorithm allows us to find non-linear relationships between variables, without overfitting, by aggregating mean predictions from a number of regression trees generated over sample subsets of both observations and input variables.[11] In most settings, random forest out-of-sample predictions outperform those of linear models.

Columns 1-3 describe the prediction error from a simple model that includes the measures from columns 7 and 8 in Panel A of Table 1.[12] Column 1 excludes the two lagged measures of social proximity to cases and column 2 includes them. Columns 1 and 2 show the root mean squared error (RMSE) from a model trained using data from all periods before the period of interest, then tested on that next period; each prediction period is shown as a separate row. The RMSE for both models generally decreases as the training sample gets larger, ending at 0.667 and 0.652 log new cases per 10,000 residents. Column 3 shows the difference in RMSE between the two models, with negative numbers indicating an improvement in out-of-sample fit from including social proximity to cases. In every row, the RMSE is lower when including social proximity to cases, suggesting it does significantly improve predictions.

In columns 4-9 we add information on Google symptom searches and mobility based on smartphone locations. Doing so allows us to benchmark the predictive value of social

[11] In our analysis we use 500 trees. For more information on random forests, see Breiman (2001).

[12] Here, we use non-binned measures of population density and median household income. Because random forests are able to identify non-linear relationships between input and outcome measures, we no longer need to split the measures into percentiles to allow for non-linear relationships.

Table 2: Predicting COVID-19 cases in U.S., with and without Social Proximity to Cases

	RMSE: Baseline Model			RMSE: Best Available Model			RMSE: Counties w/ Google + LEX Only		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Without Social Proximity	With Social Proximity to Cases	Diff. from Social Proximity	Without Social Proximity	With Social Proximity to Cases	Diff. from Social Proximity	Without Social Proximity	With Social Proximity to Cases	Diff. from Social Proximity
(1) April 14 - April 27	1.636	1.534	-0.102	1.488	1.387	-0.102	1.399	1.299	-0.100
(2) April 28 - May 11	0.900	0.838	-0.062	0.954	0.889	-0.066	0.887	0.835	-0.053
(3) May 12 - May 25	0.746	0.722	-0.024	0.771	0.746	-0.025	0.671	0.646	-0.025
(4) May 26 - June 8	0.704	0.680	-0.024	0.687	0.675	-0.012	0.584	0.581	-0.003
(5) June 9 - June 22	0.800	0.776	-0.024	0.779	0.766	-0.013	0.669	0.660	-0.010
(6) June 23 - July 6	0.859	0.838	-0.021	0.809	0.798	-0.011	0.665	0.667	0.002
(7) July 7 - July 20	0.793	0.780	-0.013	0.733	0.730	-0.003	0.530	0.526	-0.004
(8) July 21 - Aug. 10	0.755	0.719	-0.036	0.725	0.701	-0.024	0.508	0.509	0.002
(9) Aug. 11 - Aug. 24	0.770	0.740	-0.030	0.741	0.720	-0.022	0.530	0.517	-0.014
(10) Aug. 25 - Sep. 7	0.725	0.719	-0.005	0.728	0.722	-0.006	0.503	0.503	0.000
(11) Sep. 8 - Sep. 21	0.699	0.691	-0.008	0.694	0.686	-0.009	0.495	0.494	-0.001
(12) Sep. 22 - Oct. 5	0.748	0.719	-0.029	0.726	0.705	-0.021	0.513	0.511	-0.002
(13) Oct. 6 - Oct. 19	0.688	0.662	-0.026	0.684	0.658	-0.025	0.475	0.479	0.004
(14) Oct. 20 - Nov. 2	0.667	0.652	-0.015	0.647	0.628	-0.018	0.462	0.455	-0.007

Note: Table shows results from county-level predictions of COVID-19 case growth. The predicted outcome is log of one plus the number of new COVID-19 cases per 10,000 residents. All columns show root mean squared errors (RMSEs) from a random forest model trained on data from all periods prior to the period of interest. The model inputs in column 1 are population density; median household income; and log of growth in physical proximity to cases and actual cases, lagged by two and four weeks (one and two time periods). Columns 4 and 7 include information on one and two period lagged measures of LEX proximity to cases, and one period lagged percent changes in Google searches related to fever, cough, and fatigue. Column 4 includes predictions for 3,136 counties using, for each county, a model that utilizes the most available information possible. Column 7 limits to the 1,976 counties for which we have both Google symptom search and LEX data. Columns 2, 5, and 8 add lagged measures of social proximity to cases to columns 1, 4, and 7. Columns 3, 6, and 9 show the change in RMSE from adding social proximity to cases.

proximity to cases over and above these other predictors. Columns 4-6 include predictions for all counties included in the COVID-19 case data. We make a prediction for each county using the "best" model (in terms of model features) based on data availability. For example, for a county with LEX and Google data, we will predict cases using a model trained with LEX and Google data. For a county with only LEX data, we will predict using a model trained without Google data, and so on.[13] Column 6 shows that, once again, RMSE decreases in every period after including social proximity to cases, highlighting its incremental predictive value over and above other measures one might have used.

