

# Saving Lives versus Saving Livelihoods: Can Big Data Technology Solve the Pandemic Dilemma?\*

Kairong Xiao<sup>†</sup>

April 30, 2020

## Abstract

This paper studies the effectiveness of big data technology in mitigating the economic and health impacts of the COVID-19 outbreak. I exploit the staggered implementation of contact-tracing apps called “health code” in 322 Chinese cities during the COVID-19 pandemic. Using high-frequency variations in population movements and greenhouse emission across cities before and after the introduction of health code, I disentangle the effect of big data technology from confounding factors such as public sentiments and government responses. I find that big data technology significantly improves the tradeoff between human toll and economic costs. Cities adopt health code experience a significant increase in economic activities without suffering from higher infection rates. Overall, big data technology creates an economic value of 0.5%-0.75% of GDP during the COVID-19 outbreak in China.

*Keywords:* Big data, Coronavirus

---

\*I benefit from valuable comments by Stijn van Nieuwerburgh. I thank Meichen Qian for the excellent research assistance.

<sup>†</sup>Assistant Professor of Business, Columbia Business School. [kx2139@gsb.columbia.edu](mailto:kx2139@gsb.columbia.edu)

# 1 Introduction

Pandemics such as COVID-19 present an impossible choice to policymakers between saving lives and saving livelihoods. On the one hand, population movement restrictions such as social distancing and lockdown are deemed necessary to contain the rapid spread of the disease. On the other hand, such restrictions inflict steep economic costs as normal activities are disrupted. The painful tradeoff between human toll and economic costs has led to heated and sometimes divisive debate in the policy domain.

How can we solve the pandemic dilemma between saving lives and saving the economy? Many believe that the answer lies in big data technology (Ferretti, Wymant, Kendall, Zhao, Nurtay, Abeler-Dörner, Parker, Bonsall, and Fraser, 2020). Using the enormous amount of real-time location data produced by smartphones, we may detect potential carriers of the disease and break the transmission chains. At the same time, big data technology can also identify the group of people who are unlikely to carry the disease so that they can resume normal work and life, which limits the economic damage of the disease. Advocates often cite the successful experiences in China and South Korea where big data technology was aggressively deployed to combat the virus.<sup>1</sup>

However, big data technology is also highly controversial. Critics argue that some countries such as Singapore have seen little success from using contact-tracing apps.<sup>2</sup> Implementing big data technology could also divert critical resources from proven containment methods such as aggressive testing. Big data technology may also disproportionately impact the rights of those under- or misrepresented by the data.<sup>3</sup> Finally, big data technology also raises concerns about privacy infringement and government surveillance. Therefore, using

---

<sup>1</sup>See Forbes April 9, 2020 article “The Vital Role Of Big Data In The Fight Against Coronavirus”.

<sup>2</sup>See WSJ April 22, 2020 article, “Singapore Built a Coronavirus App, but It Hasn’t Worked So Far”.

<sup>3</sup>For instance, the April 12, 2020, AP News article “Europe eyes smartphone location data to stem virus spread” reports that Israeli government’s cell phone location-tracking program has caused complaints that the authorities are erroneously confining people to their homes based on inaccurate location data.

big data to address the public health crisis can potentially do more harm than good.

This paper sheds light on this debate by studying the effectiveness of big data technology in mitigating the economic and human costs of the COVID-19 outbreak in China. I exploit the staggered implementation of contact-tracing apps called “health code” in 322 Chinese cities amid the COVID-19 pandemic. Using high-frequency variations in population movements and greenhouse emission across cities, I disentangle the effect of big data technology from confounding factors such as public sentiments and government responses. I show that big data technology allows the vast majority of the population to resume economic activities without risking the public health condition. The estimated benefits of this technology seem to dominate the potential costs related to privacy.

I start by describing the institutional background. On February 9, 2020, 17 days after the lockdown of Wuhan, the first “health code” was developed by Ant Financial, a FinTech company affiliated with Alibaba, and adopted by the Hangzhou municipal government, where Ant Financial’s headquarter locates. This app uses real-time location data produced by smartphones to predict holders’ risks of being infected based on whether the holders are in close contact with confirmed patients. This app assigns a QR code for each holder, which functions as a “traffic permit” within the city. Holders can travel in the city freely if they obtain green codes but face quarantine if their codes are yellow or red. Health code was subsequently expanded to other cities in China. By the end of March 31, all the cities in China have implemented this system. The adoption of health code represents the largest experiment of big data technology in the public health domain. It offers an invaluable opportunity to examine the effectiveness of big data technology in mitigating economic and human costs inflicted by pandemics.

I collect the adoption dates of health code in each of 322 cities from local government websites and news reports. To measure high-frequency variations in economic activities at the city level, I use within-city population movements constructed from smartphone locations

by Baidu and daily emission of greenhouse gases related to industrial activities. I also use daily numbers of confirmed COVID-19 cases, cured patients, and death tolls for each city provided by the Chinese Center for Disease Control and Prevention (CCDC). The sample period spans January 1 to March 31, 2020, covering the lockdown of Wuhan on January 23, 2020, and the introduction of the first health code in Hangzhou on February 9, 2020. The staggered implementation of health code across cities allows me to disentangle the causal effects of the big data technology from confounding factors.

The empirical analysis yields three main results. First, I find that the introduction of health code significantly increases economic activities by 2-3%. Second, cities that implemented health code attract greater population inflows and experience smaller population outflows. Third, I find that the increase in economic activities does not lead to an increase in infection rates in cities with health code. Overall, the use of big data technology has significantly improved the tradeoff between economic activities and public health, creating an economic value of 0.5%-0.75% GDP during the COVID-19 outbreak in China.

One potential concern on the empirical approach is that cities that adopt health code early could have higher economic importance for the country, thus are forced to reopen before other cities. I address this concern by matching the treated cities to control cities with similar pre-COVID-19 economic activities and find the results are robust. One may also worry that the timing of implementation of health code may be correlated with the successful containment of the outbreak in a city. I address this concern by matching the treated cities to control cities with similar active cases when health code was introduced. The results are also robust to this alternative matching scheme.

Finally, I compare the estimated economic benefits of big data technology with potential costs. I find that the introduction of health code creates an economic value of \$50-75\$ per capita. Comparing this estimate to the value of privacy estimated in the literature (Athey, Catalini, and Tucker, 2017; Tang, 2019), I find that the benefits of big data technology seem

to outweigh the potential costs on privacy.

Why is big data technology an effective tool to mitigate the economic and human costs of pandemics? The answer is that it can address the key amplification mechanism of pandemics: information frictions. Because of the hidden virus, people are afraid of going out, which brings offline consumption to a standstill. Governments have to impose quarantines on the whole population just to stop a few hidden carriers. In such a situation, big data technology can be a powerful tool. By leveraging the enormous amount of data produced in our digital age, big data technology can help to identify the hidden carriers, thus contains the transmission of the virus. Furthermore, it can reduce people's fear of being infected, thus restores economic activities depressed by pandemics.

This paper contributes to the fast-growing literature on the optimal policy response to the pandemic shock (Alvarez, Argente, and Lippi, 2020; Barro, Ursúa, and Weng, 2020; Correia, Luck, and Verner, 1918; Eichenbaum, Rebelo, and Trabandt, 2020; Hall, Jones, and Klenow, 2020; Dewatripont, Goldman, Muraille, and Platteau, 2020; Fang, Wang, and Yang, 2020; Piguillem, Shi, et al., 2020; Jones, Philippon, and Venkateswaran, 2020). This paper is closely related to Alvarez, Argente, and Lippi (2020) and Jones, Philippon, and Venkateswaran (2020) which study the optimal lockdown policy to control the fatalities of a pandemic while minimizing the output costs of the lockdown. Alvarez, Argente, and Lippi (2020) suggest that 60% of the population should be under tight lockdown to contain pandemics like COVID-19. The economic costs of such lockdown are estimated to be at least 8% of the GDP. This paper shows that big data technology can significantly improve the tradeoff between economic and human costs of a pandemic.

This paper also contributes to the literature on the effect of big data on the economy. Farboodi and Veldkamp (2019) and Jones and Tonetti (2019) construct neoclassical growth models in which big data are an important contributor to economic growth. This paper provides micro-level evidence that big data can address frictions that limit economic growth.

Athey, Catalini, and Tucker (2017) and Tang (2019) use field experiments to estimate the value of privacy. This paper contributes to this literature by showing that personal location data can create substantial economic value in pandemics, which seems to outweigh the value of privacy. Finally, this paper also sheds light on the regulation of big data technology, an issue studied by Acemoglu, Makhdoui, Malekian, and Ozdaglar (2019), Bergemann, Bonatti, and Gan (2020), and Campbell, Goldfarb, and Tucker (2015).

## 2 Background and Data

**Health code.** Health code is a big data technology that uses smartphone location data to predict the risk of an individual to be infected by a disease. It was developed by several tech companies in China such as Ant Financial and Tencent in the height of the COVID-19 outbreak. Health code are used as “traffic permits” by numerous local governments. Holders of green code can freely travel in the city; holders of yellow or red code have to be quarantined for 7 or 14 days, respectively. The codes turn back to green after the the quarantine. Figure 1 shows the three levels of color codes used in China.

# 杭州健康码



## 防控疫情 人人有责

Figure 1: **Hangzhou Health Code**

This figure shows the first health code introduced in China, the Hangzhou health code. Individuals with green code can freely travel in the city. Individuals with yellow code have to be quarantined for 7 days. Individuals with red code have to be quarantined for 14 days. The code turn back to green after the corresponding quarantine periods.

