

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

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

We use anonymized and aggregated data from Facebook to show that areas with stronger social ties to two early COVID-19 hotspots (Westchester County, NY, in the U.S. and Lodi province in Italy) generally have more confirmed COVID-19 cases as of March 30, 2020. These relationships hold after controlling for geographic distance to the hotspots as well as for the income and population density of the regions. These results suggest that data from online social networks may prove 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). Yet, the geographic structure of social networks and interactions is usually hard to measure on a national or global scale. In Bailey et al. (2018b), we showed how data from online social networking services can be used to measure and understand the geographic structure of social networks. We introduced a new data set, the *Social Connectedness Index*, which captures the relative probability that individuals across two regions are connected through a friendship link on Facebook, a global online social network. At the time, we suggested that such a measure of the geographic structure of social networks may be helpful to epidemiologists hoping to forecast the spread of communicable diseases. The idea was that two regions connected through many friendship links are likely to see more physical interactions between their residents, providing increased opportunities for the spread of communicable diseases.

In this note, we explore the relationship between the geographic spread of COVID-19 and the geographic structures of social networks in the United States and in Italy. We show that regions with stronger social ties to early COVID-19 “hotspots” in each country — Westchester County, NY, in the United States, and Lodi province in Italy — have more

*Date: April 8, 2020. The data on social connectedness data used in this paper (as well as similar data for a wide range of other geographies) are accessible to other researchers by emailing sci_data@fb.com.

[†]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)

documented COVID-19 cases per resident as of March 30, 2020. This relationship is robust to controlling for the geographic distance to these early “hotspots”, as well as a number of demographic characteristics of the regions.

Our objective is not to incorporate social connectedness data into a state-of-the-art epidemiological model, but instead to provide a “proof of concept” by highlighting that social connectedness as measured by our *Social Connectedness Index* is correlated with COVID-19 prevalence in a statistically meaningful way. This finding suggests to us that the geographic structure of social network as measured by Facebook may indeed provide a useful proxy for the type of social interactions that epidemiologists have long known to contribute to the spread of communicable diseases.¹ We thus hope that the *Social Connectedness Index* can help epidemiologists with forecasting the spread of communicable diseases, in particular given that these data are easily accessible to researchers by emailing sci_data@fb.com.

Data Description

To measure the intensity of social connectedness between locations, we use an anonymized snapshot of all active Facebook users and their friendship networks from March 2020. 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 Asian-Pacific, and 817 billion in the rest of the world. The data therefore have extremely wide coverage, and provide 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. Our measure of the social connectedness between two locations i and j is provided by the *Social Connectedness Index (SCI)* introduced by [Bailey et al. \(2018b\)](#):

$$SocialConnectedness_{ij} = \frac{FB_Connections_{ij}}{FB_Users_i * FB_Users_j}. \quad (1)$$

Here, $FB_Connections_{ij}$ is the total number of Facebook friendship links between Facebook users living in location i and Facebook users living in location j .² FB_Users_i and FB_Users_j are the number of active users in each location. $SocialConnectedness_{ij}$ thus measures the

¹A number of other researchers have explored how different aspects of social media and internet-usage patterns can be used for tracking and preventing disease (for one overview see [Aiello et al., 2020](#)). Our hope is to add to this exciting work.

²Establishing 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.

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 of social connectedness is useful for describing real-world social networks. We also documented that it 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., 2020a), patent citations (Bailey et al., 2018b), travel flows (Bailey et al., 2019b, 2020b), and investment decisions (Kuchler et al., 2020). More generally, we have found that information on individuals’ Facebook friendship links can help understand their product adoption decisions (Bailey et al., 2019c) and their housing and mortgage choices (Bailey et al., 2018a, 2019a).

In the next section, we use these data to 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 COVID-19 outbreak in the eastern United States (NPR, March 10, 2020). As of March 20th, the county had over 9,300 cases, second only to nearby New York City. Additionally, a number of articles have reported wealthy residents from Westchester and the New York area fleeing to other parts of the U.S. (New York Times, March 25, 2020), providing a vector that could potentially spread the disease across the country. Social connections to Westchester may thus provide particularly important information for tracking the spread of COVID-19, especially if individuals’ travel patterns follow their social networks, as suggested by Bailey et al. (2019b, 2020b). 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 has been at the center of Italy’s outbreak (New York Times, March 21, 2020).

Data on confirmed COVID-19 cases in the United States by county come from Johns Hopkins University Center for Systems Science and Engineering.³ Similarly, data for confirmed COVID-19 cases for each Italian province come from the Italian Dipartimento della Protezione Civile.⁴ We use data from March 30th, 2020, but our results are robust to using data from prior days. At this stage it is important to note that, as with any data on confirmed cases, some bias may be introduced by differential testing across regions.