In columns 7-9 we limit to the 1,976 counties which have both LEX and Google symptom search data. Column 9 shows that in 10 of 14 periods, predictions using social connectedness do outperform the comparison model. However, the differences are generally small, suggesting that when limiting to *only* counties with LEX and Google data, social proximity to cases may provide only a small degree of additional predictive value. This is perhaps unsurprising: our proposed mechanism by which social connectedness helps forecast COVID-19 spread is through predicting in-person interactions, which are more directly measured in LEX data.[14]

Tthe fact that social connectedness consistently improves predictions in the full set of U.S. counties (columns 4-6) highlights an important availability advantage of the data. While the LEX and Google data are limited to counties with a sufficient number of devices or searches in a period, the relatively stable nature of social connectedness over time (combined with Facebook's large user base) allows the SCI to be available in more counties, and potentially also at finer levels, such as zip codes.[15] Furthermore, Facebook's global reach allows for SCI measures *within and between* most parts of the world. We are unaware of smartphone location data that can similarly measure, for example, connections between GADM1 regions in Africa, NUTS3 regions in Europe, and U.S. counties — and information on these connections may aid in forecasting the global spread of communicable diseases.

[13]This requires training four models: (1) baseline; (2) baseline + LEX; (3) baseline + Google; and (4) baseline + LEX + Google. We train each model on every county that has the necessary non-missing data. For example, a county with LEX and Google data will be used to train all four models.

[14]In Appendix C we repeat this exercise using COVID-19 related deaths. We find similar results. Baseline models that include social proximity to deaths have smaller prediction errors than models that exclude the measure. Social proximity to deaths does not appear to add sizable predictive value to LEX and Google measures when limiting to counties with both sets of data; however, when using the "best available" model in any county, social proximity to deaths *does* sizably improve predictions.

[15]The SCI data used for this paper include 3,141 counties, the Google data include 2,572 counties, and the LEX data include 2,018 counties.

4 Conclusion

In the context of threats from communicable diseases such as COVID-19, a region's ability to determine optimal public health responses depends on its ability to forecast the risk of an outbreak (Reich et al., 2019). A primary determinant of this risk is the likelihood of physical interactions between the region's residents and residents of other areas with severe outbreaks. Information on the geography of social connections, which shape patterns of physical interactions, are therefore crucially important for public health officials. In this paper, we use de-identified and aggregated data from Facebook to measure social connections between regions, and find those connections to be an important predictor of outbreaks during the COVID-19 pandemic. We show that areas that are more connected to early pandemic hotspots in the U.S. and Italy had, on average, higher case counts by March 30, 2020, even after controlling for physical distance and other demographics. Furthermore, due to its broad geographic coverage, social connectedness data improves out-of-sample predictions of COVID-19 spread during the U.S. pandemic beyond smartphone location and Google symptom search data.

The methodologies we use should not be interpreted as an attempt to create a state-of-the-art epidemiological model. However, our results strongly suggest that our measure of social connectedness may prove useful in future epidemiological work. In particular, its high-degree of availability — in terms of both geographic coverage and granularity — allow social connectedness to provide predictive power over and above other available measures.

References

A. E. Aiello, A. Renson, and P. N. Zivich. Social media- and internet-based disease surveillance for public health. *Annual Review of Public Health*, 41:101–118, 2020.

C. Avery, W. Bossert, A. Clark, G. Ellison, and S. F. Ellison. Policy implications of models of the spread of coronavirus: Perspectives and opportunities for economists. Working Paper 27007, National Bureau of Economic Research, 2020.

M. Bailey, R. Cao, T. Kuchler, and J. Stroebel. The economic effects of social networks: Evidence from the housing market. *Journal of Political Economy*, 126(6):2224–2276, 2018a.

M. Bailey, R. Cao, T. Kuchler, J. Stroebel, and A. Wong. Social connectedness: Measurements, determinants, and effects. *Journal of Economic Perspectives*, 32(3):259–80, 2018b.

M. Bailey, E. Dávila, T. Kuchler, and J. Stroebel. House price beliefs and mortgage leverage choice. *The Review of Economic Studies*, 86(6):2403–2452, 2019a.

M. Bailey, D. M. Johnston, T. Kuchler, J. Stroebel, and A. Wong. Peer effects in product adoption. Working Paper 25843, National Bureau of Economic Research, 2019b.

M. Bailey, P. Farrell, T. Kuchler, and J. Stroebel. Social connectedness in urban areas. *Journal of Urban Economics*, page 103264, 2020a.