Registering a health code is voluntary and can be easily done in smartphone apps. Because many cities in China have imposed movement restrictions, people have a strong incentive to register the code if they want to enter public space such as supermarkets and subways. Although there is no systematic record on the adoption rates in the population, anecdotal evidence suggests that the adoption rates are quite high. For instance, in Zhejiang province where health code was first introduced, around 90% of the provincial population has obtained health code 15 days after the introduction according to the disclosure of the local government officials.<sup>4</sup> Among all the health codes, 98.2% are green, and 1.8% are yellow or red.

The first health code app was developed by Ant Financial and implemented in its

---

<sup>4</sup>See New York Times article on March 1, 2020: “In Coronavirus Fight, China Gives Citizens a Color Code, With Red Flags”, by Paul Mozur, Raymond Zhong and Aaron Krolik.

headquarter city, Hangzhou, on February 9, 2020, 17 days after the lockdown of Wuhan. 2 days later, a different version of health code app was developed by another tech giant in China, Tencent, and implemented in its headquarter city, Shenzhen. Following Hangzhou and Shenzhen, many provinces and cities adopt their version of health code. It is worth noting that the adoption of health code was not coordinated by the central government. Instead, it is largely initiated by local governments. The decentralized adoption created a patchwork of policies. Cities do not recognize each other's health code. Different versions of health code sometimes show inconsistent results for the same individual. Some people have been required to scan multiple health code from different providers at a single location.

The uncoordinated implementation of health code have created inconvenience and confusion for people who travel across cities to the extent that the central government warned local governments not to go overboard by launching too many versions of health code.<sup>5</sup> However, it is good news for identification purpose. I collect the implementation dates of health code of 322 Chinese cities from the local government websites and local news media. Figure 2 shows the number of cities that adopted health code over time. The adoption process lasts for two months after the initial adoption in Hangzhou.

---

<sup>5</sup>See South China Morning Post article on March 9, 2020, "National version of China's controversial health code isn't ready".

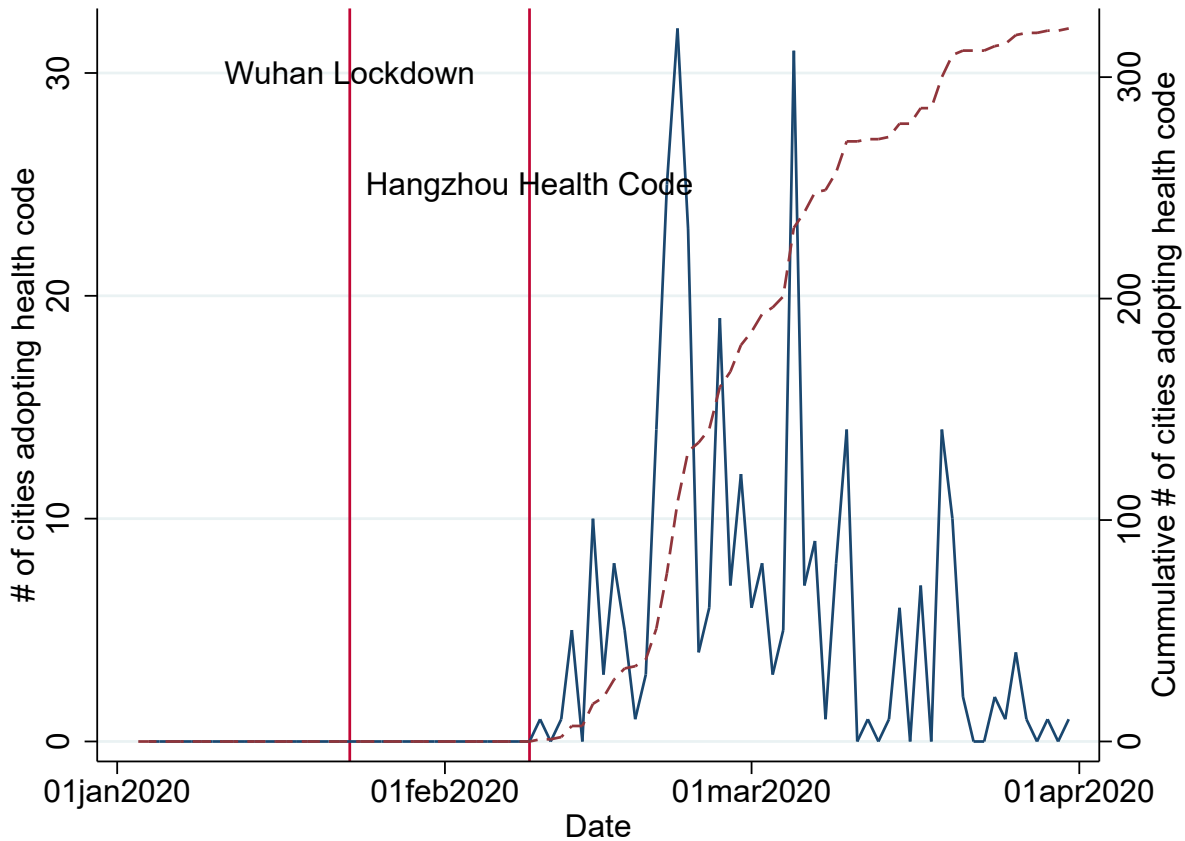


Figure 2: **Implementation of Health Code in Chinese Cities**

This figure plots the number of cities that adopt health code. The first vertical line indicates January 23, 2020, the date of Wuhan lockdown. The second vertical line indicates February 9, 2020, the date when the first health code was introduced in Hangzhou. Data source: government websites, local news report.

Figure 3 shows the fraction of cities that have adopted health code by February 15, February 29, March 15, and March 31, respectively. The adoption appears to be quite idiosyncratic: it is not related to the geographical proximity to the epicenter of the virus outbreak, Hubei. The coastal and inland provinces seem to have a balanced tendency to adopt health code. The staggered introduction of health codes across Chinese cities will provide a great laboratory to identify the causal effect of big data technology on the economy and public health.

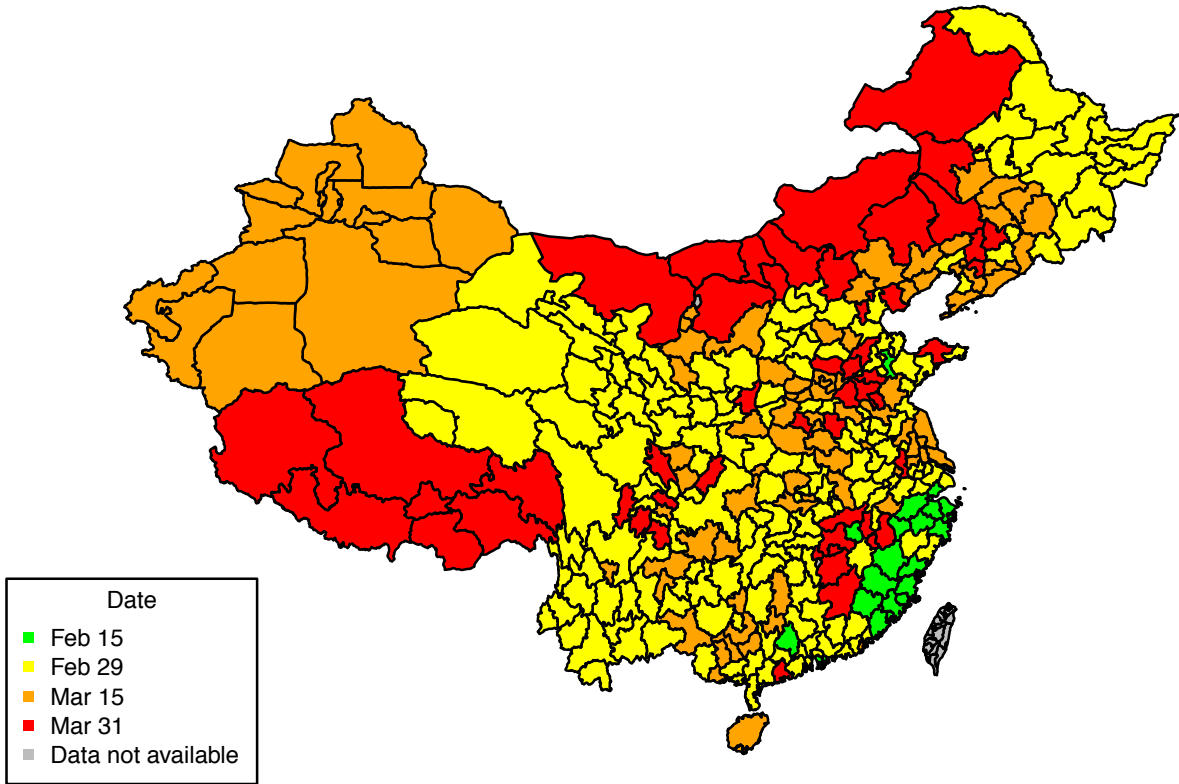


Figure 3: **Implementation of Health Code in Chinese Cities**

This figure shows the cities that adopted health code before four dates: February 15, February 29, March 15, and March 31, 2020. Data source: government websites, local news report.

**Economic activities.** I use daily within-city population movements as a high-frequency measure of economic activities across cities. The data are created using real-time smartphone phone location data from the largest Chinese search engine in China, Baidu.<sup>6</sup> The population movement data covers 322 Chinese cities between January 1 and April 10 in 2020. The final data is a panel consisting of 28,658 city-day level observations. The high-frequency nature of this data is important for identification because health code was rolled out within 2 months. Therefore, typical macroeconomic data at quarterly or monthly frequency may not capture the effect of the adoption.

<sup>6</sup>The source of the data can be found on the website of CNEMC: <http://https://qianxi.baidu.com/>.

