U.S. HOUSE OF REPRESENTATIVES
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
SUBCOMMITTEE ON INVESTIGATIONS AND OVERSIGHT

HEARING CHARTER

The Disinformation Black Box: Researching Social Media Data

Tuesday, September 28, 2021
10:00 a.m. EDT – 12:00 p.m. EDT
Zoom

PURPOSE

The purpose of this hearing is to discuss how researchers are able to access and analyze data from social media companies. Researchers will testify about their work looking into the spread of misinformation and disinformation on social media platforms and how platforms drive traffic to advertisements and promoted posts. The hearing will also explore the limitations of current tools, techniques, and datasets for researching social media platforms and how researchers have utilized information available to advertisers to flag privacy concerns to the platforms. The hearing will examine how the Federal government can contribute to the ethical study of social media's impact on society while protecting the privacy of users.

WITNESSES

- **Dr. Alan Mislove**, Professor and Interim Dean, Khoury College of Computer Sciences, Northeastern University
- **Ms. Laura Edelson**, Ph.D. Candidate and Co-Director of Cybersecurity for Democracy at New York University
- **Dr. Kevin Leicht**, Professor, University of Illinois Urbana-Champaign Department of Sociology

OVERARCHING QUESTIONS

- What kind of data can and should be made available by social media companies in order to understand the spread of misinformation and disinformation and its impact on society?
- What kind of research is possible without privileged access to data from social media companies, and why is it important that researchers independent of social media companies have access to data?
- What are the limitations of current tools, techniques, and data sets used to analyze social media?
- What do we know about how misinformation and disinformation spreads on social media platforms and the effectiveness of platforms' monitoring and moderation techniques?
- How can the Federal government assist researchers in accessing data from social media companies that can help shed light on the spread of misinformation and disinformation?

Cambridge Analytica

The Cambridge Analytica scandal thrust into focus the issues of access to social media data for research purposes and the privacy breaches and political manipulation that ensued. Cambridge Analytica was a voter-profiling company that partnered with an outside researcher to collect data allegedly for academic purposes, but the data was in fact used in contracts with the 2016 presidential campaigns of Ted Cruz and Donald Trump.[1]

Cambridge Analytica harvested Facebook user data via personality quizzes. The quizzes were developed with an academic who got an app approved by Facebook on the basis that the data collected would be used for academic purposes. The app harvested data from users – informed about the collection via fine print – and from their Facebook friends, who were not informed. 270,000 users consented to participate in the personality quizzes and 50 million users' data were swept up in the collection, with 30 million containing enough personally identifiable information to create "psychographic profiles" incorporating records outside Facebook.

The U.S. Federal Trade Commission (FTC) initiated an investigation into Facebook in March 2018 following the allegations that Cambridge Analytica's actions violated a 2012 decree requiring notification when user data is shared beyond the agreed upon privacy settings.[2] The inquiry concluded in July 2019, when the FTC commissioners approved a $5 billion penalty for violating the 2012 order and established new accountability mechanisms for protecting user privacy, including an internal committee and compliance officers as well as biennial external assessors. The privacy requirements explicitly include third-party apps.[3]

State of the Available Data

A majority of Americans report using social media, with YouTube and Facebook drawing the eyes of 81 and 69 percent of Americans, respectively.[4] This makes social media platforms a wealth of information on individuals' habits, preferences, and beliefs. It also makes these platforms extremely valuable to advertisers – social media ad revenues totaled $41.5 billion in 2020.[5] Social media data is extremely valuable to researchers looking to understand what users are consuming on these platforms and how that content shapes their beliefs and behavior on- and offline.

The primary way social media platforms publicize data for public use is through Application Programming Interfaces (APIs). APIs are platforms on which companies publish data for use by third parties, including app developers, business partners, and researchers.[6] Researchers can write code that combs through the data available through the API, which they typically can gain access to after being verified by the platform.