M. Bailey, A. Gupta, S. Hillenbrand, T. Kuchler, R. Richmond, and J. Stroebel. International trade and social connectedness. Working Paper 26960, National Bureau of Economic Research, 2020b.

M. Bailey, D. Johnston, M. Koenen, T. Kuchler, D. Russel, and J. Stroebel. Social distancing during a pandemic: The role of friends. Working paper, National Bureau of Economic Research, 2020c.

M. Bailey, D. M. Johnston, T. Kuchler, D. Russel, B. State, and J. Stroebel. The determinants of social connectedness in europe. In *International Conference on Social Informatics*, Lecture Notes in Computer Science. Springer, 2020d.

L. Bengtsson, J. Gaudart, X. Lu, S. Moore, E. Wetter, K. Sallah, S. Rebaudet, and R. Piarroux. Using mobile phone data to predict the spatial spread of cholera. *Scientific reports*, 5:8923, 2015.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

J. K. Brueckner and A. G. Largey. Social interaction and urban sprawl. *Journal of Urban Economics*, 64:18–34, 2008.

K. Büchel and M. V. Ehrlich. Cities and the structure of social interactions: Evidence from mobile phone data. *Journal of urban economics*, 119:103276, 2020.

B. Carey and J. Glanz. Travel from new york city seeded wave of u.s. outbreaks. *New York Times*, 2020. URL https://www.nytimes.com/2020/05/07/us/new-york-city-coronavirus-outbreak.html.

M. Carlisle. Rhode island governor announces national guard will go 'door-to-door' to identify new yorkers to slow covid-19 spread. *Time*, 2020. URL https://time.com/5812069/rhode-island-new-york-coronavirus/.

B. Chappell. Coronavirus: New york creates 'containment area' around cluster in new rochelle. *NPR*, 2020. URL https://www.npr.org/sections/health-shots/2020/03/10/814099444/new-york-creates-containment-area-around-cluster-in-new-rochelle.

B. Charoenwong, A. Kwan, and V. Pursiainen. Social connections with covid-19–affected areas increase compliance with mobility restrictions. *Science Advances*, 6(47), 2020.

R. Chetty, J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter. The opportunity atlas: Mapping the childhood roots of social mobility. Working Paper 25147, National Bureau of Economic Research, 2016.

R. B. Correia, L. Li, and L. M. Rocha. Monitoring potential drug interactions and reactions via network analysis of instagram user timelines. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 492–503. World Scientific, 2016.

V. Couture, J. I. Dingel, A. Green, J. Handbury, and K. Williams. Measuring movement and social contact with smartphone data: a real-time application to covid-19. Working Paper 27560, National Bureau of Economic Research, 2020.

J. Coven, A. Gupta, and I. Yao. Urban flight seeded the covid-19 pandemic across the united states. *Available at SSRN 3711737*, 2020.

L. Danon, A. P. Ford, T. House, C. P. Jewell, M. J. Keeling, G. O. Roberts, J. V. Ross, and M. C. Vernon. Networks and the epidemiology of infectious disease. *Interdisciplinary perspectives on infectious diseases*, 2011, 2011.

Facebook. Facebook form 10-k, 2019 annual report, 2020. URL http://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/45290cc0-656d-4a88-a2f3-147c8de86506.pdf.

Facebook Symptom Survey. URL https://dataforgood.fb.com/tools/symptommap/.

N. Generous, G. Fairchild, A. Deshpande, S. Y. Del Valle, and R. Priedhorsky. Global disease monitoring and forecasting with wikipedia. *PLoS Comput Biol*, 10(11):e1003892, 2014.

J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

S. Gittelman, V. Lange, C. A. G. Crawford, C. A. Okoro, E. Lieb, S. S. Dhingra, and E. Trimarchi. A new source of data for public health surveillance: Facebook likes. *Journal of medical Internet research*, 17(4):e98, 2015.

P. Giuliano and I. Rasul. Compliance with social distancing during the covid-19 crisis, 2020. URL https://voxeu.org/article/compliance-social-distancing-during-covid-19-crisis.

E. L. Glaeser, H. D. Kallal, J. A. Scheinkman, and A. Shleifer. Growth in cities. *Journal of Political Economy*, 100(6):1126–1152, 1992.

Google LLC. Google covid-19 search trends symptoms dataset. URL http://goo.gle/covid19symptomdataset.

D. Holtz, M. Zhao, S. G. Benzell, C. Y. Cao, M. A. Rahimian, J. Yang, J. Allen, A. Collis, A. Moehring, T. Sowrirajan, D. Ghosh, Y. Zhang, P. S. Dhillon, C. Nicolaides, D. Eckles, and S. Aral. Interdependence and the cost of uncoordinated responses to covid-19. *Proceedings of the National Academy of Sciences*, 117(33):19837–19843, 2020.

