Stanford | Cyber Policy Center
Freeman Spogli Institute

OPENING A WINDOW INTO TECH: THE CHALLENGE AND OPPORTUNITY FOR DATA TRANSPARENCY

STANFORD UNIVERSITY

Nathaniel Persily

OPENING A WINDOW INTO TECH: THE CHALLENGE AND OPPORTUNITY FOR DATA TRANSPARENCY

In the first year of the Biden administration, we should expect several new initiatives to be proposed relating to technology regulation. Content moderation, privacy, antitrust, and cybersecurity exist on a crowded agenda, although many devils exist in the details of any policy proposals in these areas. Moreover, as anyone who has engaged honestly with these issues recognizes, the impulses that drive policies in these domains often conflict with each other. Navigating the inescapable tradeoffs presents a real challenge to those with authority willing to jump into this political thicket.

Data transparency, however, represents a condition precedent to effective regulation in all of these areas. At present, we do not know even what we do not know concerning a host of pathologies attributed to social media and digital communication technologies. Pundits and policy makers think they have a handle on phenomena as varied as disinformation, hate speech, political bias in content regulation, and microtargeted advertising, but the publicly available data relevant to these problems represents a tiny share of what the platforms possess. The first step toward regulation of these platforms is to grant access to outsiders to bring to light the prevalence and character of the problems that are the target of regulation.

When critics describe Facebook and Google as “data monopolies”, they usually mean it in the antitrust sense. That is, the anticompetition “problem” with those companies is that they have amassed an enormous amount of data, which puts them in a privileged position to deliver targeted advertising as well as tweak their algorithms to maximize engagement. Of course, this amassing of data is also the source of their privacy and surveillance problems, but what sets them apart in the marketplace is the economic chokehold they have on would-be competitors, none of whom can ever

The research that drips out from the companies represents a tiny share of the potential insights that could be gained from their data were access more broadly available.

achieve parity given the years of data these companies have on billions of users.

In a different sense from their economic dominance, though, their status as data monopolies poses a distinct threat to democracy arising from their exclusive access to insights from the mass of data they have collected. Unfortunately, the insights of most value to them often concern how to keep people on the platform and how to target them better with advertising. To be sure, sometimes, after serious vetting from multiple authorities within the companies, internal researchers publish research that has great value to society, on issues such as polarization, news consumption, or even on the effect of certain platform interventions on the health of the information ecosystem. But the research that drips out from the companies represents a tiny share of the potential insights that could be gained from their data were access more broadly available. (These arguments are given fuller treatment in Chapter 13 of Nathaniel Persily & Joshua Tucker, eds., **Social Media and Democracy: The State of the Field and Prospects for Reform** (Cambridge Press, 2020).)

Researcher access is not a luxury good for academics; it is a precondition for sound policy concerning the information ecosystem and economy. The U.S. government, like its counterparts around the world, is rushing headstrong and blind toward regulation without a complete understanding of the problems they wish to solve. Legislators need better information about what

Researcher access is not a luxury good for academics; it is a precondition for sound policy concerning the information ecosystem and economy.

is happening on the major internet platforms and what is happening behind the scenes. For both legal and commercial reasons, the platforms are not going to provide that information willingly.

Of course, the platforms cannot make public all the sensitive data they have on their users. Doing so would be against the law and would be a fundamental violation of user privacy even if it were not. Nor should the platforms (or society) simply trust the government to hold onto the data for its purposes, which likely would include surveillance and criminal investigation. The policy challenge, therefore, involves creating a regime that respects user privacy, keeps user data out of the hands of government, and allows for public facing research that could lead to policy-relevant insights concerning the nature of the online information ecosystem.

Doing so requires legislation. The beginning of such an effort — and it is only a beginning – should include three components. The first concerns immunity from civil and criminal liability when platforms share data with vetted academics under prescribed circumstances. The second involves compulsion of the largest platforms, namely Facebook and Google, to share their data under the circumstances for which they would receive immunity. The third would immunize qualified researchers who scrape publicly available data for research purposes. A new "Platform Transparency and Accountability Act" with these three components could help turbocharge research on the harms and benefits of new communication technologies with a goal of producing well-informed public policy.