Figure 4 plots the national average within-city movement in the sample period. I report the value as a percentage of the average value in the first week of 2020. Note that the sample period contains the Lunar New Year holiday, during which the economic activities would naturally decrease. To control for the effect of Lunar New Year, I normalize the level of within-city movement using the same day value of the 2019 lunar calendar. Figure 4 shows a steep drop in economic activities after January 23, 2020, the day of Wuhan lockdown. The second vertical red line indicates February 9, 2020, the date when Hangzhou health code was introduced. The within-city movement slowly recovers in mid-February. By the end of the sample period, the within-city movement has rebounded to about 95% of the pre-COVID-19 level.

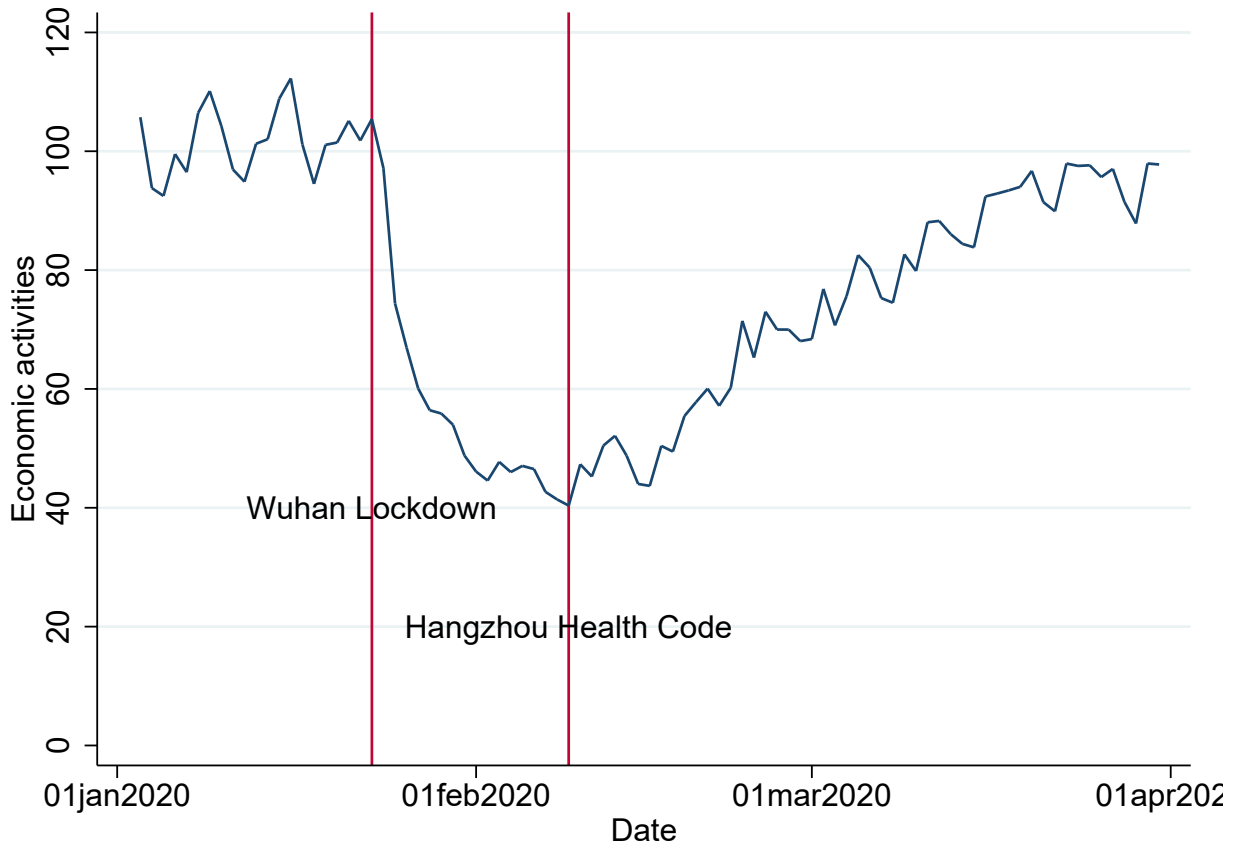


Figure 4: **Economic Activities of Chinese Cities**

This figure plots the average economic activities as measured by within-city population movements. The sample period is from January 1 to March 31, 2020. The first vertical line indicates January 23, 2020, the date of Wuhan lockdown. The second vertical line indicates February 9, 2020, the date when the first health code was introduced in Hangzhou. Data source: Baidu.

The Baidu data also provides a between-city migration pattern. For each city in the sample, the data shows the top 100 cities of inflows and outflows and the corresponding intensities.

**Greenhouse gas emission.** One may worry that within-city movements may not capture economic activities that can be conducted without human movements. To address this concern, I use the daily level of Nitrogen Dioxide (NO<sub>2</sub>) as an alternative high-frequency measure of economic activities. NO<sub>2</sub> is a green house gas created by factories and automobiles

burning fossil fuels. Because Chinese economy heavily relies on coal as a source of energy, the amount of Nitrogen Dioxide is a good measure of economic activities of China. I collect daily level of NO<sub>2</sub> from the China National Environmental Monitoring Center (CNEMC) for each city.<sup>7</sup> Figure 5 plots the average NO<sub>2</sub> of the sample cities where the values are normalized by the average of the first two weeks in 2020. A sharp drop occurs after the Wuhan lockdown on January 23, 2020. Economic activities decreased by 40% of the pre-lockdown level at the peak of outbreak. The magnitude of the reduction is similar to the within-city movements. The NO<sub>2</sub> level started to slowly recover in March, 2020. The sharp decrease in the NO<sub>2</sub> level during the COVID-19 outbreak documented in the data from the China National Environmental Monitoring Center (CNEMC) is consistent with the satellite images produced by the National Aeronautics and Space Administration (NASA) as shown in Figure 6.

---

<sup>7</sup>The source of the data can be found on the website of CNEMC: <http://www.cnemc.cn/>.

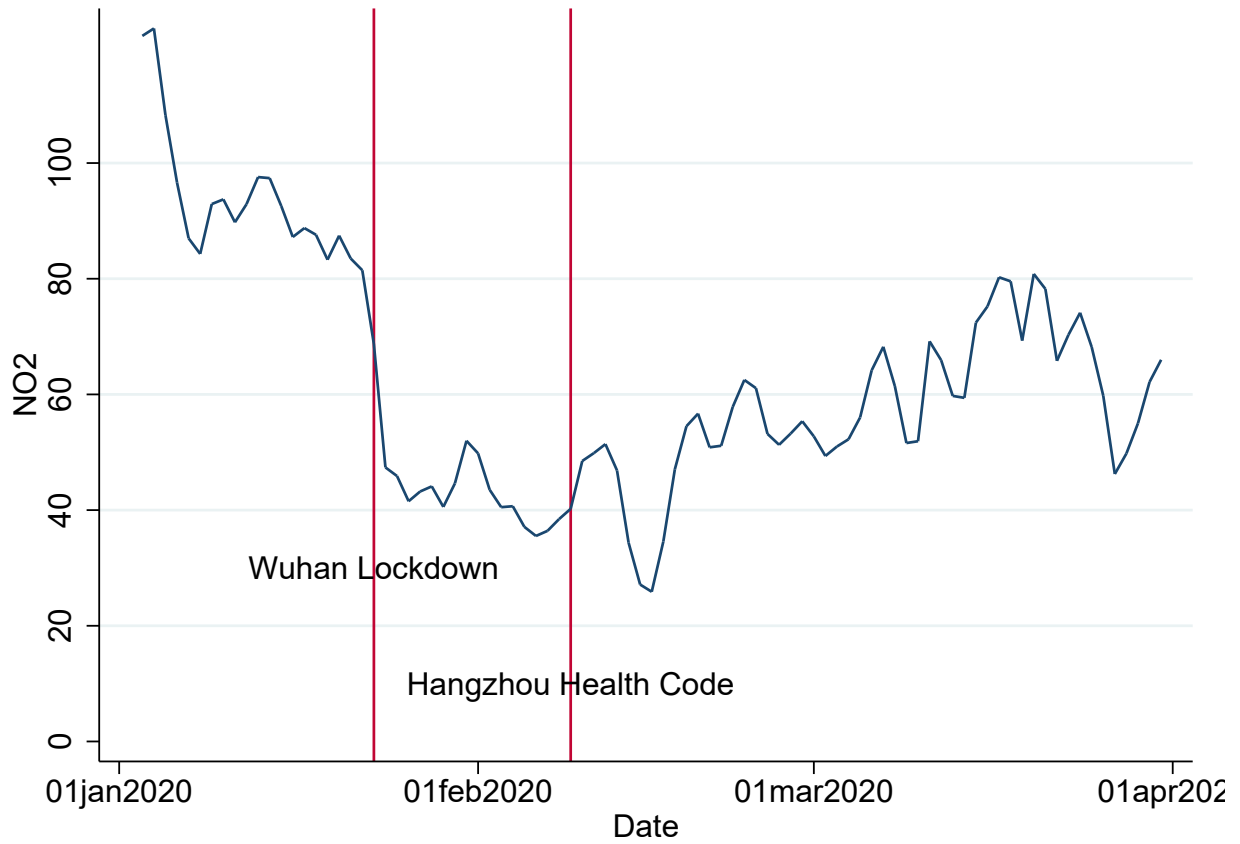


Figure 5: **Average NO2 Level of Chinese Cities**

This figure plots the average NO2 level of Chinese cities. The values are normalized by the average of the first two weeks in 2020. Data source: the China National Environmental Monitoring Center (CNEMC).