³Available at <https://github.com/CSSEGISandData/COVID-19>.

⁴Available at <https://github.com/pcm-dpc>. Because Italian provinces on the island of Sardinia do not align with current European NUTS3 regions (the level at which we aggregate social connectedness), we include Sardinia as a single observation throughout our analysis.

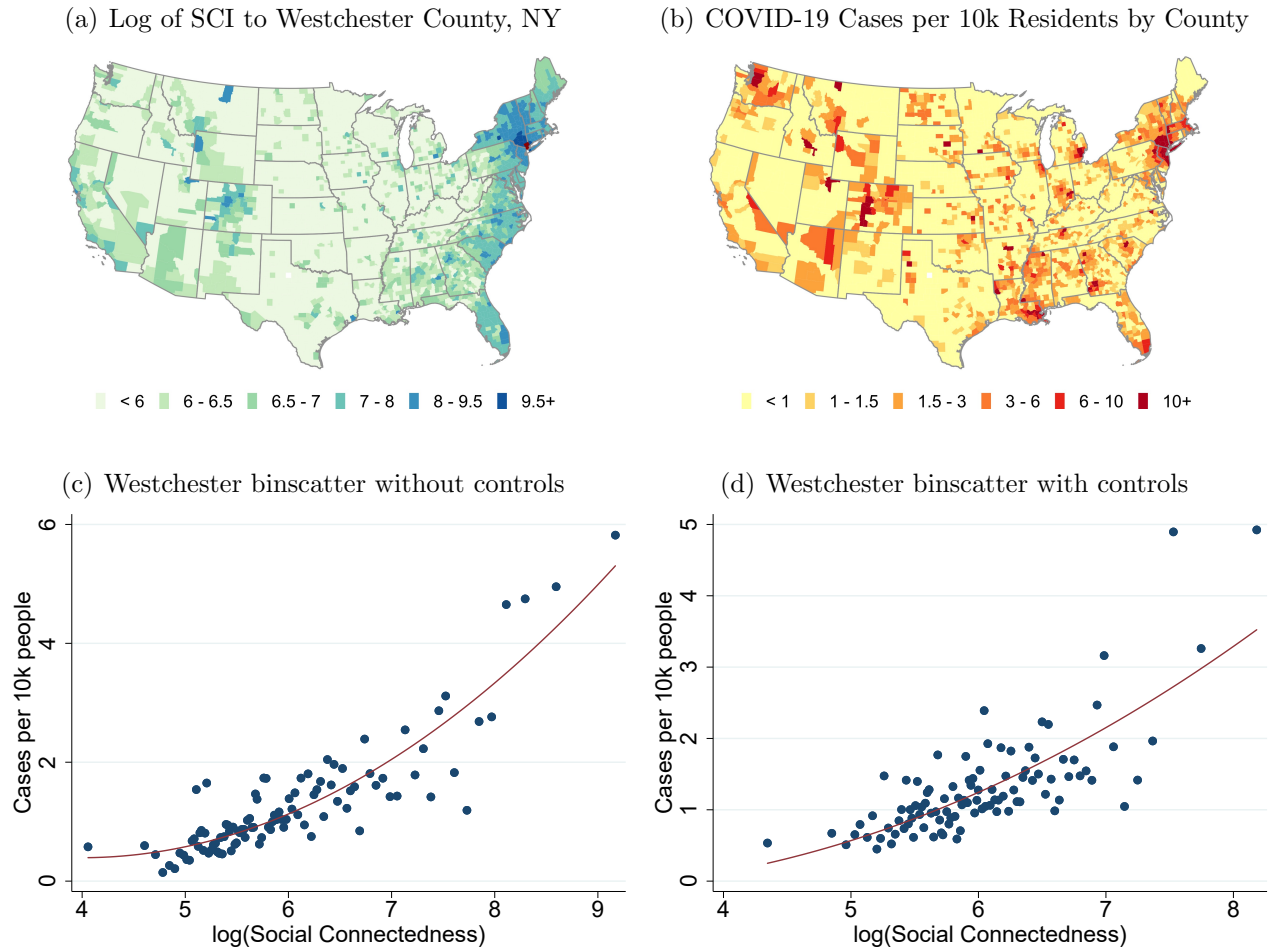
Results

Panel (a) of Figure 1 shows a heatmap of the social connectedness of Westchester County, NY, to all 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, again 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 emerge as well. Both measures are high in the communities along the coasts of Florida (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 have both publicly lamented the number of New York area residents fleeing to their states and spreading COVID-19 ([Tampa Bay Times, March 23, 2020](#); [Time, March 28, 2020](#)). By contrast, many areas that are geographically closer but less socially connected to Westchester, such as in western Pennsylvania and West Virginia, have fewer confirmed COVID-19 cases. There are also a number of patterns of COVID-19 prevalence that connectedness to Westchester alone cannot explain. Areas surrounding King County, WA (Seattle), for example, have relatively low levels of connectedness to Westchester, but were an independent early hotspot of COVID-19. Some states in the southern U.S. where residents were slower to limit travel also have higher case densities than would be predicted purely by social connectedness to Westchester ([New York Times, April 2, 2020](#)).