The policy challenge, therefore, involves creating a regime that respects user privacy, keeps user data out of the hands of government, and allows for public facing research that could lead to policy-relevant insights concerning the nature of the online information ecosystem.

I. SCOPING THE CHALLENGE AND POLICY RESPONSE

Enacting a compulsory data sharing regime is easier said than done. It is all well and good to say that platforms should share data with researchers, but legally defining which platforms, which data, which researchers, and under what circumstances proves especially challenging. Some models can provide clues, such as the protocols in place for the sharing of census, IRS, or sensitive health data, but none of them quite capture the breadth of the potential data available or the unique position and character of the relevant platforms.

A. 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. 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** 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." That law might capture more than just the "big three" social media platforms, but the form of the definition could be instructive in refining it further for purposes of researcher access.

It may be that different firms should be compelled to provide data than should be given immunity for voluntarily providing data. In other words, we should encourage a large number of firms to cooperate with approved researchers and be immune from liability for doing so. But when it comes to compelling certain firms to grant researcher access, that extreme measure should be reserved for Google, Facebook, and Twitter. Compelling smaller firms, such as Gab and Parler, let alone traditional or "new" media companies, to grant outside researcher access would raise constitutional concerns as to the First Amendment rights of these companies. (Indeed, as discussed later, such constitutional concerns will also be present for compulsion of the large companies, but at least their monopoly status might augur toward greater access in the public interest.)

B. Which Researchers?

One of the most difficult questions in considering researcher access concerns the selection and vetting process for researchers who will be granted access. “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. But categories such as journalists or think tanks are not amenable to any limiting principle.

Focusing a data access regime on university-affiliated researchers has several advantages. 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 that can provide ethics and Human Subjects review for research proposals. 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 framework.

Assuming universities are the universe from which to draw the researchers to be granted access, how should the researchers be selected and vetted? The platforms, the government, or some academic association, in theory, could be in the position of deciding which researchers get access. The law should prescribe a process for designating “qualified researchers” and qualified research projects, which could involve familiar procedures initiated by the National Science Foundation (NSF). The precise body doing the vetting is not critical, so long as the process is transparent and is one step removed from influences from both the platforms and the government.

C. Which 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.

As a threshold matter, any available dataset should be anonymized and stripped of personally identifiable information. To be sure, social media data are so rich that enterprising analysts might be able to reidentify people if they were hell-bent on doing so. But the datasets must be delivered in a format with protections that make it extremely difficult to do so. Moreover, as described below, monitoring of the research and researchers should be in place to prevent any reidentification.

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.

However, to get a handle on the prevalence of the most notorious problems on these platforms, information about user exposure, engagement, and other behaviors will be essential, as will data about the producers of content and the policies of the platforms. It would be difficult for the law to specify in advance the range of data that platforms must make available or for which their disclosure to academics would not incur liability. Here too, an outside party, such as the NSF, could be in a better position to evaluate which datasets can reasonably be provided by platforms to answer the most important research questions.

Such was the vision for Social Science One, the outside academic research initiative that I co-chaired until last year and that was established to serve as a broker between firms, such as Facebook, and the research community. In its original vision, the academics affiliated with Social Science One would pose questions to firms like Facebook and the firm would generate datasets that independent researchers could use to answer those questions. For various reasons, including those related to privacy, that model did not succeed. Instead, we moved to a model in which Facebook would provide a privacy-protected dataset and then researchers would apply, through the Social Science Research Council, to have access to it. This, too, proved suboptimal, because, in the end, Facebook said it could not find a way to provide broad access to the data it possessed while complying with its obligations under the FTC consent decree and applicable privacy laws around the world. The growing pangs of Social Science One, however, can be instructive for federal legislation that might compel platforms to provide researcher access and for the development of an agency with the power to define the datasets that might be made available for outside analysis. The experience has demonstrated that the necessary researcher access will not emerge voluntarily from the platforms. Their economic incentives counsel against it, and the applicable privacy laws (and consent decrees) create liability risks that far exceed the benefits – PR, public-spirited, or otherwise – of giving access to data to a bunch of academics.

the purpose of academic access is to combat the privileged monopoly position that insiders at the firms have over socially meaningful insights derived from the data in their possession. Private data has been and will continue to be analyzed by employees of the internet companies. The question is whether anyone else detached from the profit-making motives of the firms will have access to those same data to produce research in the public interest.