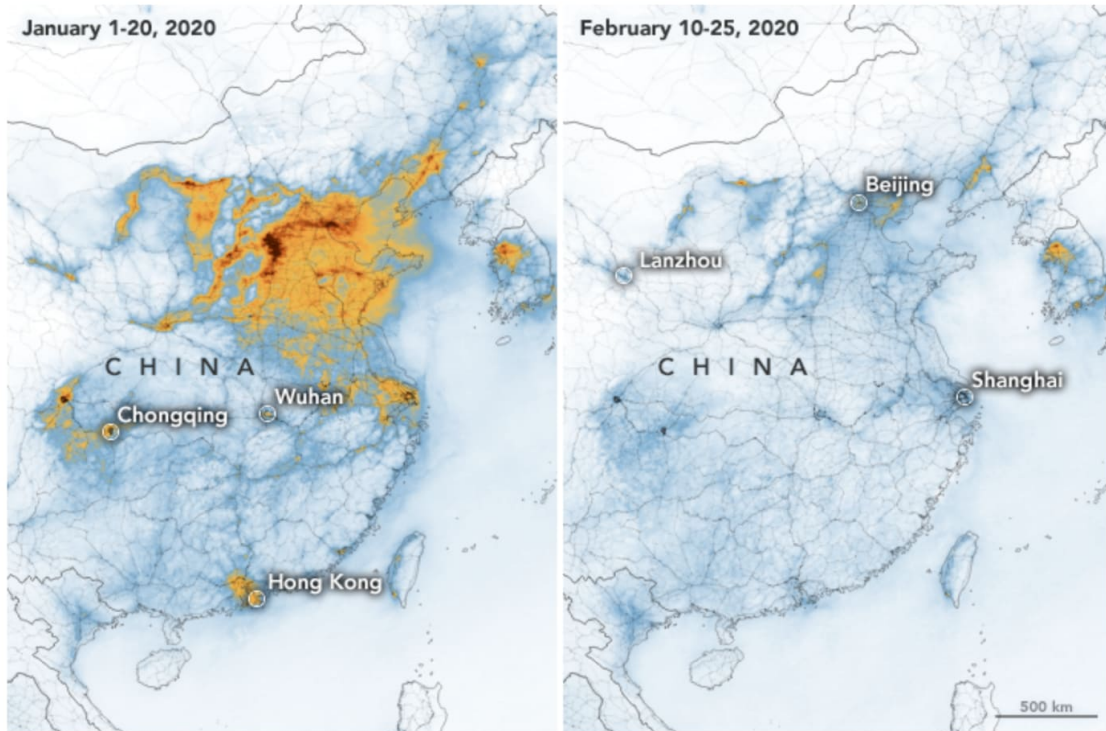


Figure 6: **Nitrogen Dioxide (NO<sub>2</sub>) Level over China**

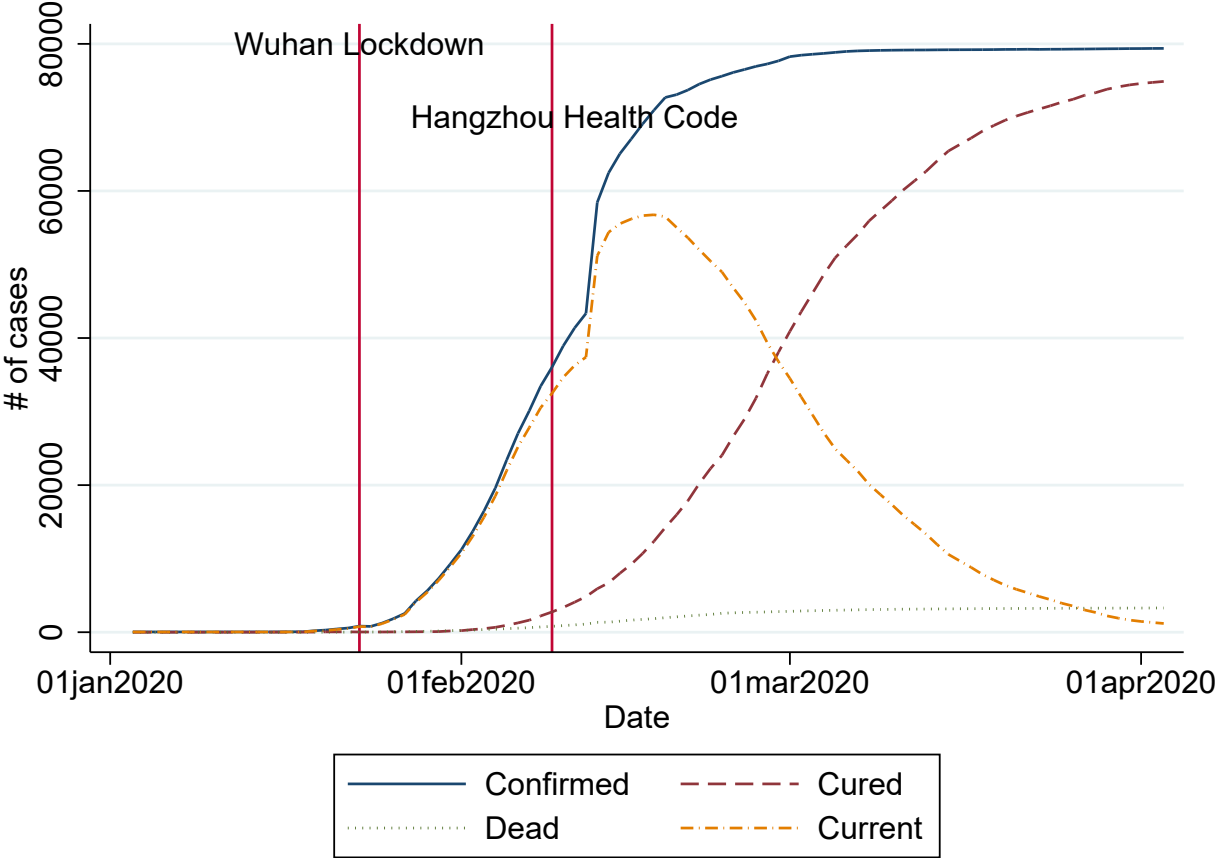
This figure shows the heat maps of Nitrogen Dioxide (NO<sub>2</sub>) level over China. The heat maps are created by the National Aeronautics and Space Administration (NASA).

In addition to NO<sub>2</sub>, I also use the daily level of fine particulate matter (PM<sub>2.5</sub>), which is produced by chemical reactions between gases such as Sulfur Dioxide, Nitrogen Oxides, and volatile organic compounds with dust from industrial activities, as a high-frequency measure of economic activities. I also collect daily level of PM<sub>2.5</sub> for each city from the China National Environmental Monitoring Center (CNEMC).

**Virus outbreak.** I collect the daily count of confirmed, dead, and recovered COVID-19 cases of each of 322 cities from the Centers for Disease Control and Prevention of China (CDC).<sup>8</sup> Figure 7 plots the time series of COVID-19 cases in the sample. From January 11 to April 3, 2020, the data cover 81,198 confirmed COVID-19 cases, 3,302 dead cases, and 75,887 recovered cases. The fatality rate amount the confirmed cases is around 4%, which is

<sup>8</sup>The source of the data can be found on the website of CDC: <http://2019nCoV.chinacdc.cn/2019-nCoV/>.

in line with the fatality rates in other countries. The increase in the confirmed cases levels off in early March. Using this data, I calculate the infection rate, defined as the ratio of newly confirmed cases over the active cases as a measure of the severity of the outbreak.



**Figure 7: Confirmed, Cured, Dead, and Current Cases of COVID-19**  
 This figure plots the cumulative confirmed, cured, dead, and current cases of COVID-19 in the sample. Data source: Chinese Center for Disease Control and Prevention.

One may worry that the numbers of cases in Wuhan and other cities of Hubei can be underestimated because testing capacity was limited at the early stage of the outbreak. Furthermore, government officials in the epicenter cities initially may have also downplayed the severity of the outbreak. Fang, Wang, and Yang (2020) found that there were substantial undocumented infection cases in the early days of the COVID-19 outbreak in cities of Hubei province. Still, they find the gap between the officially reported cases and their estimated

actual cases narrows significantly as the testing capacity was strengthened in Wuhan. To address this concern, I conduct robustness checks for all the regressions by excluding the observations in Hubei province. It is worth noting that economic activity measures are based on smartphone location or air pollution data, which are unlikely to be subject to the same measurement issue as the confirmed COVID-19 cases.

**Summary statistics.** Table 1 provides the summary statistics of the final sample. Panel A reports the city-date sample. I use this sample to study the impact of health code on city-level economic activities and the COVID-19 infection rates. The sample period starts from January 1, 2020, and ends on March 31, 2020. All three measures suggest consistent reduction in economic activities: the average within-city movements, NO<sub>2</sub>, and PM<sub>2.5</sub> are around 78%, 63%, and 78% of their normal levels. The average daily infection rate is 2%, which implies that the new confirmed case grows by 2% of the active cases each day.

In addition to the three main data sources, I collect information on the level of the emergency response of each province from the local government websites and news reports. A higher level of emergency gives local governments greater power to impose exceptional measures such as lockdown and social distancing rules. This system classifies the emergency event into four levels. The lowest level is coded as 0 and the highest as 4. The average emergency level in the sample is 2.

Panel B and C report two city pair-day samples on population inflows and outflows, respectively. I use these two samples to study the impact of health code on the between-city migration pattern. The inflows and outflows are expressed as the percentage of the total flows of the corresponding city. The average inflow and outflows are both 1%.