J. Horowitz, E. Bubola, and E. Povoledo. Italy, pandemic's new epicenter, has lessons for the world. *New York Times*, 2020. URL https://www.nytimes.com/2020/03/21/world/europe/italy-coronavirus-center-lessons.html.

K. Jahanbin and V. Rahmanian. Using twitter and web news mining to predict covid-19 outbreak. *Asian Pacific Journal of Tropical Medicine*, 13, 2020.

J. S. Jia, X. Lu, Y. Yuan, G. Xu, J. Jia, and N. A. Christakis. Population flow drives spatio-temporal distribution of covid-19 in china. *Nature*, pages 1–5, 2020.

M. J. Keeling and K. T. Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, 2005.

M. J. Keeling and P. Rohani. Spatial models. In *Modeling infectious diseases in humans and animals*, pages 232–290. Princeton University Press, 2011.

J. S. Kim, E. Patacchini, P. M. Picard, and Y. Zenou. Urban interactions. *Working Paper*, 2017.

T. Kington. As italy extends quarantine zone, many flee; angry official tell them to go back. *Los Angeles Times*, 2020. URL https://www.latimes.com/world-nation/story/2020-03-08/italy-extends-quarantine-across-north-many-flee.

A. S. Klovdahl. Social networks and the spread of infectious diseases: the aids example. *Social science & medicine*, 21(11):1203–1216, 1985.

A. S. Klovdahl, J. J. Potterat, D. E. Woodhouse, J. B. Muth, S. Q. Muth, and W. W. Darrow. Social networks and infectious disease: The colorado springs study. *Social science & medicine*, 38(1):79–88, 1994.

T. Kuchler, Y. Li, L. Peng, J. Stroebel, and D. Zhou. Social proximity to capital: Implications for investors and firms. Working Paper 27299, National Bureau of Economic Research, 2020.

C. Makridis and T. Wang. Learning from friends in a pandemic: Social networks and the macroeconomic response of consumption. *Available at SSRN 3601500*, 2020.

P. Mossay and P. M. Picard. On spatial equilibria in a social interaction model. *Journal of Economic Theory*, 146(6):2455–2477, 2011.

J. Mossong, N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med*, 5(3):e74, 2008.

L. Mower. New yorkers flying to florida to self-quarantine for 14 days, gov. desantis says. *Tampa Bay Times*, 2020. URL https://www.tampabay.com/news/health/2020/03/23/huge-amounts-of-new-yorkers-flocking-to-florida-gov-desantis-says-in-refusing-lock-down/.

M. E. Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen. Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Comput Biol*, 9(10):e1003256, 2013.

D. Paolotti, A. Carnahan, V. Colizza, K. Eames, J. Edmunds, G. Gomes, C. Koppeschaar, M. Rehn, R. Smallenburg, C. Turbelin, et al. Web-based participatory surveillance of infectious diseases: the influenzanet participatory surveillance experience. *Clinical Microbiology and Infection*, 20(1):17–21, 2014.

P. S. Peixoto, D. Marcondes, C. Peixoto, and S. M. Oliva. Modeling future spread of infections via mobile geolocation data and population dynamics. an application to covid-19 in brazil. *PloS one*, 15(7):e0235732, 2020.

A. P. Piontti, N. Perra, L. Rossi, N. Samay, and A. Vespignani. *Charting the Next Pandemic: Modeling Infectious Disease Spreading in the Data Science Age*. Springer, 2018.

N. G. Reich, L. C. Brooks, S. J. Fox, S. Kandula, C. J. McGowan, E. Moore, D. Osthus, E. L. Ray, A. Tushar, T. K. Yamana, et al. A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the united states. *Proceedings of the National Academy of Sciences*, 116(8):3146–3154, 2019.

M. Rodríguez-Martínez and C. C. Garzón-Alfonso. Twitter health surveillance (ths) system. In *Proceedings: IEEE International Conference on Big Data*, volume 2018, page 1647. NIH Public Access, 2018.

M. S. Smolinski, A. W. Crawley, K. Baltrusaitis, R. Chunara, J. M. Olsen, O. Wójcik, M. Santillana, A. Nguyen, and J. S. Brownstein. Flu near you: crowdsourced symptom reporting spanning 2 influenza seasons. *American journal of public health*, 105(10):2124–2130, 2015.

T. Tully and S. Stowe. The wealthy flee coronavirus. vacation towns respond: Stay away. *New York Times*, 2020. URL https://www.nytimes.com/2020/03/25/nyregion/coronavirus-leaving-nyc-vacation-homes.html.

A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

A. Wesolowski, T. Qureshi, M. F. Boni, P. R. Sundsøy, M. A. Johansson, S. B. Rasheed, K. Engø-Monsen, and C. O. Buckee. Impact of human mobility on the emergence of dengue epidemics in pakistan. *Proceedings of the National Academy of Sciences*, 112(38): 11887–11892, 2015.