II. A THREE-PRONGED APPROACH TO REGULATION

Regulation to promote transparency through academic access to platform data must reckon with the serious privacy concerns that surround release of any social media data. Indeed, a transparency bill, such as that proposed here, should be adopted as part of a larger comprehensive privacy bill that makes clear how the balance shall be struck in limited, protected circumstances between privacy and other competing values. However, the purpose of academic access is to combat the privileged monopoly position that insiders at the firms have over socially meaningful insights derived from the data in their possession. Private data has been and will continue to be analyzed by employees of the internet companies. The question is whether anyone else detached from the profit-making motives of the firms will have access to those same data to produce research in the public interest.

A. Platform Immunity for Granting Researcher Access

Unless platforms are given immunity from suit, the data they willingly provide, if any, will not be amenable to the kind of detailed analysis that will produce the necessary public benefits. It is all well and good for academics to preach the value of public facing research, but the platforms are staring down multi-billion dollar fines if they leak user data. Moreover, the general counsels at the platforms tend to adopt the maximalist and most risk-averse interpretation of applicable privacy laws, sometimes dismissing research or public interest exceptions in such laws as vague and insufficiently shielding the company from liability. In short, if platforms are going to share their data, they need to know they will not be sued as a result.

At the same time, these privacy protections exist for a reason. The major platforms have a sorry history of protecting user privacy, and their business models depend on massive surveillance of users to gather information that enables targeted advertising. They possess data on some of the most private aspects of people's lives, and in some respects, they understand user psychology and behavior better than users themselves. Any pathway for research must ensure, to the extent possible, that an individual user's data will not be leaked to the public or even to the researcher.

To be clear, though, if researchers have data access akin to that of firm insiders, there is a risk that they will abuse their position. The task of policy, therefore, is to make sure that does not happen. This can be done at every stage of the research process. The law needs to specify, in detail, the privacy-protecting prerequisites for a platform to receive legal immunity when it shares data.

First, the law must identify a process for selecting/vetting researchers, research questions, and research designs. As noted above, this can be done by the NSF or a comparable institution. Similar methods related to IRS data, although not terribly routinized, have allowed researchers to do pathbreaking work on social mobility.

Second, the law must specify the environment in which the research will be conducted. Three options warrant consideration for where the data will be contained and the research conducted: (1) at the firm itself; (2) in a government supervised depository; or (3) at a university or other data hub, akin to the Federal Statistical Research Data Centers. Depending on the nature of the data, it may make the most sense for the data to remain at the firm under its control. Delivering private social media data to a government facility runs the risk of actual or perceived government surveillance of users. Depositing the data with secure, government-approved university facilities would make the data more broadly accessible and keep the data one step removed from the government. But at least in the short term, if for no other reason than to build confidence, the firms themselves could be made responsible for the secure facilities for the data environments for research.

What should those data environments look like? Again, analogies from similar secure facilities related to health, financial, or even national defense data will prove helpful. Among the key features of the environment, apart from all the expected cybersecurity measures, would be real time observation and recording of the researchers. Researchers need to know that they are being observed and that every key stroke is recorded. Doing so sends a signal to the researchers and the public alike that any attempt to misuse data will set off early alarms and that safety measures are built in to preserve evidence of the actions taken by the researchers as they analyze the data.

For similar reasons of privacy protection, all results garnered from the research should be evaluated before submission for publication. The firm should only be able to “object” to public release if the research will necessarily leak private data or otherwise violate the law. Any objections must be made in writing and explanations sent to the appropriate body (e.g., the NSF) overseeing the research.