Table 1: **Summary Statistics**

| Panel A: City-level economic activities |       |      |      |    |     |     |     |     |
|---|-------|------|------|----|-----|-----|-----|-----|
|   | N     | mean | sd   | p5 | p25 | p50 | p75 | p95 |
| Within-city movements                   | 28658 | 78   | 26   | 32 | 56  | 84  | 99  | 109 |
| NO2                                     | 24742 | 63   | 30   | 23 | 40  | 58  | 81  | 119 |
| PM2.5                                   | 24742 | 78   | 51   | 20 | 43  | 68  | 102 | 171 |
| Infection rate                          | 28658 | 2    | 5    | 0  | 0   | 0   | 0   | 20  |
| Confirmed cases                         | 28658 | 144  | 1986 | 0  | 0   | 8   | 31  | 213 |
| Cured cases                             | 28658 | 83   | 1252 | 0  | 0   | 3   | 18  | 140 |
| Dead cases                              | 28658 | 5    | 91   | 0  | 0   | 0   | 0   | 3   |
| Emergency level                         | 28658 | 2    | 1    | 0  | 0   | 2   | 3   | 3   |

| Panel B: City-to-city population inflows |         |      |      |    |     |     |     |     |
|--|---------|------|------|----|-----|-----|-----|-----|
|  | N       | mean | sd   | p5 | p25 | p50 | p75 | p95 |
| Outflow                                  | 1964052 | 1    | 4    | 0  | 0   | 0   | 0   | 4   |
| Confirmed cases (source)                 | 1964052 | 143  | 2020 | 0  | 0   | 8   | 30  | 169 |
| Cured cases (source)                     | 1964052 | 82   | 1270 | 0  | 0   | 3   | 18  | 126 |
| Dead cases (source)                      | 1964052 | 5    | 93   | 0  | 0   | 0   | 0   | 2   |
| Existing cases (source)                  | 1964052 | 56   | 1004 | 0  | 0   | 0   | 6   | 54  |
| Emergency level                          | 1964052 | 2    | 1    | 0  | 0   | 2   | 3   | 3   |

| Panel C: City-to-city population outflows |         |      |      |    |     |     |     |     |
|---|---------|------|------|----|-----|-----|-----|-----|
|   | N       | mean | sd   | p5 | p25 | p50 | p75 | p95 |
| Outflow                                   | 1896486 | 1    | 4    | 0  | 0   | 0   | 0   | 4   |
| Confirmed cases (source)                  | 1896486 | 116  | 1548 | 0  | 0   | 17  | 49  | 252 |
| Cured cases (source)                      | 1896486 | 75   | 1083 | 0  | 0   | 6   | 35  | 173 |
| Dead cases (source)                       | 1896486 | 3    | 72   | 0  | 0   | 0   | 0   | 3   |
| Existing cases (source)                   | 1896486 | 37   | 725  | 0  | 0   | 0   | 11  | 76  |
| Emergency level                           | 1896486 | 2    | 1    | 0  | 0   | 2   | 3   | 3   |

*Note:* This table reports summary statistics of the regression sample. The sample is a panel of 322 cities from January 1, 2020 to March 31, 2020. Data sources: Baidu, Chinese Center for Disease Control and Prevention.

### 3 Empirical Results

In this section, I exploit the staggered implementation of health code in 322 Chinese cities to identify the causal effects of big data technology on economic activities and public health conditions. Specifically, I use a difference-in-differences (DID) research design to test three main hypotheses: (1) whether the introduction of health code increases local economic activities, (2) whether the introduction of health code affects the migration pattern between cities, and (3) whether the introduction of health code reduces the infection rates of COVID-19.

#### 3.1 Economic activities

I study the effects of big data technology on economic activities measured by within-city population movements:

$$\text{EconomicActivity}_{i,t} = \beta \text{HealthCode}_{i,t} + \gamma X_{i,t} + \epsilon_{i,t}, \quad (1)$$

where  $\text{EconomicActivity}_{i,t}$  is measured by within-city population movements of city  $i$  on date  $t$ ,  $\text{HealthCode}_{i,t}$  is a dummy variable which equals to 1 if city  $i$  has health code at time  $t$ , and 0 otherwise. The vector of the control variables,  $X_{i,t}$ , includes the emergency level of the city and the log number of confirmed, cured, and dead cases. I also include city fixed effects to absorb time-invariant city characteristics, time fixed effects to absorb aggregate shocks. Therefore, the empirical design effectively compares differential changes in economic activities of the treated cities with those of the untreated cities before and after the introduction of health code.

Column 1 of Table 2 presents the baseline results. I find that the introduction of health code significantly increases local economic activities by around 2%. The regression also shows that the severity of the outbreak also significantly affects local economic activities.

In particular, an increase in confirmed and dead cases significantly reduce local economic activities while an increase in cured cases increases local economic activities.

One may worry whether the above result is purely driven by comparing cities in Hubei province, the epicenter of the virus outbreak, with the rest of the country. Column 2 of Table 2 presents the results, excluding cities in Hubei province. I find the result is largely the same as the baseline.

Another potential concern on the empirical approach is that cities that adopt health code early could have higher economic importance for the country, thus are forced to reopen before other cities. I address this concern by matching treated cities to control cities with similar pre-COVID-19 economic activities. The result is presented in Column 3 of Table 2. The result is robust in the matching sample.

Finally, one may worry that the timing of implementation of health code may be correlated with a differential trajectory of outbreaks in each city. Cities may choose to implement health code because the outbreak is over. I address this concern by matching the treated cities to control cities with a similar number of active cases when health code was introduced. The result is presented in Column 4 of Table 2. The results are also robust to this alternative matching scheme.

Table 2: **Health Code and Economic Activities**

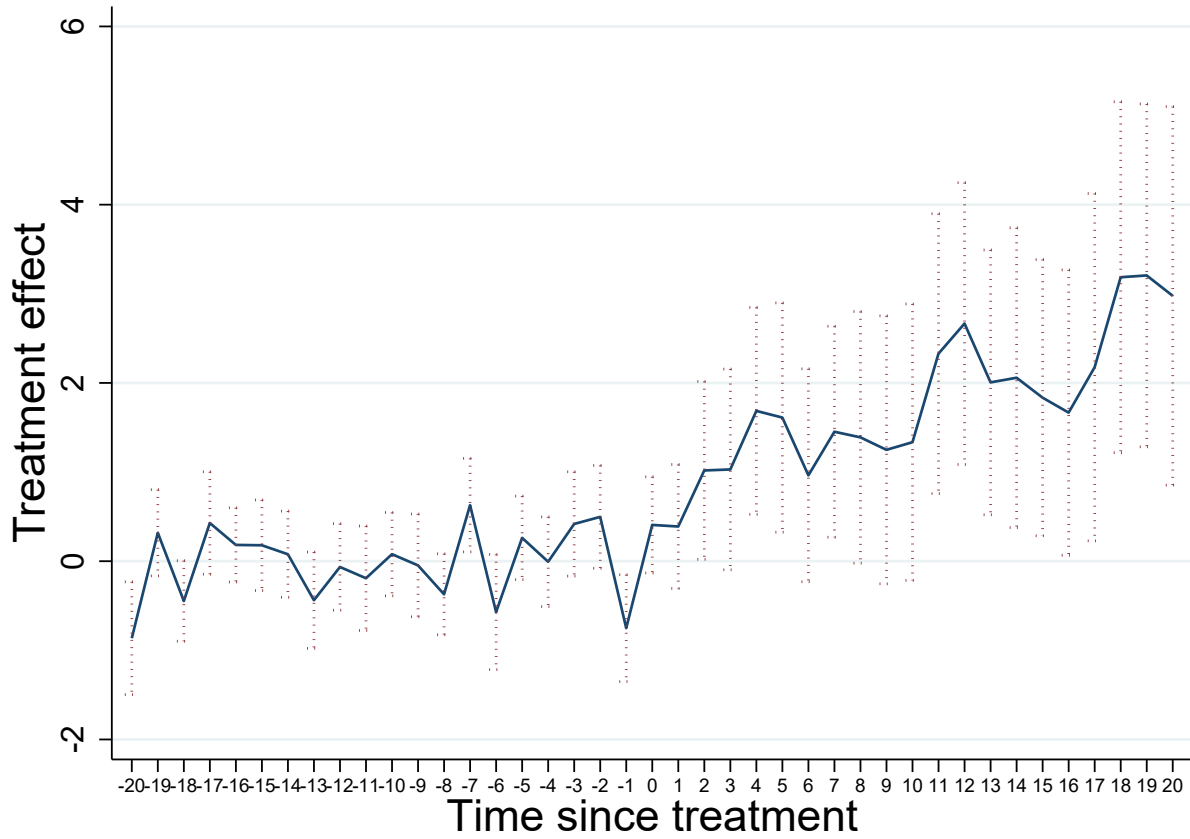
|                 | (1)                  | (2)                  | (3)                  | (4)                  |
|-----------------|----------------------|----------------------|----------------------|----------------------|
|                 | Movement             | Movement             | Movement             | Movement             |
| Health code     | 2.859***<br>[0.410]  | 2.687***<br>[0.345]  | 2.552***<br>[0.437]  | 3.118***<br>[0.399]  |
| Confirmed cases | -4.560***<br>[0.195] | -4.515***<br>[0.276] | -4.477***<br>[0.177] | -4.423***<br>[0.187] |
| Cured cases     | 2.191***<br>[0.203]  | 2.147***<br>[0.215]  | 2.005***<br>[0.207]  | 2.284***<br>[0.201]  |
| Dead cases      | -2.713***<br>[0.649] | -6.125***<br>[0.706] | -1.617**<br>[0.645]  | -2.901***<br>[0.639] |
| City F.E.       | Yes                  | Yes                  | Yes                  | Yes                  |
| Time F.E.       | Yes                  | Yes                  | Yes                  | Yes                  |
| Emergency F.E.  | Yes                  | Yes                  | Yes                  | Yes                  |
| Sample          | Full sample          | Excl. Hubei          | Match by cases       | Match by act.        |
| Observations    | 28,658               | 27,145               | 28,658               | 28,658               |
| Adj. R-squared  | 0.852                | 0.862                | 0.867                | 0.852                |

*Note:* This table reports the results of the following regression

$$\text{EconomicActivity}_{i,t} = \beta \text{HealthCode}_{i,t} + \gamma X_{i,t} + \epsilon_{i,t}$$

where  $\text{EconomicActivity}_{i,t}$  is measured by within-city movement of city  $i$  on date  $t$ ,  $\text{HealthCode}_{i,t}$  is a dummy variable which equals to 1 if city  $i$  has health code at time  $t$ , and 0 otherwise. The vector of the control variables,  $X_{i,t}$ , includes the emergency level of the city, the log number of confirmed/dead/cured cases, city fixed effects, and time fixed effects. The sample is a panel of 322 cities from January 1, 2020 to March 31, 2020. Standard errors are clustered at date level. Data sources: Baidu, Chinese Center for Disease Control and Prevention.