E. Williamson and K. Hussey. Party zero: How a soiree in connecticut became a 'super spreader'. *New York Times*, 2020. URL https://www.nytimes.com/2020/03/23/us/coronavirus-westport-connecticut-party-zero.html.

C. Yang, R. Wang, F. Gao, D. Sun, J. Tang, and T. Abdelzaher. Quantifying projected impact of social distancing policies on covid-19 outcomes in the us. *arXiv preprint arXiv:2005.00112*, 2020.

Appendices

A Out-of-Sample Hotspot Analysis

In this section, we conduct a simple prediction exercise to test the value of social connectedness in forecasting how a disease will propagate from a hotspot early in a pandemic. To do so, we train a model using an Italian hotspot and test it using a subsequent U.S. hotspot.

On March 10, 2020, New York state created a "containment area" around New Rochelle, a community in Westchester County that had the first major COVID-19 outbreak in the eastern United States. At this time, U.S. local officials around the country were likely worried about the extent of their own regions' exposures to COVID-19. Yet, in the absence of existing data on domestic COVID-19 spread in the U.S., it may have been difficult to calibrate a local forecasting model. One potential solution would have been to train a model using information from the Italian outbreak — which by mid-March had been spreading for some time — to predict COVID-19 spread in the U.S.

We mimic this use case by first training linear regression and random forest models using data from Italy. Specifically, for each Italian province, we predict the number COVID-19 cases per 10,000 people on March 10 using population density, a measure of income, and distance and social connectedness to the Lodi hotspot. We exclude provinces within 80 km (50 miles) of Lodi. We train versions of these models with and without including social connectedness to Lodi as a predictor. In each model, we normalize every measure by subtracting the mean and dividing by the standard deviation. This normalization, which is common in machine learning prediction applications, ensures that our predictive measures are scaled similarly in the U.S. and Italian settings. In our prediction exercise we use, for each U.S. county, the Italy-trained models to predict the number of COVID-19 cases per 10,000 people on March 30 (i.e., 20 days in the future) with and without data on social connectedness to Westchester.

Table A1 shows that our predictions improve when adding social connectedness to the hotspots as a model input. Row (1) shows that the RMSE drops from 0.99 to 0.97 (in standard deviations from the mean cases per 10,000 residents) for the linear regression model and from 1.04 to 1.01 for the random forest model. Row (2) compares the rank of counties' predictions with the true rank. For both models the rank-rank correlations increase when including social connectedness to the hotspot as a predictor.

Table A1: Predicting U.S. Hotspot COVID-19 spread, trained on Italian Hotspot spread

	Linear Regression			Random Forest		
	(1)	(2)	(3)	(4)	(5)	(6)
	Without SCI to Hotspot	With SCI to Hotspot	Diff. from SCI to Hotspot	Without SCI to Hotspot	With SCI to Hotspot	Diff. from SCI to Hotspot
(1) RMSE	0.990	0.972	-0.018	1.041	1.010	-0.031
(2) Rank-Rank Corr. w/ Truth	0.238	0.350	0.112	0.254	0.315	0.061

Note: Table shows results from county-level predictions of COVID-19 cases per 10,000 residents. Columns 1-3 and 4-6 show results from linear regression and random forest models, respectively. The models are trained using information from Italy on March 10 and tested using information from the U.S. on March 30. All measures are normalized by subtracting the mean then dividing by the standard deviation. Row (1) shows the prediction root mean squared errors (RMSEs) and row (2) shows prediction rank-rank correlation with the truth. The model inputs in columns 1 and 4 are $log(distance)$ to the hotspot (Lodi in Italy, Westchester in the U.S.), population density, and household income / (GDP per inhabitant). Columns 2 and 5 add $log(SCI)$ to the hotspots. Columns 3 and 6 show the change in each measure from adding $log(SCI)$.

B Additional Time Series Regressions

Table A2 shows the results from running the regression in Table 1's column 8 separately for every two-week period betwen March 30 and November 2. In every period, a two week lagged measure of social proximity to cases was a statistically significant predictor of actual case growth. The magnitudes of the coefficients suggest a doubling in social proximity to cases in one two-week period corresponds to between a 10.9% and 66.4% increase in actual cases in the next period, after controlling for physical proximity to cases and other controls.

Table A3 shows the results from adding additional predictive measures to Table 1. Columns 1 and 2 are the same as columns 7 and 8 in Table 1. Columns 3 and 4 show that a "nowcast" of changes in Google symptom search trends is strongly correlated with changes in case growth. Columns 5 and 6 show that this relationship persists, though somewhat less strongly, for a one-period lagged measure of changes in symptom searches. The latter measure, unlike the former, could be used to predict *future* case growth. Columns 7 and 8 show that the change in LEX proximity to cases in one period also has a strong positive relationship with actual case growth in the next. Even in the presence of each of these other measures, social proximity to cases remains a significant predictor of future case growth.