Finally, the researchers, themselves, must be under threat of criminal punishment if they misuse data and attempt to invade the privacy of individual users. This may seem extreme, but researchers need to understand how seriously they must consider user privacy in their research. Any replay of Cambridge Analytica needs to be met with the strongest sanction. The public also needs to know that malfeasance will lead to fine and imprisonment.

If these conditions are met by the platform, however, then they should be immune from liability for the release of data to the researchers. To be clear, this immunity would not extend to the lessons learned from the data itself. For example, if the researcher discovers or publishes information pointing toward criminal or civil liability on the part of the platform, then such information could be used in a lawsuit or prosecution. The immunity for the platform should extend only to potential liability arising from the mere fact of releasing data to the researchers. In other words, they cannot be punished (under privacy laws or otherwise) for giving researchers data so long as the stringent privacy-protecting conditions are met.

B. Compelled Data Access for Major Platforms

Even if the law spells out a clear safe harbor for research, some platforms (perhaps even most) will still resist giving researchers access. As described above, even apart from potential legal liability, platforms worry about the PR or financial risk of what research might reveal. For data monopolies, though, such as Google and Facebook, this compelled access should be seen as a price they need to pay as the quasi-public utilities they have come to resemble.

Forcing any platform, big or small, to grant access to its data poses potential constitutional problems. The government could not, for example, require every website to reveal to government-approved researchers information about individuals who use its service. Indeed, doing so might not only raise privacy concerns but also First Amendment issues for both the platforms and their users.

For the largest platforms, though, this kind of law should be viewed as regulation designed to protect First Amendment rights, rather than threaten them. Because Facebook and Google exert unprecedented control over the speech marketplace, understanding what is happening on those platforms is critical to ensuring that users' speech rights are respected. This is not to say that the companies have the same responsibilities as the government. In fact, their First Amendment rights include the right to deplatform speakers and ban speech in ways that governments cannot. As with other corporate disclosure regulations in the public interest, however, requiring limited academic access to proprietary data should be viewed as a necessary step in preventing potential harms caused by the products themselves.

This kind of regulation could be seen as part of antitrust enforcement or as a condition for receiving the legal immunity platforms receive under Section 230 of the Communications Decency Act. In other words, for platforms that have achieved a status comparable to regulators of the public square, their scale comes with certain obligations, including allowing independent access to researchers who will gauge the effect of such platforms on democracy. If direct regulation is seen as legally precarious for one reason or another, then the large platforms could be given a choice. If they wish to enjoy the legal immunity for user-generated speech provided by CDA 230, then they must also agree to the academic access conditions detailed above. On the other hand, if protecting their data from outside analysis is sufficiently important to the firm, then they will be liable for the content on the platform. To be sure, such a "choice" raises questions of "unconstitutional conditions," but that is a court fight worth fighting, especially with respect to the largest data monopolies.

C. Immunity for qualified researchers who use publicly available data from the largest platforms

In addition to compelling the largest platforms to make their data available for research and shielding other platforms from legal liability were they to voluntarily make data available under restrictive circumstances, the law should protect researchers who amass publicly available data from the largest platforms. "Web scraping" and similar methods have been used by researchers when authorized research pathways, such as APIs, have been shut down. Often, these methods violate platform terms of service and in an extreme case, could lead to criminal liability on the part of the researcher. Applicable law, including the Computer Fraud and Abuse Act (CFAA), should be amended to carve out immunity, at least for approved researchers on the major platforms.

A similar impulse underlies "Aaron's Law" 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.

A similarly aggressive move occurred this summer when Facebook sought to shut down the NYU Ad Observatory. The project of NYU's engineering school scraped Facebook's political ad archive with a browser plug-in to perform research on political ads and to evaluate whether Facebook was following its declared disclosure policies. Facebook sent the researchers a warning letter threatening additional "enforcement action." It did so, it said, because these terms of service protect user privacy and comply with the terms of Facebook's consent decree with the FTC.

The rules against scraping and unauthorized accessing of available data derive from real concerns about hacking, cyber security, and privacy protection. For their products to function, for example, platforms may need to limit the numbers of queries per user or the degree to which users can download all the content on the service. By placing content on the web, companies are not consenting to hacking the entire back end of their systems by clever users or the reproduction of all of their content on someone else's server.