To investigate the dynamic effects of health code introduction, Figure 8 plots the difference in the economic activities between treated and control cities 20 days before and after the implementation of health code. Before the introduction of health code, there is no pre-trend between the treated and control cities, suggesting that the parallel trend assumption seems to hold in the data. After the introduction of health code, the economic activities of the treatment cities increases by around 2%-3% compared to the controlled cities.



**Figure 8: Difference in Economic Activities in Treated and Control Cities**  
 This figure plots the difference in economic activities in treated and control cities. The horizontal axis is the date since the adoption of health code. Standard errors are clustered at date level. Data source: Baidu, Chinese Center for Disease Control and Prevention,

Within-city movements may not capture the economic activities can be conducted without population movements. To address this concern, I use the concentration level of NO<sub>2</sub> and PM<sub>2.5</sub> as alternative measures of economic activities. The results are reported in Table 3. Consistent with the baseline measure, I find that the introduction of health code significantly increases economic activities as proxied by the concentration level of NO<sub>2</sub> and PM<sub>2.5</sub>. The economic magnitude is also quite similar: economic activities increase by 2-3% in cities where health code is implemented.

Table 3: **Health Code and Economic Activities (Alternative Measures)**

|                 | (1)                  | (2)                  | (3)                 | (4)                   |
|-----------------|----------------------|----------------------|---------------------|-----------------------|
|                 | NO2                  | NO2                  | PM2.5               | PM2.5                 |
| Health code     | 1.792*<br>[1.006]    | 1.973*<br>[1.057]    | 4.989***<br>[1.671] | 4.830***<br>[1.767]   |
| Confirmed cases | -1.707**<br>[0.661]  | -2.393***<br>[0.659] | -0.146<br>[1.821]   | -0.685<br>[1.752]     |
| Cured cases     | 3.939***<br>[0.654]  | 4.640***<br>[0.655]  | -0.782<br>[1.750]   | -0.155<br>[1.729]     |
| Dead cases      | -5.741***<br>[0.765] | -8.721***<br>[1.315] | -2.454*<br>[1.412]  | -10.415***<br>[3.484] |
| City F.E.       | Yes                  | Yes                  | Yes                 | Yes                   |
| Time F.E.       | Yes                  | Yes                  | Yes                 | Yes                   |
| Emergency F.E.  | Yes                  | Yes                  | Yes                 | Yes                   |
| Sample          | Full sample          | Excl. Hubei          | Full sample         | Excl. Hubei           |
| Observations    | 24,742               | 23,674               | 24,742              | 23,674                |
| Adj. R-squared  | 0.541                | 0.534                | 0.358               | 0.352                 |

*Note:* This table reports the results of the following regression

$$\text{EconomicActivity}_{i,t} = \beta \text{HealthCode}_{i,t} + \gamma X_{i,t} + \epsilon_{i,t}$$

where  $\text{EconomicActivity}_{i,t}$  is measured by the NO2 or PM2.5 levels of city  $i$  on date  $t$ ,  $\text{HealthCode}_{i,t}$  is a dummy variable which equals to 1 if city  $i$  has health code at time  $t$ , and 0 otherwise. The vector of the control variables,  $X_{i,t}$ , includes the emergency level of the city, the log number of confirmed/dead/cured cases, city fixed effects, and time fixed effects. The sample is a panel of 322 cities from January 1, 2020 to March 31, 2020. Standard errors are clustered at date level. Data sources: Baidu, Chinese Center for Disease Control and Prevention.

### 3.2 Between-city migration pattern

Next, I investigate how health code affects the migration pattern between cities. Specifically, I first examine the effect of health code on inflows into cities. The regression model is the following:

$$\text{Inflow}_{i,j,t} = \beta \text{DestinationHealthCode}_{j,t} + \gamma X_{i,j,t} + \epsilon_{i,t},$$

where  $\text{Inflow}_{i,j,t}$  is the flow from city  $i$  to city  $j$  on date  $t$ ,  $\text{DestinationHealthCode}_{j,t}$  is a dummy variable which equals to 1 if destination city  $j$  has health code at time  $t$ , and 0

otherwise. The vector of the control variables,  $X_{i,j,t}$ , includes the emergency level, and the log number of confirmed, cured, and dead cases in the destination city. I also include destination-city fixed effects to absorb time-invariant city characteristics. Finally, I include source city-time fixed effects. The regression compares the flows from the same source city to two similar destination cities, of which one has health code, but the other does not. Table 4 shows the results. I find that the introduction of health code significantly increase the inflows to cities with health code by around 13%. This result suggests that cities with health code become more attractive as most residents can move freely and economic activities recover.

Table 4: **Health Code and Population Inflows**

|                               | (1)                   | (2)                   | (3)                   | (4)                   |
|-------------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
|                               | Inflow                | Inflow                | Inflow                | Inflow                |
| Health Code (destn)           | 12.694***<br>[1.700]  | 13.185***<br>[1.696]  | 11.183***<br>[1.709]  | 12.557***<br>[1.687]  |
| Confirmed cases (destination) | 3.479**<br>[1.470]    | 3.937**<br>[1.505]    | 2.486*<br>[1.426]     | 3.422**<br>[1.475]    |
| Cured cases (destination)     | 3.158**<br>[1.278]    | 3.119**<br>[1.319]    | 2.262*<br>[1.216]     | 3.102**<br>[1.281]    |
| Dead cases (destination)      | -18.900***<br>[1.021] | -20.139***<br>[1.259] | -17.484***<br>[1.052] | -18.764***<br>[1.037] |
| City pair F.E.                | Yes                   | Yes                   | Yes                   | Yes                   |
| Source-time F.E.              | Yes                   | Yes                   | Yes                   | Yes                   |
| Emergency level F.E.          | Yes                   | Yes                   | Yes                   | Yes                   |
| Sample                        | Full sample           | Excl. Hubei           | Match by cases        | Match by act.         |
| Observations                  | 1,888,652             | 1,798,439             | 1,888,652             | 1,888,652             |
| Adj. R-squared                | 0.857                 | 0.859                 | 0.863                 | 0.858                 |

*Note:* This table reports the results of the following regression

$$\text{Inflow}_{i,j,t} = \beta \text{DestinationHealthCode}_{j,t} + \gamma X_{i,j,t} + \epsilon_{i,j,t},$$

where  $\text{Inflow}_{i,j,t}$  is the flow from city  $i$  to city  $j$  on date  $t$ ,  $\text{DestinationHealthCode}_{j,t}$  is a dummy variable which equals to 1 if destination city  $j$  has health code at time  $t$ , and 0 otherwise. The vector of the control variables,  $X_{i,j,t}$ , includes the emergency level, and the log number of confirmed/dead/cured cases in the destination city, destination-city fixed effects, and source city-time fixed effects. The sample is a panel of 322 cities from January 1, 2020 to March 31, 2020. Standard errors are clustered at date level. Data sources: Baidu, Chinese Center for Disease Control and Prevention.

I then examine how health code affects the population outflows from a city. The regression model is the following:

$$\text{Outflow}_{i,j,t} = \beta \text{SourceHealthCode}_{i,t} + \gamma X_{i,j,t} + \epsilon_{i,t}$$

where  $\text{Outflow}_{i,j,t}$  is the flow from city  $i$  to city  $j$  on date  $t$ ,  $\text{SourceHealthCode}_{i,t}$  is a dummy variable which equals to 1 if source city  $i$  has health code at time  $t$ , and 0 otherwise. The vector of the control variables,  $X_{i,j,t}$ , includes the emergency level, and the log number of confirmed, cured, and dead cases in the source city. I also include source-city fixed effects to absorb time-invariant city characteristics. Finally, I include source city-time fixed effects. The regression compares the flows from two similar source cities to the same destination city. One of the source cities has health code, but the other does not. Table 5 shows the results. I find that the introduction of health code significantly decrease the outflows from cities with health code by 14%. This result suggests that residents in cities with health code seem to be more willing to stay in the cities, presumably due to the recovery of economic activities.

Table 5: **Health Code and Population Outflows**

|                          | (1)                   | (2)                   | (3)                   | (4)                   |
|--------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
|                          | Outflow               | Outflow               | Outflow               | Outflow               |
| Health Code (source)     | -14.840***<br>[1.778] | -14.647***<br>[1.744] | -13.681***<br>[1.612] | -14.549***<br>[1.749] |
| Confirmed cases (source) | -12.896***<br>[1.678] | -15.319***<br>[1.806] | -12.232***<br>[1.661] | -12.195***<br>[1.628] |
| Cured cases (source)     | -4.449***<br>[1.272]  | -3.501***<br>[1.229]  | -3.880***<br>[1.206]  | -4.321***<br>[1.243]  |
| Dead cases (source)      | 7.602***<br>[1.658]   | -11.448***<br>[1.841] | 10.099***<br>[1.714]  | 7.272***<br>[1.617]   |
| City pair F.E.           | Yes                   | Yes                   | Yes                   | Yes                   |
| Destination-time F.E.    | Yes                   | Yes                   | Yes                   | Yes                   |
| Emergency level F.E.     | Yes                   | Yes                   | Yes                   | Yes                   |
| Sample                   | Full sample           | Excl. Hubei           | Match by cases        | Match by act.         |
| Observations             | 1,887,544             | 1,834,631             | 1,887,544             | 1,887,544             |
| Adj. R-squared           | 0.860                 | 0.862                 | 0.861                 | 0.860                 |

*Note:* This table reports the results of the following regression

$$\text{Outflow}_{i,j,t} = \beta \text{SourceHealthCode}_{i,t} + \gamma X_{i,j,t} + \epsilon_{i,j,t},$$

where  $\text{Outflow}_{i,j,t}$  is the flow from city  $i$  to city  $j$  on date  $t$ ,  $\text{SourceHealthCode}_{i,t}$  is a dummy variable which equals to 1 if source city  $i$  has health code at time  $t$ , and 0 otherwise. The vector of the control variables,  $X_{i,j,t}$ , includes the emergency level, and the log number of confirmed/dead/cured cases in the source city, source-city fixed effects, and destination city-time fixed effects. The sample is a panel of 322 cities from January 1, 2020 to March 31, 2020. Standard errors are clustered at date level. Data sources: Baidu, Chinese Center for Disease Control and Prevention.