Table A2: COVID-19 Case Growth and Prior Proximity to Cases, by Two-Week Period

	log(Change in Cases per 10k Residents + 1)														
	March 31 - April 13	April 14 - April 27	April 28 - May 11	May 12 - May 25	May 26 - June 8	June 9 - June 22	June 23 - July 6	July 7 - July 20	July 21 - Aug. 10	Aug. 11 - Aug. 24	Aug. 25 - Sep. 7	Sep. 8 - Sep. 21	Sep. 22 - Oct. 5	Oct. 6 - Oct. 19	Oct. 20 - Nov. 2
2 Week Lag:	0.735***	0.411***	0.150**	0.204***	0.580***	0.178**	0.287***	0.225***	0.243***	0.305***	0.314***	0.152**	0.371***	0.149**	0.313***
log(Change in Social Proximity to Cases + 1)	(0.093)	(0.088)	(0.060)	(0.061)	(0.062)	(0.074)	(0.057)	(0.067)	(0.072)	(0.077)	(0.080)	(0.068)	(0.073)	(0.069)	(0.069)
4 Week Lag:	0.339	-0.190	0.157*	0.053	-0.116*	0.196***	0.057	0.097	0.084	0.024	-0.046	0.041	0.128*	0.049	-0.141**
log(Change in Social Proximity to Cases + 1)	(0.434)	(0.127)	(0.082)	(0.060)	(0.062)	(0.074)	(0.058)	(0.062)	(0.069)	(0.075)	(0.083)	(0.072)	(0.070)	(0.068)	(0.065)
Share of Friends within 50 Miles	0.250	-0.167	0.069	-0.180	-0.218	0.039	0.484*	-0.424*	0.774***	0.203	0.060	-0.547**	-0.455*	0.398*	0.776***
	(0.247)	(0.292)	(0.278)	(0.272)	(0.261)	(0.281)	(0.262)	(0.251)	(0.232)	(0.253)	(0.267)	(0.259)	(0.250)	(0.229)	(0.214)
Share of Friends within 100 Miles	0.066	0.657*	0.259	0.913***	0.191	-0.043	-0.514*	0.824***	-0.213	0.300	-0.018	0.629**	0.845***	-0.475*	-0.655***
	(0.284)	(0.336)	(0.320)	(0.311)	(0.298)	(0.322)	(0.300)	(0.286)	(0.264)	(0.285)	(0.301)	(0.292)	(0.282)	(0.258)	(0.243)
2 Week Lag:	1.125***	0.486	2.089***	1.207***	-0.112	2.281***	1.401***	1.821***	2.129***	2.098***	1.977***	0.985*	2.723***	4.118***	3.876***
log(Change in Physical Proximity to Cases + 1)	(0.189)	(0.388)	(0.284)	(0.256)	(0.316)	(0.436)	(0.355)	(0.428)	(0.607)	(0.761)	(0.752)	(0.521)	(0.637)	(0.456)	(0.480)
4 Week Lag:	-2.193***	-0.156	-1.686***	-1.072***	0.429	-2.705***	-1.551***	-1.802***	-2.127***	-1.929**	-1.824**	-0.949*	-2.452***	-3.748***	-3.664***
log(Change in Physical Proximity to Cases + 1)	(0.724)	(0.430)	(0.289)	(0.274)	(0.287)	(0.443)	(0.332)	(0.405)	(0.618)	(0.806)	(0.768)	(0.544)	(0.633)	(0.448)	(0.496)
2 Week Lag:	0.172***	0.381***	0.554***	0.463***	0.276***	0.361***	0.328***	0.346***	0.371***	0.315***	0.283***	0.423***	0.255***	0.367***	0.327***
log(Change in Cases per 10k Residents + 1)	(0.059)	(0.050)	(0.037)	(0.036)	(0.036)	(0.041)	(0.033)	(0.036)	(0.037)	(0.041)	(0.042)	(0.035)	(0.037)	(0.035)	(0.035)
4 Week Lag:	-0.083	0.124*	-0.026	0.046	0.128***	-0.005	0.003	0.016	0.004	0.033	0.131***	0.056	0.014	0.076**	0.155***
log(Change in Cases per 10k Residents + 1)	(0.247)	(0.075)	(0.047)	(0.037)	(0.035)	(0.040)	(0.033)	(0.034)	(0.036)	(0.040)	(0.043)	(0.039)	(0.036)	(0.033)	(0.033)
Pop. Density FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Median Household Income FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
State FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Sample Mean	1.239	1.257	1.334	1.372	1.429	1.586	2.038	2.530	2.707	2.675	2.627	2.629	2.840	3.040	3.356
R-Squared	0.608	0.574	0.644	0.648	0.666	0.615	0.673	0.705	0.732	0.657	0.606	0.617	0.637	0.680	0.713
N	3,135	3,135	3,135	3,135	3,135	3,135	3,135	3,135	3,135	3,135	3,135	3,135	3,135	3,135	3,135

