A Proposal for Researcher Access to Platform Data: The Platform Transparency and Accountability Act

Nathaniel Persily

The disclosures of whistleblower Frances Haugen have provided a unique glimpse into Facebook's internal research and the ways that the company evaluates and addresses different harms on the platform. As explosive as the content contained in Haugen's revelations may have been, most of the reaction may have arisen from the mere fact that outsiders got an opportunity to see what Facebook knows (or could know) about its users and the information ecosystem it controls. Every inadvertent disclosure that comes out of Facebook gains such notoriety because most of what the public normally sees is subjected to rigorous vetting, corporate-speak and spin.

We should not need to wait for whistleblowers to blow their whistles, however, before we can understand what is actually happening on these extremely powerful digital platforms. Congress needs to act immediately to ensure that a steady stream of rigorous research reaches the public on the most pressing issues concerning digital technology. No one trusts the representations made by the platforms themselves, though, given their conflict of interest and understandable caution in releasing information that might spook shareholders. We need to develop an unprecedented system of corporate data-sharing, mandated by government for independent research in the public interest.

This is easier said than done. Not only do the details matter, they are the only thing that matters. It is all well and good to call for "transparency" or "data sharing," as an uncountable number of academics have, but the way government might set up this unprecedented regime will determine whether it can serve the grandiose purposes tech critics hope it will.

As with so many areas of tech regulation, transparency laws come with tradeoffs. In some cases, for instance, transparency might inhibit necessary security or harm prevention measures, as public disclosures about platform standards' enforcement might lead to gamesmanship by bad actors. When it comes to data access for research, the chief risk that needs to be addressed is user privacy. The shadow of Cambridge Analytica is cast over any academic access to user data, as that scandal involved a university researcher mishandling user data for the benefit of a private political consulting firm. If user data cannot be protected, then the public will not have faith in any government-mandated data-sharing program.

It is critical to understand at the outset, though, that user data is already collected and analyzed—but only by employees at the firms themselves. The threshold question when it comes to outside researcher access is whether the firms (and their employees who are tied to their profit maximizing mission) should have a monopoly on the insights that access to such data guarantees. Perhaps the firms should be prevented from gathering so much user data, but once they do, the public needs to be aware of it and to benefit from the insights that independent analysis will provide.

doi:10.54501/jots.v1i1.22

These benefits will be substantial. Most importantly, the mere fact that outsiders will have access to platform data will affect platform policies and behavior. Digital platforms, like any other association, institution or individual, will alter their behavior if they know they are being watched. Second, researcher access will enable evaluation and auditing of platform rules and interventions to gauge the responsibility of firms for problems that occur on their platforms. In other words, researcher access can enable outside auditing of actions taken by platform against users and content. Third, such access will inform policy makers seeking to regulate the platforms: only if they understand what is actually going on online might they be able to craft the appropriate regulations related to antitrust, privacy, advertising, child safety, content moderation or anything else. Finally, research on digital trace data is absolutely critical to understanding the sociology of the online information ecosystem, irrespective of potential links to policy. A large share of the human experience is taking place online. To understand it we need access to the relevant data.

The proposed legislation that follows—the Platform Transparency and Accountability Act—intends to design a data sharing program that protects user privacy to the extent possible while ensuring outside independent research on platform data. There are many ways to craft such a regime, and I hope this proposal sparks alternative approaches. The key features of any such system, though, must be (1) access by researchers not chosen by the firm to (2) the same data that the firms' own data analysts can analyze but (3) in a secure environment that minimizes any risks of disclosure of user private data.

Any proposal for outside access to platform data must wrestle with several questions (and this list is necessarily underinclusive). First, to which companies or platforms should such a regulatory regime apply? Second, who should have access? Third, to what data should they have access? Fourth and most important, how shall such access be regulated to protect both user privacy and research integrity?

1 Which Platforms?

Google and Facebook are first among (un)equals when it comes to the sheer volume of social media and digital trace data the firms possess. Any regulatory regime aimed at researcher access should be reverse engineered to capture those two firms in particular, as well as TikTok, which is quickly becoming a real competitor to YouTube. Twitter, which already provides more data than any other firm for researcher access, could also be added to the list, if the focus of the regulation is social media, per se.

But what about Amazon, Apple, and Microsoft? Researchers could gain enormous insight from access to those firms' data. Amazon, in particular, represents a monopoly of a different sort with data on users that could be extremely helpful to understanding the digital economy. Moreover, if the communications ecosystem is the target for research, what about the cable and cell phone companies, such as Comcast and Verizon? Surely, they possess data farther down the stack that could be helpful in assessing some relevant problems. A similar argument could be made for traditional media companies, e.g., Fox, or "new media" companies, such as Netflix.