[1] https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

[2] https://www.washingtonpost.com/news/the-switch/wp/2018/03/20/ftc-opens-investigation-into-facebook-after-cambridge-analytica-scrapes-millions-of-users-personal-information/

[3] https://www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restrictions

[4] https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

[5] https://www.cnbc.com/2021/04/07/digital-ad-spend-grew-12percent-in-2020-despite-hit-from-pandemic.html

[6] https://www.ibm.com/cloud/learn/api

APIs are limited in terms of data transparency because social media companies control what is shared and who has access. Researchers must also rely on social media companies to fix technical glitches when they occur. Researchers have expressed their desire for platforms to share more data and access, including items that advertisers have access to. This includes:

- the number of views on an ad (called "impressions");
- information on audience characteristics, such as gender;
- historical archives of advertisements beyond political ads;[7]
- the ability to run scripts to evaluate the vast contents of the Facebook Political Ad Library and other public databases provided by platforms; and
- information on how advertisements are targeted to specific individuals.

At the moment, only non-political advertisements that are currently running are available on Facebook's Ad Library. Political advertisements are available in the library for seven years, though Facebook itself classifies what qualifies as an ad about "issues, elections, or politics."[8] Furthermore, Facebook is making some impression data available in a quarterly report, but these reports show very little granularity, excludes impressions via private groups and pages, and only profiles the top 20 posts that were viewed in that period.

The tension between the public pressure to provide more data and the potential backlash on social media companies was exemplified in coverage of Facebook grappling with how to handle its CrowdTangle platform. CrowdTangle is a Facebook tool that provides access to public content on Facebook, Instagram, and Reddit. It allows third parties to track how and where public posts are shared and interacted with, though it does not provide access to impressions or demographics.[9] After journalists used the tool to show how frequently top performing posts originated from extremist and unreliable pages, some executives pushed a pivot to curated data sets, and reorganized the CrowdTangle team.[10] The CrowdTangle tool is still active, and Facebook has published curated data sets through its Facebook Open Research and Transparency (FORT) program.[11]

Facebook Revoking Access to Researchers

On August 4, Facebook revoked a team of New York University (NYU) researchers' access to the platform.[12] The researchers, including witness and PhD candidate Laura Edelson, were collecting data about political advertisements through a browser extension called the Ad Observatory. The extension had been running since September 2020. Volunteers downloaded the browser extension and consented to data collection on the political ads shown to them on

[7] https://blog.mozilla.org/en/mozilla/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like/
[8] https://www.facebook.com/help/259468828226154
[9] https://help.crowdtangle.com/en/articles/4201940-about-us
[10] https://www.nytimes.com/2021/07/14/technology/facebook-data.html
[11] https://research.fb.com/data/
[12] https://about.fb.com/news/2021/08/research-cannot-be-the-justification-for-compromising-peoples-privacy/

Facebook.[13] After Facebook issued a cease-and-desist letter in October,[14] the parties negotiated access terms for nine months and reached a standstill, at which point NYU turned its collection back on. Facebook then disabled the researchers' accounts, informing the team via an automated email message.[15]

Facebook alleges that the NYU team was engaging in unauthorized scraping that jeopardized the privacy of its users. Data scraping is when an automated program is used to collect information from another website or app.[16] The information can then be made available to third parties. Scraping itself is not inherently problematic. It is behind the tools we use every day, enabling search engines to deliver relevant results and price comparison tools to aggregate information across e-commerce platforms. Facebook's policy bans all unauthorized scraping, regardless of whether the data being accessed is widely available.[17] NYU researchers object to the characterization of their collection methods as scraping, saying that their extension collects only the ads seen by consenting users and not private information of users or their friends. Regardless of the particulars of NYU's research, scraping is a research technique that can be explicitly authorized by Facebook.[18] However, at the time it disabled the NYU researchers' accounts, Facebook posted a blog claiming that under its privacy program established pursuant to the 2019 post-Cambridge Analytica FTC order, the Ad Observer extension posed too serious a risk to user privacy, and opted to revoke researchers' access instead of authorizing the activity.