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 COVID-19 prevalence and social ties to Westchester. 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

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



Note: Panel (a) shows the social connectedness to Westchester for U.S. counties. Panel (b) shows the number of confirmed COVID-19 cases by U.S. county on March 30th, 2020. Panels (c) and (d) show binscatter plots with provinces more than 50 miles from Westchester 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. We then group $\log(SCI)$ into 100 equal-sized bins and plot the average $\log(SCI)$ 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 distance to Westchester from the Nation Bureau of Economic Research; population density and median household income made available from (Chetty et al., 2016); and dummies for the six National Center for Health Statistics Urban-Rural county classifications.

picking up other factors that affect the spread of COVID-19, and that are correlated with social connectedness. 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 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.

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. We chose the current set of control variables to 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. The observed increase in predictive power thus suggests that the *Social Connectedness Index* might serve as a valuable measure above some existing proxies for social interactions.⁵

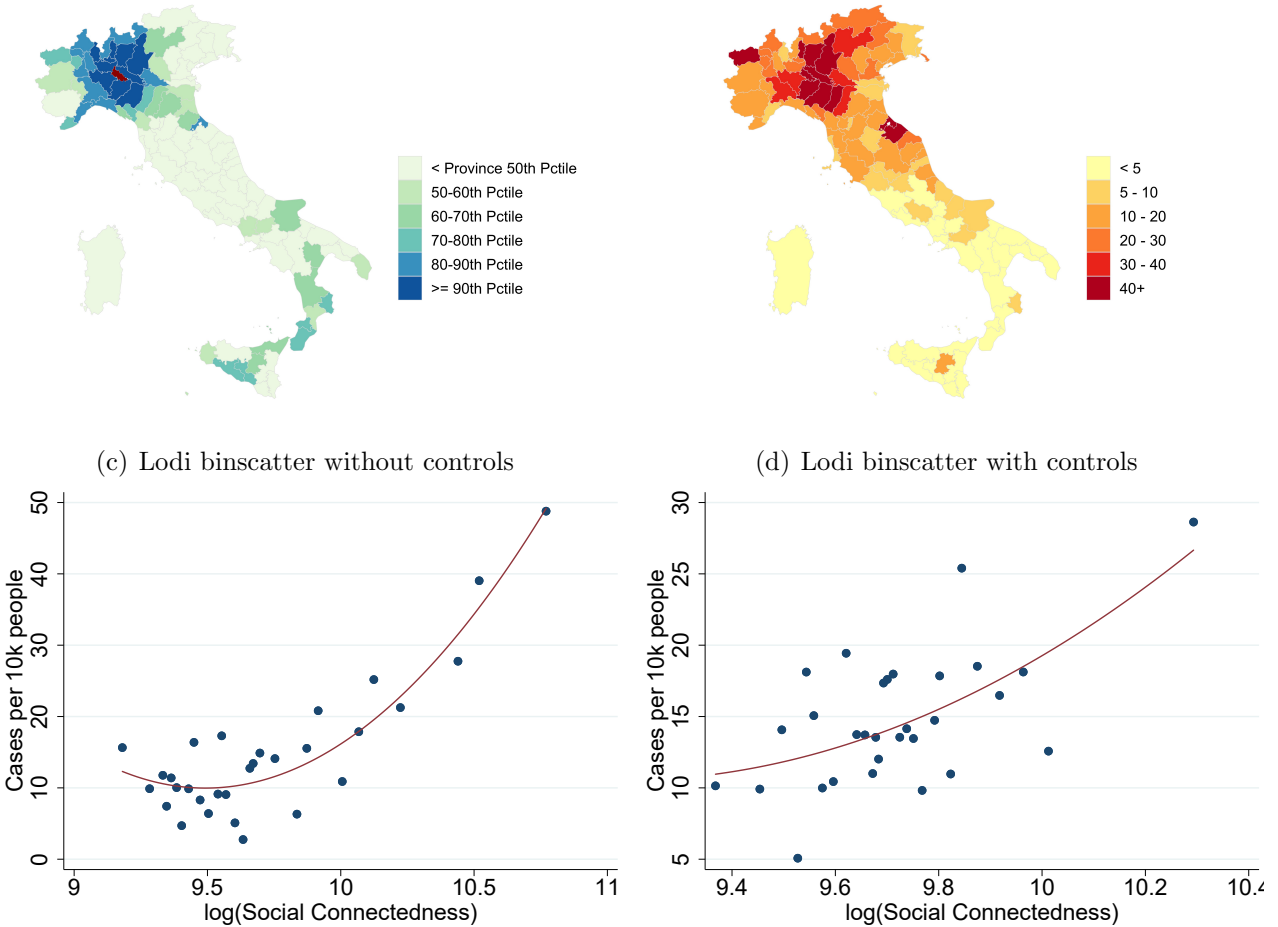
Figure 2 explores the analogous relationships for Lodi province in Italy. The provinces with highest COVID-19 case densities and connectedness to Lodi are in the surrounding Lombardy region, as well as the nearby Piemonte and Veneto regions. There are also relatively high levels of both connectedness 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 disproportionately larger, perhaps reflecting the efforts of Italian authorities to

