World view

By Simon Hegelich

Facebook needs to share more with researchers

Private companies get free access to data that are more informative than what researchers are forced to compete for.

Smarting from the Cambridge Analytica debacle in 2018, Facebook promised a research initiative to give academics access to its data. I lead one of a dozen research groups that were granted access in February to a large data set, which is still less than was promised, after more than a year of delays. Our project aimed to determine whether disinformation campaigns have a measurable effect on polls. But the data provided are nearly useless for answering this and many other research questions, and are far inferior to what Facebook gives private companies. The company must do better.

For me, the scandal of Cambridge Analytica was that the data could be used to manipulate elections. (The company acquired Facebook users' data without consent, to build profiles of voters. It also worked for the campaign to elect US President Donald Trump.) For Facebook, the scandal was that a researcher gave data obtained in accordance with the company's policy to a third party.

Subsequently, Facebook considered ways to share data with academics while keeping user information private. Maintaining that privacy is a real concern. But my opinion is that Facebook is working with university researchers mainly to gain positive news coverage and to reduce political pressure on the company.

In April 2018, Facebook and a group of academics helped to set up a non-profit initiative called Social Science One to act as a data broker between the company and researchers. Its announced goal was to shed light on how Facebook affects society, particularly democracy and elections. But the released data set includes no information about the 2016 US presidential election or the coming one, or the Brexit referendum, all of which are plagued with worries about misinformation and foreign manipulation. Social Science One should be explaining why it is failing to meet its goals, not joining with Facebook to trumpet its work.

Social Science One grants access to only a handful of certified researchers who use data only behind a firewall. The data provided are a circumscribed sampling of all the data that were collected. Only URLs that have been shared publicly more than 100 times between January 2017 and August 2019 are available. Instead of information from single users, researchers see numbers of shares, likes and so on from all users in a specific country and demographic. For example, one URL of nature.com could have been viewed 3,987 times by females aged 25–35 in the United States. To ensure further anonymity, statistical 'noise' is added.

My first reaction to these precautions was that Facebook is taking privacy really seriously. But looking more closely, I noticed contradictions. Facebook's free service CrowdTangle allows searches by URL to show who has reacted to a post and how, although it is not that useful to researchers like me because it is difficult to scan for patterns that you do not know about in advance.

> "My opinion is that Facebook is working with university researchers mainly to gain positive news coverage."

Simon Hegelich is a professor of political data science at the Technical University of Munich in Germany. He is not affiliated with the university's artificial-intelligence ethics centre, funded by Facebook.
e-mail: simon.hegelich@hfp.tum.de

Other services are more relevant. Graph API, for example, allows private companies, but not university researchers, to apply for access (doing so requires a VAT registration). But from a privacy perspective, all services deal with the same user information, which is not accessible in the URL data set released to researchers.

When I learnt about this discrepancy in the levels of privacy in December 2019, I thought it must be a mistake. I flagged it to Facebook and the chairs of Social Science One. Nothing changed, although Social Science One complained publicly about lack of access.

Anyone with statistical knowledge who reads the codebook carefully will realize that the data have been rendered nearly useless for tracking the influence of misinformation or manipulation. There is hardly any variation in the data, especially compared with the noise. That results in nonsense values, such as URLs viewed by minus one million users. So it's impossible to ask important questions about, say, URLs referring to fake news regarding Angela Merkel, the German chancellor, or about the increase in angry reactions in a specific country in a given week. In other words, crucial questions — especially those that might prompt further scrutiny of Facebook — cannot be addressed with the platform that's supposed to help answer them. Social Science One is not doing enough to challenge Facebook on its approach. Part of the problem, I think, is that the initiative is not truly independent. Its chairs were co-appointed by the non-profit organizations that support it and by Facebook. I am not saying that Social Science One has done something wrong, only that Facebook can apply pressure to it in ways that would be impossible for other scientific organizations that have institutional barriers to secure their independence.

Where should we go from here? Social Science One should reveal its direct and indirect connections to Facebook; its chairs and directors should be appointed by outside experts. Facebook should make available to certified researchers, as it originally promised, an 'advanced' data set of URLs that extends to at least January 2016. The current Social Science One data set should be released publicly to the entire research community. Similar commitments should be required of other digital platforms.

We still do not know whether Facebook and other platforms have a negative effect on our democracies. If these companies do not release sufficient data for us to analyse such questions soon, democratic governments must pass regulations requiring that they do.

EDI&SEPPTUM