On August 5, the FTC sent a letter to Facebook noting that the action taken against the NYU researchers was not, in fact, required pursuant to the Facebook's consent decree with the FTC. The Acting Director of the Bureau of Consumer Protection noted that the FTC was not notified prior to Facebook's erroneous invocation of the decree. Furthermore, the letter explicitly condoned "good-faith research in the public interest" and noted Facebook's ability to make access exceptions to support such research.[19]

The dispute between Facebook and NYU serves as a helpful and timely illustration of the control Facebook has over the access researchers have to the platform. NYU's browser extension is still active and collecting data from users who have downloaded it, but the researchers' accounts remain locked, though Facebook has since acknowledged that this decision was not forced by the agreement with FTC.[20]

Algorithms

The transparency push on social media companies goes beyond user characteristics and ad impressions. Researchers are looking to understand how content reaches users. Along with the

[13] https://www.wsj.com/articles/facebook-cuts-off-access-for-nyu-research-into-political-ad-targeting-11628052204

[14] https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533

[15] https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html

[16] https://www.targetinternet.com/what-is-data-scraping-and-how-can-you-use-it/

[17] https://about.fb.com/news/2021/04/how-we-combat-scraping/

[18] https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html

[19] https://www.ftc.gov/news-events/blogs/consumer-blog/2021/08/letter-acting-director-bureau-consumer-protection-samuel

[20] https://www.wired.com/story/facebooks-reason-banning-researchers-doesnt-hold-up/

benefits of access to a near limitless amount of information, the Internet has also created information fatigue. To combat this and sell advertisements, social media platforms use algorithms to sort posts for users based on relevancy in order to prioritize which content a user sees and to increase the likelihood of the user engaging with that content. For example, Twitter, Facebook, Instagram, Reddit, TikTok, YouTube, Snapchat, and LinkedIn all create a personalized "feed"—also known as a timeline or newsfeed—from the content generated by the accounts followed by the user and/or the content the user has browsed previously. The choices that these algorithms make significantly dictate a user's experience. For example, a platform's algorithm may prioritize posts from user's friends, family, and groups to which they belong over sources of accredited news. Moreover, algorithms may place users in "filter bubbles" by only presenting content from like-minded people or amplifying selective exposure to information.[21]

The algorithms that make content decisions on social media are functionally black boxes, which means it is difficult to understand why algorithms make the decisions that they do. Training data is fed to the bottom layer of the algorithm's network, and as it passes through the succeeding layers it gets multiplied and added together in complex ways until it finally arrives at the output layer in its transformed final state. Due to the complex middle layers, observers can only effectively assess this process by reviewing an algorithm's inputs and outputs. Researchers and companies are working to improve algorithmic explainability, a topic of AI research that focuses on getting algorithms to explain their decisions, in contrast to the opaque "black box" model wherein even AI designers may not be able to track the AI's decision-making process. However, improving the explainability of algorithms has often come at the cost of accuracy of outputs.[22]

Researchers also lack the access necessary to study social media algorithms. This is in part because companies consider their algorithms to be trade secrets. Companies also benefit from the opaqueness of their algorithmic content decisions because they do not have to justify each decision. As a result, researchers are only able to assess certain outputs of social algorithmic curation, which has significantly limited this field of study. Even outputs are difficult to assess. For example, current tools offered by social media platforms do not allow researchers to retrieve the information that users see in their social feed in order to assess algorithmic choices. Similarly, confounding factors, such as the tendency of people to seek out others with similar preferences, make estimating the effects of algorithmic recommendations difficult to assess. As a result, researchers often must get creative to measure algorithmic effects, such as by using bots to create randomized field experiments.[23] Researchers looking into social media algorithms often conduct their studies without privileged access to platforms' data, putting out organic posts and paying for advertisements in order to track the metrics that are granted to paying customers but not researchers or ordinary users of the platforms.