### 3.3 Infection rate of COVID-19

The previous results offer evidence that the introduction of health code allow the economy to return to normal. However, one important question is whether the reopen of the economy will lead to a resurgence of virus infection in the future. To test this hypothesis, I estimate the following regression model:

$$\text{InfectionRate}_{i,t+7} = \beta \text{HealthCode}_{i,t} + \gamma X_{i,t} + \epsilon_{i,t}$$

where  $\text{InfectionRate}_{i,t+7}$  is the infection rate of COVID-19 in city  $i$  on date  $t+7$ ;  $\text{HealthCode}_{i,t}$  is a dummy variable which equals to 1 if city  $i$  has health code at time  $t$ , and 0 otherwise. The vector of the control variables,  $X_{i,t}$ , includes the emergency level of the city and the log number of confirmed, cured, and dead cases. I also include city fixed effects to absorb time-invariant city characteristics, time fixed effects to absorb aggregate shocks. It is worth noting that I use the infection rate in 7 days because 7 days is the median time when symptoms appear after exposure to the virus.

Column 1 of Table 6 presents the baseline results. I find that the introduction of health code does not seem to increase the infection rates of COVID-19 despite that economic activities have significantly increased. Column 2 of Table 6 presents the results excluding cities in Hubei province. I find the result is largely the same as the baseline. This result alleviates the concern that under-reporting in the epicenter of the virus outbreak could drive the result. One may also worry that the timing of implementation of health code may be endogenous to whether the virus outbreak was successfully stopped in a city. I address this concern by matching the existing confirmed cases at the time of the introduction of health code. Column 3 of Table 6 shows that the result in the matched sample is quite similar to the baseline regression as well. Finally, in the sample matched by economic activities, as shown in Column 4, I find the result is virtually the same.

Table 6: **Health Code and Infection Rates**

|                 | (1)                  | (2)                  | (3)                  | (4)                  |
|-----------------|----------------------|----------------------|----------------------|----------------------|
|                 | Infection rate       | Infection rate       | Infection rate       | Infection rate       |
| Health code     | 0.024<br>[0.078]     | 0.028<br>[0.081]     | -0.000<br>[0.092]    | 0.027<br>[0.082]     |
| Confirmed cases | -0.526*<br>[0.314]   | -0.746**<br>[0.317]  | -0.529<br>[0.325]    | -0.544*<br>[0.311]   |
| Cured cases     | -0.886***<br>[0.120] | -0.756***<br>[0.113] | -0.856***<br>[0.140] | -0.878***<br>[0.113] |
| Dead cases      | 0.411**<br>[0.204]   | 1.272**<br>[0.493]   | 0.440**<br>[0.180]   | 0.431**<br>[0.173]   |
| City F.E.       | Yes                  | Yes                  | Yes                  | Yes                  |
| Time F.E.       | Yes                  | Yes                  | Yes                  | Yes                  |
| Emergency F.E.  | Yes                  | Yes                  | Yes                  | Yes                  |
| Sample          | Full sample          | Excl. Hubei          | Match by cases       | Match by act.        |
| Observations    | 26,404               | 25,010               | 26,404               | 26,404               |
| Adj. R-squared  | 0.411                | 0.388                | 0.412                | 0.408                |

*Note:* This table reports the results of the following regression

$$\text{InfectionRate}_{i,t+7} = \beta \text{HealthCode}_{i,t} + \gamma X_{i,t} + \epsilon_{i,t}$$

where  $\text{InfectionRate}_{i,t+7}$  is infection rate of city  $i$  on date  $t + 7$ ,  $\text{HealthCode}_{i,t}$  is a dummy variable which equals to 1 if city  $i$  has health code at time  $t$ , and 0 otherwise. The vector of the control variables,  $X_{i,t}$ , includes the emergency level of the city, the log number of confirmed/dead/cured cases, city fixed effects, and time fixed effects. The sample is a panel of 322 cities from January 1, 2020 to March 31, 2020. Standard errors are clustered at date level. Data sources: Baidu, Chinese Center for Disease Control and Prevention.

Figure 7 reports the dynamic effect of the introduction of health code on infection rates of COVID-19. Again, there is no pre-trend between treated and control cities suggesting the parallel trend assumption is stratified in the data. Furthermore, the introduction of health code do not significantly increase the infection rate.

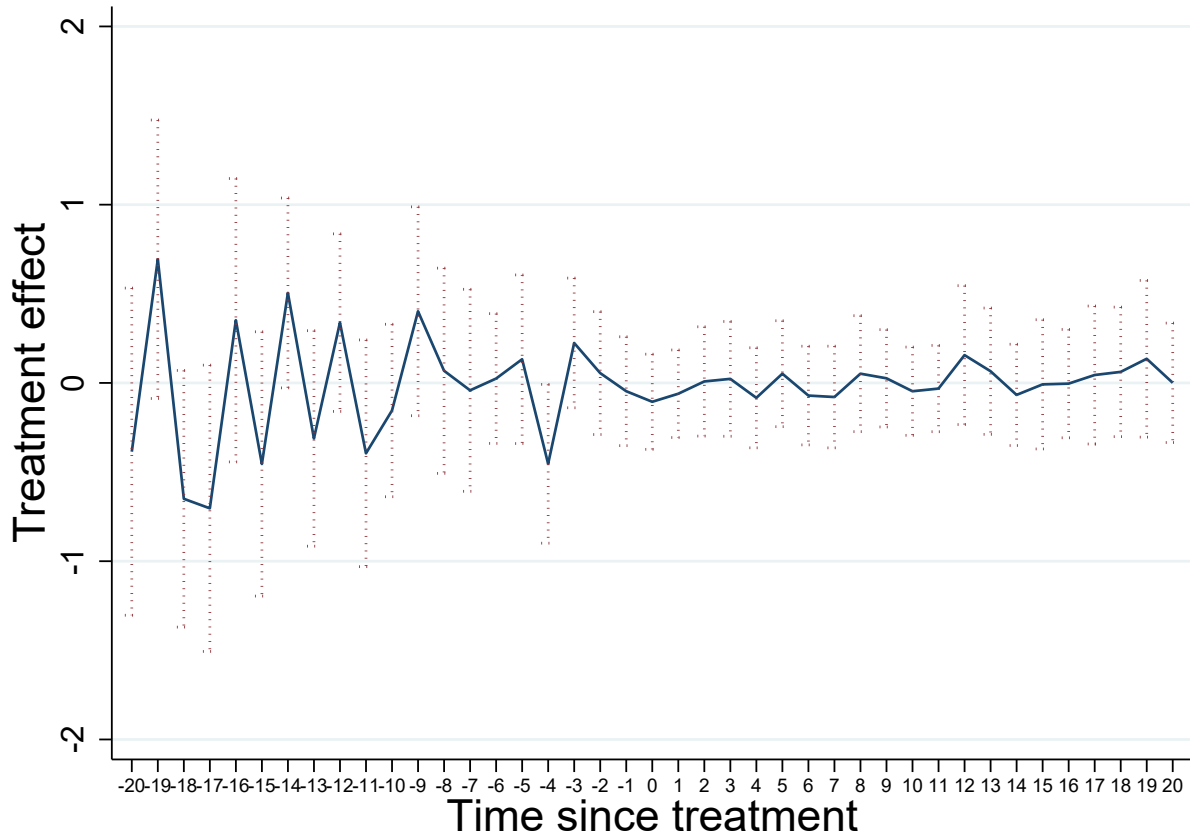


Figure 9: **Difference in Infection Rates in Treated and Control Cities**  
 This figure plots the difference in infection rates in treated and control cities. The horizontal axis is the date since the adoption of health code. Standard errors are clustered at date level. Data source: Baidu, Chinese Center for Disease Control and Prevention,

### 3.4 Trade-off between lives and livelihoods?

Finally, I examine whether the big data technology helps to improve the trade-off between lives and livelihoods. Specifically, I estimate the relationship between economic activities and future infection rates with and without health code.

$$\text{InfectionRate}_{i,t+7} = \beta_1 \text{HealthCode}_{i,t} \times \text{EconomicActivity}_{i,t} + \gamma X_{i,t} + \epsilon_{i,t}$$

where  $\text{InfectionRate}_{i,t+7}$  is the infection rate of COVID-19 in city  $i$  on date  $t+7$ ;  $\text{HealthCode}_{i,t}$  is a dummy variable which equals to 1 if city  $i$  has health code at time  $t$ , and 0 otherwise;  $\text{EconomicActivity}_{i,t}$  is measured by the index of within-city movement of city  $i$  on date  $t$ . The vector of the control variables,  $X_{i,t}$ , includes the current infection rate, emergency level fixed effects, city fixed effects, and time fixed effects.

Figure 10 plots the predicted infection rates for each level of economic activities with and without health code, respectively. I find that health code improve the tradeoff between economic activities and virus infection. Specifically, without health code, economic activities have to decrease by 19% to reduce the daily infection rate by 1% from its average level. In contrast, with health code, economic activities only need to decrease by 17% to achieve the same amount of reduction in infection rate.

