For the public, data collection during COVID-19 offers benefits and poses hazards

27 July 2020, by Patrick Ercolano

Credit: CC0 Public Domain

As American workplaces and schools shifted operations online to comply with social distancing during the coronavirus pandemic, a new area of concern arose: online privacy and user data collection.

Economist Itay Fainmesser of the Johns Hopkins Carey Business School focuses his research on social media and social networks. In recent months, he has expanded his focus to consider how the epochal COVID-19 pandemic has affected digital communication and how individuals' personal data can be collected and used by external parties during the crisis.

"There are benefits [to rising online activity], but there might also be severe consequences," Fainmesser says. "Even before COVID-19, there were concerns that information from various platforms and apps may end up being used by health insurance companies to determine premiums, or by potential employers for hiring decisions. There is also the worry that databases will be hacked—leading, for example, to identity thefts. The pandemic amplifies all of these concerns by a significant factor."

The Carey Business School reached out to Fainmesser, who is co-authoring a new working paper that examines the incentives of digital businesses to collect and protect users' information, for more insights into COVID-19 and digital privacy.

Have you seen data or heard anecdotal information that might indicate how the pandemic has affected online activity?

It is safe to say that there has been a sharp increase in online activity. As early as late March, *Forbes* reported that Internet usage went up 70%. Later on, in early April, a *New York Times* analysis showed that much of that increase happened in video chat, video games, and other forms of social networking. Facebook, for example, experienced a 27% increase in usage.

There have been COVID-19–related apps and websites on which people have provided personal information—to find testing sites, to get general information about the virus, to provide data for crowdsourcing platforms that aim to track the spread of the virus, and so on. The benefits seem evident, but might there also be harmful effects from providing this type of data?

Yes, there are benefits, but there might also be severe consequences. Even before COVID-19, there were concerns that information from various platforms and apps may end up being used by health insurance companies to determine premiums, or by potential employers for hiring decisions. There is also the worry that databases will be hacked—leading, for example, to identity thefts. The pandemic amplifies all of these concerns by a significant factor.

More generally, it is true that information is useful in

managing the pandemic: location data, interactions data, health history, and even current blood pressure, fever, and oxygen levels of individuals, or their medical concerns. The problem is that once the data is collected, it can have many other uses. For example, if scammers get their hands on a person's interactions over a week, or even a day, they stand a good chance of being able to execute a successful fraud, and if the U.S. Immigration and Customs Enforcement has access to this data, they can easily infer the likely immigration status of a user, as well as their whereabouts.

The way in which data is collected and protected matters a lot, and some websites and apps try to mitigate the risk. But this comes with a cost. A high-profile example is the Apple-Google contact-tracing app. The main idea is that the data is collected locally on a user's device using Bluetooth technology. If executed correctly, this means that even Apple and Google will not have access to much of the data. The downside is that the data also cannot be used by businesses and governments to learn about the spread patterns of COVID-19 and how people respond to government recommendations. The upside is that people will feel safer to use the app. This will increase adoption and therefore may improve contact tracing.

Your paper proposes "a two-pronged policy, which combines a minimal data protection requirement with a tax proportional to the amount of data collected." How would this work, and why do you advocate for it? Would such a policy become a function of a federal agency, such as the Federal Trade Commission?

We know that different digital businesses have different revenue models. Some businesses are more usage-driven—they rely on users being active on their platforms in order to derive revenues. Think of a service like Uber, or numerous online-dating platforms.

Other businesses are more data driven. Think of your average weather app (like Accuweather)—such businesses make most of their revenues from selling location data to the data aggregators. Google, Facebook, and many other platforms fall somewhere in between these two extremes: They rely on data for targeting ads to users but also need the users to be active and view or click on ads.

My work with Andrea Galeotti from London Business School and Ruslan Momot from HEC Paris shows that any business with a revenue model that is even partly data-driven has an incentive to collect more information than users and society would ideally want. One could imagine that such businesses might compensate users by providing more data protection: for example, by using a better firewall or by restricting API [or application program interface] access. However, it turns out that this is not necessarily the case. Many businesses have an incentive to invest less in data protection relative to what is socially desirable.

One of the policies that we propose to fix such inefficiencies (over-collection and under-protection of data) is to introduce a tax on data collection and a requirement of a minimal data protection level. When in place, the tax pushes businesses to internalize the effect that over-collecting data has on users, whereas a minimal data protection requirement ensures that businesses invest sufficiently in protecting the data they collect. We also find that an alternative solution could be to replace the tax on data collection with fines on data breaches—imposing a liability of sorts on businesses for the damage that data misuse cause to users.

Currently, in the United States, the FTC has a mandate to enforce a minimal protection level, which is a good thing. Nevertheless, our work suggests that the government can do better. One issue is that, nationally, there is no formal regulation of data collection. In practice, the FTC most frequently investigates firms that suffered data breaches and pursues fines, often through a combination of costly litigation and negotiations. This practice partially mimics our second policy suggestion of fines on data breaches. Our work suggests that a more systematic policy along these lines could increase consumer welfare.

The paper refers to "adversaries"—that is, entities whose use of data harms users. You

give the examples of a hacker attempting identity thefts and a government agency seeking to use the data to crack down on dissent. Are there adversarial actions that might be specific to the pandemic crisis?

An obvious one involves employers trying to obtain protected medical information on potential hires—for example, whether a potential hire has risk factors for COVID-19, or whether a potential hire was previously infected and is now immune. It is easy to see why employers would be interested in such information; yet this can lead to discrimination and other adverse effects. The fear of having their personal information exposed can therefore deter individuals from using online services that would otherwise be beneficial for them. This can be a big problem, and, in fact, our work shows that the loss to society from people's fear to use online services may be much bigger than just the direct damage from adversarial activity.

Who do you think is most responsible when data is used in ways that harm private individuals—the individuals themselves for surrendering their data, the businesses/platforms that collect and sell the data to third parties, or the third parties that use the data in sometimes troubling ways?

It may be easy to blame the third parties that misuse data and digital businesses that collect data, or even individuals who surrender their data. However, it is important to remember that, broadly speaking, many of the services offered online are beneficial and, to some extent, require that users surrender some information to businesses and platforms. The problem is that there are externalities. When a platform collects data, it can improve the services it provides, which is good for users. However, the data also attracts third parties, some of whom use the data in ways that harm users. That is, a platform's decision to collect more or less data affects users in unintended ways, and there is a constant trade-off. While it will be great to have the platform internalize these externalities by itself, this is exactly the type of situations in which regulation is helpful and can provide the right incentives for businesses.

Are there ways in which data is collected and used that ameliorate or exacerbate social, racial, and economic inequalities?

Definitely yes. In the context of COVID-19, you can think about the resource and treatment allocation problem, such as providing stem cell therapy to COVID-19 patients. Because treatment is scarce, we may like to allocate it to people who will benefit the most from the treatment. Now suppose that patients of certain social and economic characteristics are, on average, less likely to recover even with the treatment. Detailed data on social and economics characteristics can then be used to withhold treatment from such patients, thus reducing further their chance of recovery. On the other hand, detailed information on individual patients' health can help doctors make the decision based on individuals' health records rather than based on their perceived social and economic characteristics.

One of the tricky parts is that all of this is compounded by another issue. If users believe that information that they reveal in one context could be used to harm them in another context, they may choose not to reveal it to begin with, which may lead to a lose-lose situation. This is why every contact-tracing app and every treatment assignment protocol must trade off collecting data more aggressively with motivating individuals to reveal information to begin with.

Provided by Johns Hopkins University