The Spread of Misinformation and Disinformation

Many social media researchers focus on how misinformation and disinformation spread across platforms. Particularly since the 2016 Presidential election, when the public became aware of the impact of "fake news" on the broader political discourse, researchers have sought to examine

[21] https://5harad.com/papers/bubbles.pdf
[22] https://www.nature.com/articles/s42256-019-0048-x
[23] https://dl.acm.org/doi/fullHtml/10.1145/3447535.3462491

how untrustworthy information is spread, how platforms do or do not monitor and moderate it, and how it impacts society at large. A recent study found that on Facebook, publishers that share misinformation get six times as much engagement as trustworthy news sources.[24] While Facebook has pushed back and said that engagement data is not indicative of how many people view the misinformation relative to trustworthy posts, the company does not make impression data available to researchers.

Without more extensive data on the spread of misinformation and disinformation, it is difficult for third parties to assess how effective social media platforms' monitoring and moderation techniques are at managing dangerous content. A research group using Facebook's CrowdTangle tool found that pages sharing election misinformation tripled their interactions from October 2019 to October 2020, despite Facebook's partnership with third-party fact checkers to mark these posts as misinformation.[25] The same group used non-privileged access to Facebook to demonstrate the algorithm's push of anti-vaccine content via "related pages" suggestions.[26]

In July, the U.S. Surgeon General issued a report classifying misinformation as a public health threat, calling out the confusion over COVID-19 vaccines, preventative measures, and unproven treatments frequently stoked by malicious actors looking to profit financially or politically.[27] The report notes that health misinformation is not a new problem – it contributed to over 330,000 AIDS deaths between 2000-2005 – but that the changing information environment enabled by social media escalates the threat to unprecedented levels. Social media companies have pledged to combat misinformation and disinformation, taking steps like banning political advertisements around elections.[28] But it is imperative that third-party researchers who do not financially benefit from the spread of malicious content have sufficient access in order examine the potential threat of social media misinformation to public health and to democracy.

Ethics and Social Media Research

As with other forms of research focused on human subjects, the use of social media data in research poses important ethical concerns. To date, there is no clear consensus on an ethical framework for researchers entering this field like there is in other disciplines such as bioscience and health research. This situation is further complicated by the differing ways that social media platforms work with researchers as well as the ad hoc way in which disputes are resolved. As a result, different institutions and institutional review boards (IRBs) have created different guidance and recommendations for ethical social media research.

There are many ethical challenges regarding research that uses social media data, including privacy and consent, anonymity and confidentiality, authenticity of subjects, data security and management, and more. For example, informed consent can be difficult to acquire in social media research. In more traditional research approaches, informed consent is usually built into the research design, such as through consent forms. On the other hand, participants in social

[24] https://www.washingtonpost.com/technology/2021/09/03/facebook-misinformation-nyu-study/
[25] https://secure.avaaz.org/campaign/en/facebook_election_insurrection/
[26] https://secure.avaaz.org/campaign/en/fb_algorithm_antivaxx/
[27] https://www.hhs.gov/sites/default/files/surgeon-general-misinformation-advisory.pdf
[28] https://www.nytimes.com/2021/03/03/technology/facebook-ends-ban-on-political-advertising.html

media-based research are often unaware of their participation and do not give prior written, informed consent. While this consent may be given in the form of a platform’s terms and conditions, ethical questions remain whether users truly read and understand these agreements. Similarly, anonymity is a key consideration in research ethics, particularly in qualitative research practices or when data sets are shared outside of the original research team.

Grant-making agencies, such as the National Science Foundation (NSF) and the National Institute of Health, have created numerous methods and procedures to allow for research in sensitive topics while protecting the privacy and security of research participants. One example of this is the NSF’s National Center for Science and Engineering Statistics (NCSES), which has created numerous procedures for the data it licenses to researchers. Beyond directly funding research, agencies have had little direct involvement with opening secure access to social media data.