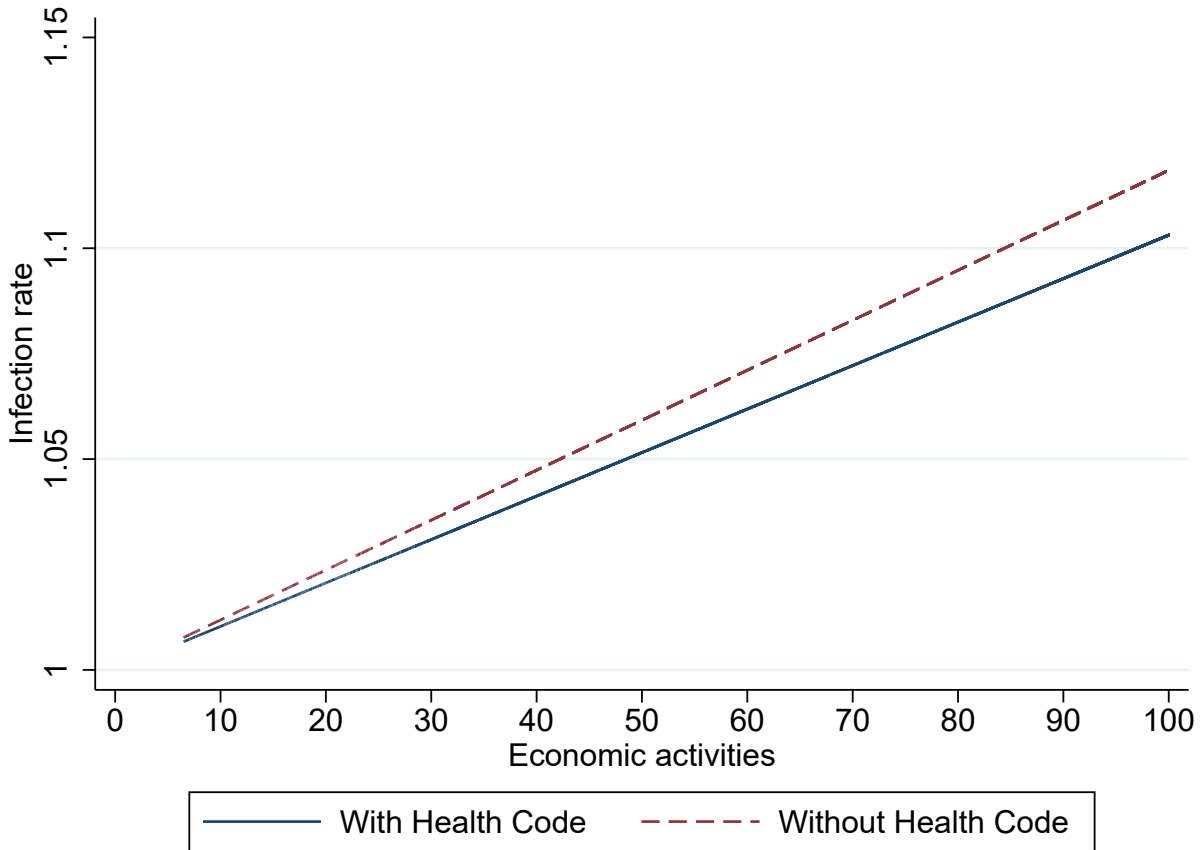


Figure 10: **Trade-off Between Economic Activities and COVID-19 Infection Rate**  
 This figure plots the trade-off between economic activities and COVID-19 infection rates. Data source: Baidu, Chinese Center for Disease Control and Prevention.

### 3.5 Do the benefits of health code justify the costs?

The above results show that health code can significantly improve economic activities without sacrificing the public health. However, using big data technology could lead to other types of costs. The most prominent concern is privacy. Do the benefits created by health code justify its costs on privacy? In this section, I conduct a back-of-the-envelope calculation of the benefits of health code and compare it with the costs of privacy estimated in the literature.

Section 3.1 shows that the introduction of health code increase economic activities by around 2-3%. Assuming that the COVID-19 outbreak lasts for a quarter, then the in-

roduction of health code creates an economic value of 0.5%-0.75% GDP. Given the GDP per capita in China is around \$10,000 as of 2018, a 0.5% increase in GDP translates to  $0.5\% \times \$10,000 = \$50$  per person.<sup>9</sup> In comparison, Tang (2019) estimate that in a field experiment that Chinese people value their privacy at a value of \$33. It seems that the benefits of the health code technology dominate the potential costs of privacy.

A caveat of this cost-benefit analysis is that people in different countries may value privacy differently. However, Athey, Catalini, and Tucker (2017) conduct experiments in the U.S. and find that, even for people who claim to value privacy, they are willing to relinquish their private data to exchange for small benefits. Therefore, big data technology could be welfare improving in many countries where people appear to value privacy a lot. The second caveat is that the above estimate is constructed based on the assumption that the COVID-19 outbreak lasts for a quarter. However, if the COVID-19 outbreak lasts longer, then the benefits should be adjusted accordingly. Third, the value of privacy estimated by Tang (2019) is based on sharing social network ID and employer contact while health code require location data, which may be valued differently by people. Fourth, using big data technology does not necessarily lead to a loss of privacy. If health code is implemented by a trusted entity or is protected by data anonymization technology, then people may be more willing to share their data. Finally, the effectiveness of every big data technology depends on the quality of data input. To make health code effective, it is estimated that three-quarter of the population needs to register. The adoption rate appears to be an issue for Singapore where only 20% of 5.7 million population has registered for their contact tracing app one month after the introduction. The low adoption rate seems to compromise the effectiveness of this technology as the number of cases keeps rising.<sup>10</sup>

---

<sup>9</sup>See the World Bank data: <https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=CN>

<sup>10</sup>See WSJ April 22 article, "Singapore Built a Coronavirus App, but It Hasn't Worked So Far".

## 4 Conclusion

Pandemics such as COVID-19 inflicts enormous costs to the economy. Policymakers are facing the impossible choice between saving human lives and saving the economy. This paper examines the effectiveness of big data technology in addressing the pandemic dilemma using a large experiment of health code in China. Exploiting the staggered implementation of health code in 322 cities in China, I find that the introduction of this technology significantly revives economic activities while keeping the outbreak under control. I find that the benefits of this big data technology seem to outweigh its potential costs on privacy. Given the medical cure of the disease is still elusive, big data technology presents a promising solution to the pandemic dilemma between lives and livelihoods.

This paper argues that big data technology addresses the key amplifier of the economic costs caused pandemics, that is, information friction. This result has many important implications because information friction lies in the hearts of many social and economic problems. By leveraging the enormous amount of data produced in our digital age, big data technology can alleviate this friction and provide better solutions for many existing problems. How to harness the power of big data technology without threatening our privacy will be a big question in the post-COVID-19 world.

## References

- Acemoglu, D., A. Makhdoumi, A. Malekian, and A. Ozdaglar, 2019, “Too much data: Prices and inefficiencies in data markets,” working paper, National Bureau of Economic Research.
- Alvarez, F., D. Argente, and F. Lippi, 2020, “A Simple Planning Problem for COVID-19 Lockdown,” *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2020-34).
- Athey, S., C. Catalini, and C. Tucker, 2017, “The digital privacy paradox: Small money, small costs, small talk,” working paper, National Bureau of Economic Research.
- Barro, R. J., J. F. Ursúa, and J. Weng, 2020, “The coronavirus and the great influenza pandemic: Lessons from the spanish flu for the coronavirus’s potential effects on mortality and economic activity,” working paper, National Bureau of Economic Research.
- Bergemann, D., A. Bonatti, and T. Gan, 2020, “The economics of social data,” .
- Campbell, J., A. Goldfarb, and C. Tucker, 2015, “Privacy regulation and market structure,” *Journal of Economics & Management Strategy*, 24(1), 47–73.
- Correia, S., S. Luck, and E. Verner, 1918, “Pandemics Depress the Economy, Public Health Interventions Do Not: Evidence from the 1918 Flu,” .
- Dewatripont, M., M. Goldman, E. Muraille, and J.-P. Platteau, 2020, “Rapid identification of workers immune to COVID-19 and virus-free: A priority to restart the economy,” working paper, Discussion paper, Universit Libre de Bruxelles.
- Eichenbaum, M. S., S. Rebelo, and M. Trabandt, 2020, “The macroeconomics of epidemics,” working paper, National Bureau of Economic Research.
- Fang, H., L. Wang, and Y. Yang, 2020, “Human mobility restrictions and the spread of the novel coronavirus (2019-ncov) in china,” working paper, National Bureau of Economic Research.
- Farboodi, M., and L. Veldkamp, 2019, “A Growth Model of the Data Economy,” working paper, Working Paper, Columbia Business School, New York, June 20.

- Ferretti, L., C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dörner, M. Parker, D. Bonsall, and C. Fraser, 2020, “Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing,” *Science*.
- Hall, R. E., C. I. Jones, and P. J. Klenow, 2020, “Trading Off Consumption and COVID-19 Deaths,” working paper, Mimeo, Stanford University.
- Jones, C., T. Philippon, and V. Venkateswaran, 2020, “Optimal Mitigation Policies in a Pandemic,” working paper, Covid Economics Issue 4.
- Jones, C. I., and C. Tonetti, 2019, “Nonrivalry and the Economics of Data,” working paper, National Bureau of Economic Research.
- Piguillem, F., L. Shi, et al., 2020, “The optimal covid-19 quarantine and testing policies,” working paper, Einaudi Institute for Economics and Finance (EIEF).
- Tang, H., 2019, “The value of privacy: Evidence from online borrowers,” working paper, Working Paper, HEC Paris.