To some extent, the universe of firms to which a data access regime would be applicable depends on the range of phenomena one considers worthy of study and the inability of researchers to gain insights from the outside. For those (like me) for whom the principal concern is the health of the information ecosystem and its impact on democracy, Google, Facebook, and Twitter reign supreme. The identification of the relevant firms, then, would include a definition of social media or search firms meeting some threshold of daily or monthly active users.

The Honest Ads Act[1] took a stab at such a definition in its attempt to force a disclosure regime on online political advertising. That bill defined an "online platform" as "any public-facing website, web application, or digital application (including a social network, ad network, or search engine) which...has 50,000,000 or more unique monthly United States visitors or users for a majority of months during the preceding 12 months." The legislative proposal that follows here lowers the bar to 40,000,000 monthly active users in order to capture TikTok as well.

2 Which Researchers?

Deciding which researchers shall have access is one of the biggest challenges to legislation in this area. "Researchers" come in many forms and a wide variety of civil society actors have an interest in the data held by internet platforms. However, some quality control must exist lest political operatives and propagandists repurpose themselves as "researchers" to gain access to platform data. It may also be that a separate regime for platform data access could be erected for think tanks or journalists, many of whom (such as Pew, ProPublica, the Markup, Buzzfeed or the Guardian) have done foundational research on these types of topics. Although categories such as journalists or think tanks may be difficult to cabin and enforce, transparency legislation should have as its goal making as much information available to as many watchdog groups, consistent with the privacy interests of users.

Focusing a data access regime on university-affiliated researchers has several advantages, however. First, a university is an identifiable "thing," and while low quality academic institutions exist, regulations can more easily specify the type of institutions that house the academics that should be granted access. Second, universities can be signatories to data access agreements with the platforms so as to add another layer of security (and retribution) against researcher malfeasance. Third, universities have Institutional Review Boards (IRBs) that can provide ethics and Human Subjects review for research proposals. Admittedly, IRBs have many well-known problems, but they are existing institutions that are in the business of evaluating research projects and the implications for human subjects. Fourth, in the wake of the Cambridge Analytica scandal, which involved an academic operating outside of his academic capacity, involving universities directly in the process of vetting and vouching for their researchers will make clear to the platforms which researchers are nested in a larger regulatory, contractual, and employment framework. Fifth, the National Science Foundation, which would play a role in vetting researchers, has established procedures in place to vet research projects and researchers from universities.

3 What Data?

In some settings, it is quite easy to define the data that should be made available for research. For instance, when drug trial data are made available for outside review, there are settled and familiar expectations for what kind of information the pharmaceutical company will provide. For Google and Facebook, though, the volume and variety of data they possess are so vast that any legally defined data access regime cannot simply say "turn over all available data to researchers." Some kind of principle should specify the range of data that should be available for research, or at least a process for deciding what data should be made available.

1. https://www.scribd.com/document/409188376/Mcg-19321

At a minimum, researchers should be allowed to analyze any data that is otherwise for sale to commercial entities or advertisers. If the datasets are available for a price, then they can be made available for academic analysis. Similarly, any data that goes into the preparation of government or other reports, such as those relating to enforcement of community standards (e.g., how many pieces of content were designated as hate speech and taken down) should be made available.

Beyond that, the key types of datasets that should be made available relate to "who" viewed/engaged with "what" content "when" and "how." In other words, to answer the most pressing questions relating to social media, we need data that can assess which types of people (though not individuals themselves) were seeing certain online content at certain times. The platforms already collect data of that nature. As part of the regulatory process, the platforms should be forced to identify datasets already in their possession, as well as data that are regularly collected. Then, the FTC, working with the NSF, should establish an application process for projects targeting those datasets. In addition, in order to prevent platforms from suddenly changing their data retention practices now that they are subject to oversight, the enforcement authority (here, the FTC) should have the authority to require the production of datasets deemed reasonably necessary for providing answers to questions researchers ask.

Moreover, the FTC should require the platforms to produce the code necessary to describe how the data were gathered and assembled, and to describe the chain of custody of the dataset. Researchers need to understand how the platform came up with the dataset. The platforms should also be fined if they misrepresent the origins of the data or otherwise produce a dataset inconsistent with what was requested.

All such data must be anonymized or pseudonomized. Moreover, if it can be done without degrading the quality of research, technologies such as differential privacy or the construction of synthetic datasets should be encouraged. In other words, user data must be presented in a format that protects user privacy as much as possible while maintaining utility for the research project.