Note: Table shows time-specific results from regression 5. Each observation is a county. The dependent variable is log of one plus the number of new COVID-19 cases per 10,000 residents in one two-week period between March 30 and November 2, 2020. All columns include log of growth in social and physical proximity to cases, as well as log of growth in actual cases, lagged by two and four weeks (one and two time periods). All columns include time-specific fixed effects for percentiles of population density and median household income, time-specific fixed effects for state, and estimations of the share of a county's Facebook connections that are within 50 and 150 miles Significance levels: *(p<0.10), **(p<0.05), ***(p<0.01).

Table A3: COVID-19 Case Growth, Prior Proximity to Cases, and Other Predictive Measures

	log(Change in Cases per 10k Residents + 1)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
2 Week Lag: log(Change in Social Proximity to Cases + 1)	0.414*** (0.041)	0.321*** (0.037)	0.362*** (0.047)	0.277*** (0.039)	0.351*** (0.047)	0.270*** (0.039)	0.141*** (0.050)	0.141*** (0.045)
4 Week Lag: log(Change in Social Proximity to Cases + 1)	-0.002 (0.036)	0.010 (0.032)	-0.008 (0.042)	0.022 (0.035)	0.001 (0.042)	0.027 (0.035)	-0.039 (0.051)	-0.014 (0.050)
Google searches related to Fever (% Change)			0.286*** (0.024)	0.231*** (0.022)				
Google searches related to Cough (% Change)			0.158*** (0.021)	0.117*** (0.018)				
Google searches related to Fatigue (% Change)			0.022 (0.019)	0.021 (0.019)				
2 Week Lag: Google searches related to Fever (% Change)					0.165*** (0.021)	0.123*** (0.018)		
2 Week Lag: Google searches related to Cough (% Change)					0.189*** (0.021)	0.139*** (0.017)		
2 Week Lag: Google searches related to Fatigue (% Change)					0.013 (0.019)	0.019 (0.018)		
2 Week Lag: log(Change in LEX Proximity to Cases + 1)							0.269*** (0.025)	0.179*** (0.022)
4 Week Lag: log(Change in LEX Proximity to Cases + 1)							0.006 (0.023)	0.006 (0.022)
Share of Friends within 50 Miles	0.050 (0.100)	0.076 (0.082)	0.006 (0.091)	0.084 (0.075)	0.010 (0.092)	0.086 (0.076)	0.036 (0.093)	0.022 (0.081)
Share of Friends within 150 Miles	-0.256** (0.124)	0.143 (0.109)	-0.213* (0.110)	0.168* (0.097)	-0.220* (0.112)	0.171* (0.098)	-0.126 (0.115)	0.287*** (0.101)
2 Week Lag: log(Change in Physical Proximity to Cases + 1)	1.244*** (0.118)	1.388*** (0.176)	1.116*** (0.105)	1.272*** (0.167)	1.117*** (0.105)	1.261*** (0.168)	0.971*** (0.104)	1.104*** (0.170)
4 Week Lag: log(Change in Physical Proximity to Cases + 1)	-1.037*** (0.121)	-1.225*** (0.187)	-0.915*** (0.108)	-1.077*** (0.178)	-0.912*** (0.108)	-1.066*** (0.179)	-0.852*** (0.107)	-0.996*** (0.180)
2 Week Lag: log(Change in Cases per 10k Residents + 1)	0.372*** (0.022)	0.351*** (0.019)	0.498*** (0.024)	0.467*** (0.019)	0.484*** (0.024)	0.456*** (0.019)	0.536*** (0.023)	0.510*** (0.020)
4 Week Lag: log(Change in Cases per 10k Residents + 1)	0.071*** (0.019)	0.056*** (0.017)	0.027 (0.021)	0.013 (0.017)	0.034* (0.021)	0.019 (0.017)	-0.005 (0.022)	0.001 (0.019)
Time x Pop. Density FEs	Y	Y	Y	Y	Y	Y	Y	Y
Time x Median Household Income FEs	Y	Y	Y	Y	Y	Y	Y	Y
Time x State FEs		Y		Y		Y		Y
Sample Mean	2.177	2.177	2.279	2.279	2.279	2.279	2.333	2.333
R-Squared	0.725	0.757	0.768	0.800	0.767	0.799	0.795	0.827
N	47,040	47,025	38,520	38,520	38,520	38,520	30,210	30,195

Note: Table shows results from regression 5. Each observation is a county × two-week period (between March 30 and November 2, 2020). The dependent variable is log of one plus the number of new COVID-19 cases per 10,000 residents. Columns 1 and 2 are the same as columns 7 and 8 in Table 1. Columns 3 and 4 add the percent growth in Google searches related to fever, cough, and fatigue from the week prior to the period to the second week of the period. Columns 5 and 6 includes analogous measures lagged by one period. Columns 7 and 8 add LEX-based proximity to cases. Standard errors are clustered at the time × state level. Significance levels: *(p<0.10), **(p<0.05), ***(p<0.01).