However well intentioned, the applicable laws should include a carve-out for legitimate research performed by university researchers. Congress should pass Aaron's law, but it should also do much more. It should make clear that researchers cannot be prosecuted for breaking the terms of service of the largest platforms—Facebook, Google and Twitter—in the course of their research. Any university researcher with a project that has been approved by the university's institutional review board should not be subject to any criminal or civil liability for scraping Facebook, Google or Twitter. The platforms could still shut down the accounts of the researchers who break the terms of service, but they cannot appeal to the courts or the U.S.

Legitimate privacy concerns exist, for example, related to researchers (or any outsiders) gaining access to platform data. But policy should address those concerns directly, rather than assume that a zero-sum game exists between privacy and transparency.

Attorneys to follow suit. Moreover, the platforms too, for reasons similar to those expressed above, should be immunized for actions taken by legitimate university researchers who scrape their sites in the course of their research.

CONCLUSION

As with Section 31 of the new Digital Services Act proposed in Europe, the kinds of proposal presented here should be bundled together in a larger package of technology reforms, including privacy protection, competition rules, cybersecurity, and content moderation regulation. Researcher access to platform data undergirds all of these policies, however, as it results in the kind of knowledge production that will inform good tech regulation. If the other values are maximized, however, transparency by way of researcher data access may be unfortunate collateral damage. Legitimate privacy concerns exist, for example, related to researchers (or any outsiders) gaining access to platform data. But policy should address those concerns directly, rather than assume that a zero-sum game exists between privacy and transparency. A first step in doing so requires a recognition that the biggest platforms, particularly Facebook and Google, are qualitatively different types of monopolies than preexisting firms. If they are going to continue to enjoy unrivaled market power, they also must take on additional responsibilities. One of the most important responsibilities in that vein is the obligation to share their data with qualified researchers seeking to answer some of the most important questions as to the impact of new digital communications on society.

ABOUT THE AUTHOR

Nathaniel Persily is the James B. McClatchy Professor of Law at Stanford Law School, with appointments in the departments of Political Science, Communication, and FSI. Prior to joining Stanford, Professor Persily taught at Columbia and the University of Pennsylvania Law School, and as a visiting professor at Harvard, NYU, Princeton, the University of Amsterdam, and the University of Melbourne. Professor Persily's scholarship and legal practice focus on American election law or what is sometimes called the "law of democracy," which addresses issues such as voting rights, political parties, campaign finance, redistricting, and election administration. He has served as a special master or court-appointed expert to craft congressional or legislative districting plans for Georgia, Maryland, Connecticut, New York, North Carolina, and Pennsylvania. He also served as the Senior Research Director for the Presidential Commission on Election Administration. In addition to dozens of articles (many of which have been cited by the Supreme Court) on the legal regulation of political parties, issues surrounding the census and redistricting process, voting rights, and campaign finance reform, Professor Persily is coauthor of the leading election law casebook, The Law of Democracy (Foundation Press, 5th ed., 2016), with Samuel Issacharoff, Pamela Karlan, and Richard Pildes. His current work, for which he has been honored as a Guggenheim Fellow, Andrew Carnegie Fellow, and a Fellow at the Center for Advanced Study in the Behavioral Sciences, examines the impact of changing technology on political communication, campaigns, and election administration. He is codirector of the Stanford Cyber Policy Center, Stanford Program on Democracy and the Internet, and Social Science One, a project to make available to the world's research community privacy-protected Facebook data to study the impact of social media on democracy. He is also a member of the American Academy of Arts and Sciences, and a commissioner on the Kofi Annan Commission on Elections and Democracy in the Digital Age. Along with Professor Charles Stewart III, he recently founded HealthyElections.Org (the Stanford-MIT Healthy Elections Project) which aims to support local election officials in taking the necessary steps during the COVID-19 pandemic to provide safe voting options for the 2020 election. He received a B.A. and M.A. in political science from Yale (1992); a J.D. from Stanford (1998) where he was President of the Stanford Law Review, and a Ph.D. in political science from U.C. Berkeley in 2002.