4 How Shall the Data be Analyzed While Protecting User Privacy?

One of the reasons that the legislative proposal presented here vests enforcement authority in the FTC is that the FTC has been on the frontlines of enforcing privacy promises (to the extent that it is authorized to do so). The consent decree with Facebook following the Cambridge Analytica scandal, for which Facebook was required to pay a $5 billion fine, was negotiated and enforced by the FTC. In an ideal world, the United States, like Europe, might have a cabinet level position that is responsible for digital services, but if any progress on researcher access is to be made in the next two years, it will need to work with existing agencies. The FTC, working with the National Science Foundation, is the logical choice. That agency, then, will be responsible for vetting researchers and research projects and specifying the conditions under which research shall be conducted.

Although the government will be heavily involved in enforcing the program of researcher access, the datasets themselves should never be placed in government hands. It is absolutely critical that there be no risk of government surveillance or privacy intrusions as a result of this program. Alternative models of access would place the datasets in a government-controlled researcher sandbox, which would allow the government to control directly the environment in which data are analyzed. Doing so would necessarily run the risk that at some point in the future, government officials would see this research

environment as a honey pot for intelligence and law enforcement activities.

Under the proposal that follows, the data reside at the firm, which is responsible for maintaining security of the research environment and monitoring all research conducted therein. Researchers need to be monitored whenever they are in touch with the data. Every keystroke must be recorded as the data analysis is conducted. Researchers may not take any data out of the research environment without a privacy review being conducted. That includes immediately prior to publication—all publication drafts must be given a privacy review to ensure no data leakage. And in the event that a researcher engages in malfeasance both the researcher and the affiliated university shall be legally liable (even criminally liable) for any privacy violation. We need to make sure measures are in place that reassure the public that no individual's data is of interest to the research project, just the aggregated findings derived from them.

If the platform follows all applicable regulations concerning protecting privacy in the research environment, then it will be immune from suit for the fact that it made such data available under this program. To be clear, this does not immunize them from harms identified by the researchers. If the platform is discovered to be acting fraudulently or contributing to offline harm, then that information might later end up in a lawsuit or even a criminal prosecution. The point about legal immunity here is that the platforms cannot simultaneously be forced by the law to provide data to researchers and then be subject, for example, to a state tort law claim for violations of privacy.

5 Conclusion

Researcher access is only one component of transparency regulation, and transparency legislation is only one component of tech regulation. Nothing in this proposal should be seen as preventing broader reporting obligations for the platforms or construction of public facing APIs. Indeed, we should strive for a system in which any data on issues of public concern relating to the online information ecosystem should be available to the public, if it can be done in a privacy-protective way without other security risks.

One provision in the proposed legislation goes in that direction by dealing with the problem of scraping data from public-facing platforms. It would shield researchers from criminal or civil liability for scraping of public data from large platforms, like Facebook and YouTube. Of course, people disagree about what data, in fact, are "public" on these platforms. However, for researchers who scrape, they cannot be subject to money damages or criminal liability. This would not solve the problem faced by the NYU Ad Observatory, which had its accounts taken down by Facebook since it promoted a plug-in that allowed users to scrape their Facebook. But it would shield them from further actions, such as lawsuits that the platforms might initiate to get damages for terms of service violations arising from scraping.

A similar impulse underlies "Aaron's Law"[2] introduced by Representative Zoe Lofgren and Senator Ron Wyden. In a now famous and tragic episode, Aaron Swartz downloaded a large number of articles from the digital repository, JSTOR. In doing so, he breached the applicable terms of service for the website. Swartz was later arrested and prosecuted under the CFAA, which could have led to a penalty of 35 years in prison and up to $1 million in fines. However, he committed suicide before he was brought to trial. Aaron's Law would remove the threat of a felony prosecution for breaching terms of service in actions like this, if they do not cause significant economic or physical damage.

Just as researcher access is not coterminous with transparency, transparency does not

2. https://www.congress.gov/bill/113th-congress/senate-bill/1196

address all problems that tech regulation seeks to solve. Nothing in this proposal should be seen as taking the place of proposals to address competition and antitrust, child safety, advertising, content moderation, cybersecurity and privacy. Indeed, a proposal like the one that follows should be bundled together with federal privacy legislation or other broad regulations of the tech industry.

Researcher access, however, is a condition precedent to effective tech regulation. Right now, we do not know what we do not know. There are fundamental inconsistencies between platform's public representations and those made by whistleblowers, let alone those that feed conventional wisdom. For example, on the critical question of whether algorithms and recommendation systems are leading users toward extremism or pro-moting disinformation, the defenders and critics of platforms fundamentally disagree on basic facts. Policy makers need and deserve answers to these kinds of questions. Only if the government develops and mandates outside researcher access might we be able to get the answers necessary to make effective policy. Otherwise, we will be left with whatever studies the platforms choose to release or whatever research whistleblowers take with them on the way out the door.

Author

Nathaniel Persily is the James B. McClatchy Professor of Law at Stanford University and Co-director of the Stanford Cyber Policy Center.