C Out-Of-Sample Prediction: COVID-19 Deaths

Table A4 show the prediction error from a simple out-of-sample forecasting exercise analogous to that in Table 2, but for COVID-19 deaths instead of COVID-19 cases. Columns 1-3 use

the variables in columns 7 and 8 of Table 1, Panel B as predictors. The model presented in column 2 includes the two lagged measures of social proximity to COVID-19 deaths, while the model presented in column 1 does not include them. The RMSE is lower in every period in the model that includes social proximity to deaths as a predictor. Columns 4-6 include predictions for all counties included in the COVID-19 case data. We make a prediction using the "best" model available (in terms of number of model features), as described in Section 3. Column 6 shows that the RMSE decreases again in every period when including social proximity to deaths as a predictor. Columns 7-9 limit to counties with Google and LEX data. Similar to our findings for county-level COVID-19 cases, social connectedness does not appear to provide substantial additional predictive value in this particular setting.

Table A4: Predicting COVID-19 deaths in U.S., with and without Social Proximity to Deaths

	RMSE: Baseline Model			RMSE: Best Available Model			RMSE: Counties w/ Google + LEX Only		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Without Social Proximity	With Social Proximity to Cases	Diff. from Social Proximity	Without Social Proximity	With Social Proximity to Cases	Diff. from Social Proximity	Without Social Proximity	With Social Proximity to Cases	Diff. from Social Proximity
(1) May 12 - June 8	0.765	0.731	-0.034	0.709	0.690	-0.019	0.709	0.711	0.002
(2) June 9 - July 6	0.480	0.435	-0.045	0.438	0.402	-0.036	0.356	0.339	-0.017
(3) July 7 - Aug. 10	0.457	0.453	-0.004	0.455	0.451	-0.003	0.431	0.428	-0.002
(4) Aug. 11 - Sep. 7	0.471	0.462	-0.008	0.458	0.454	-0.003	0.400	0.401	0.001
(5) Sep. 8 - Oct. 5	0.488	0.469	-0.019	0.475	0.460	-0.015	0.371	0.365	-0.006
(6) Oct. 6 - Nov. 2	0.549	0.543	-0.006	0.544	0.539	-0.005	0.405	0.406	0.001

Note: Table shows results from county-level predictions of COVID-19 deaths. The predicted outcome is log of one plus the number of new COVID-19 deaths per 10,000 residents. All columns show root mean squared errors (RMSEs) from a random forest model trained on data from all periods prior to the period of interest. The model inputs in column 1 are population density; median household income; and log of growth in physical proximity to deaths and actual deaths, lagged by four and eight weeks (one and two time periods). Columns 4 and 7 include information on one and two period lagged measures of LEX proximity to deaths, and one period lagged percent changes in Google searches related to fever, cough, and fatigue. Column 4 includes predictions for 3,156 counties using, for each county, a model that utilizes the most available information possible. Column 7 limits to 1,976 counties for which we have both Google symptom search and LEX data. Columns 2, 5, and 8 add lagged measures of social proximity to deaths to columns 1, 4, and 7. Columns 3, 6, and 9 show the change in RMSE from adding social proximity to deaths.

D Additional Details on Google Symptom Search Data

In Section 3, Appendix B, and Appendix C we use data on Google searches related to COVID-19 symptoms from Google LLC. The data include a county by week normalized (within county) probability a user will make a symptom-related search. These measures come in a daily series and a weekly series. If a given symptom in a given county does not meet certain Google quality or privacy thresholds at the daily level, it will be provided at

the weekly level. If it cannot meet these thresholds at the weekly level, it is not provided. If a county/symptom is provided for a period at the daily level, it is not also provided at the weekly level. To create a time series at the county/weekly level, we therefore average non-missing daily measures by week. We use searches related to three common COVID-19 symptoms: fever, cough, and fatigue.

To aggregate our measure to the bi-weekly periods used in Section 3, Appendix B, and Appendix C we define the change in searches related to a symptom as the percent change in the probability between the second week of the period and the second week of the previous period. For example, for the period March 31 - April 13, the percent change is from March 23 - March 29 to April 6 - 12. In our prediction exercise presented in Table 2, we use a one-period lagged version of this measure. Some counties are missing data for particular weeks. So that our final sample has a complete time series for every included county, we exclude counties with missing data in more than half the periods in our sample and impute zero change in any remaining missing periods. 97 percent of counties included are missing fewer than 1 in 4 periods and 77 percent are missing fewer than 1 in 20 periods.