⁵This is not to suggest that the *Social Connectedness Index* is the only such measure, and we believe that further advances can be made using other data sources, such as cell-phone location pings. But the social connectedness index has a number of advantages, including the fact that it is easily accessible to researchers and consistently available for a large number of global geographies.

Figure 2: 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



Note: Panel (a) shows a measures of Social Connectedness to Lodi for Italian provinces. Panel (b) shows the number of confirmed COVID-19 cases by Italian province on March 30th, 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 smaller 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 the quantile of the province distance to Lodi; GDP per inhabitant; and population density.

restrict the movement of individuals ([LA Times, March 8, 2020](#)). Panels (c) and (d) repeat the binscatter exercise from Figure 1. We exclude provinces within 50 kilometers. In Panel (d) we control for geographic distance using 20 dummies for the quantile of the distance from each province to Lodi, as well as GDP per inhabitant and population density. Again 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, a doubling of *SCI* corresponds to an increase of 16.6 COVID-19 cases per 10,000 residents after controlling for these relevant factors. The incremental R-Squared of including social connectedness to Lodi over the other control variables is 0.057.

Caveats

It is important, at this stage, to re-emphasize that we are not epidemiologists, and that the goal of this note is not to provide an epidemiological model of the spread of COVID-19.⁶ In normal times we would not venture this far from our primary area of expertise and study the the spread of a disease like COVID-19. Indeed, in [Bailey et al. \(2018b\)](#), we explicitly proposed the modeling of communicable diseases as a potentially fruitful direction for others to pursue, without attempting any such modeling ourselves. However, these are not normal times, and we have spent much of the last few years exploring these data on the geographic structure of social networks. In the process, we have found them to be extremely useful for understanding a large number of social and economic relationships such as trade patterns, patent citations, and travel flows. Given the urgency of the current global health crisis, we hope that our expertise in measuring social networks can therefore contribute to the worldwide interdisciplinary research effort to better understand COVID-19.

In particular, we hope that some of the initial patterns we document in this note — together with our earlier work showing how social connections as measured by Facebook can explain many important social and economic phenomena — might be sufficiently striking to epidemiologists such that they would want to incorporate the *Social Connectedness Index* data in their own work. For example, the availability of zip-code level data on social connectedness in the United States as well as similar data for many countries around the world will allow for more detailed modeling as COVID-19 case data becomes available at that level of geographic disaggregation. We would be excited to work with any interested team to help them get the most out of the *Social Connectedness Index* data.

⁶As a result, please also excuse our woefully incomplete-to-non existing review of the large related literature in epidemiology and related fields. We are grateful for any guidance in this direction.

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.
- 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, P. Farrell, T. Kuchler, and J. Stroebel. Social connectedness in urban areas. Working Paper 26029, National Bureau of Economic Research, 2019b.
- 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, 2019c.
- M. Bailey, A. Gupta, S. Hillenbrand, T. Kuchler, R. Richmond, and J. Stroebel. International trade and social connectedness. Working paper, 2020a.
- M. Bailey, T. Kuchler, D. Russel, B. State, and J. Stroebel. Social connectedness in europe. Working paper, 2020b.
- R. Chetty, J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter. The opportunity atlas: Mapping the childhood roots of social mobility. *National Bureau of Economic Research Working Paper No. 25147*, 2016.
- T. Kuchler, L. Peng, J. Stroebel, Y. Li, and D. Zhou. Social proximity to capital: Implications for investors and firms. Working paper, 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.