Why Are We Just Finding Out Now That All Two Billion Facebook Users May Have Been Harvested?

Kalev Leetaru FORBES
I write about the broad intersection of data and society.

The Facebook logo. (Jaap Arriens/NurPhoto via Getty Images)

Facebook kept the privacy headlines going yesterday when it acknowledged that "malicious actors have also abused [the platform] to scrape public profile information… given the scale and sophistication of the activity we've seen, we believe most people on Facebook could have had their public profile scraped in this way," while Zuckerberg himself offered that "I would assume if you had that setting turned on that someone at some point has access to your public information in some way." In short, the company acknowledged what I've said many times before – likely the entirety of Facebook's two billion public profiles (and quite a few private profiles) are archived in repositories all over the world by academics, companies and criminal actors, not to mention countless governments. The big story was not Facebook's confirmation of this, but rather why the company took until yesterday to confirm it.

For years many like myself have warned of the sheer magnitude of Facebook scraping that is performed everyday across the world by academic and commercial interests (government surveillance is a whole different world into itself). Academics in particular have long harvested Facebook data in bulk with the full permission of their ethical oversight boards, frequently with US federal government funding from agencies like NSF and with the results published in top academic journals. These archives are almost never deleted and are frequently shared across the world, with mailing lists and conference sidelines filled with offers of bulk downloaded data.

Bulk harvested datasets frequently find their way from academia into commercial for-profit startup enterprises as universities increasingly encourage their faculty to commercialize their research, with surprisingly few institutions asking many questions about the data freeing flowing from their institutions to their faculty members' side ventures. After all, the Cambridge Analytica story is at its core that of an academic allegedly making research data available for a for-profit company without ensuring all of

the necessary permissions had been received for the transfer – a story that happens every day at universities all over the world.

To those familiar with academic and commercial data practices, Facebook's revelation that potentially up to the entirety of its user community, all two billion of them, have had their public profile data harvested without their knowledge, is old news. The only surprising part is why Facebook is just now, in April 2018, acknowledging the scope of unauthorized data harvested and why it is focusing only on a narrow slice of that harvesting, rather than the myriad other forms of bulk harvesting that are used against its systems day.

Over the past year I have repeatedly asked Facebook for its stance on bulk harvesting and research use of its users' data. Last February I asked the company if it had comment on the mass harvesting of data by commercial enterprises for political purposes and whether it had any policies prohibiting the use of personality quizzes or other apps that bulk harvested profiles. In June I asked it, in light of all of the ways Facebook itself was conducting research on its users, whether it might consider offering users the right to opt-out of having their personal data exploited by Facebook for research. In September, in the aftermath of the controversial "gaydar" study that claimed to be able to estimate someone's sexual orientation from their photo and used a large volume of harvested Facebook data, I asked whether the work's mass harvesting of profile photos was of concern to the company. Just last month I asked whether Facebook was planning to request that large holders of data harvested from the platform delete their archives or whether it planned to request that bulk Facebook datasets available for download be restricted to university researches and exclude commercial researchers. Not to mention countless other requests for comment about various Facebook research use of private user data. In every case the company's response was silence.

If Facebook was so concerned about bulk harvesting and use of its users' data, it certainly would seem that the company would have taken every opportunity to state that bulk harvesting, archival and commercial exploitation of private user data was something it was concerned about. It could comment that it was working to identify bulk harvesting, to

request that companies and universities delete those archives or that it was asking that universities restrict access to the large harvested datasets they make available for download, limiting them to academic and not commercial uses. Instead, radio silence until the company lost control of the privacy narrative and suddenly decided now was the time to say it was shocked by how its data was being harvested and would take steps to reign it in.

Where was all this concern a year ago?

More to the point, in its statements yesterday, Facebook offered that its estimate of two billion profiles being downloaded was based on "the scale and sophistication of the activity we've seen." Why was Facebook not monitoring its system logs from the beginning looking for bulk harvesting activity?

It turns out they were. Indeed, when the Obama campaign bulk harvested data from the platform, the company's security teams immediately detected the bulk harvesting and approved it.

If Facebook was so easily able to detect the Obama campaign harvesting, why didn't see all of these other harvesting efforts? Given that it was able to identify that up to its entire user community of two billion people have had their public profiles harvested at least once, it is clear the company did not lack the logging or analytic tools to identify such activity. The company did not respond to a request for comment.

In reality, it likely comes down to the fact that Facebook's early years were defined as becoming a data hub, the plumbing around which they would remake the web in their image. By being open with their APIs and allowing harvesting of user data, they would become the invaluable must-have nexus of the evolving web. Twitter's early trajectory was very similar, in which it made its firehose freely accessible and worked hard to become a central web nexus. Today Facebook has focused instead on building a walled garden which it wields total control over and focusing on bringing data into its platform, rather than letting it out. Twitter, too, has locked up its firehose to paying customers only and tightened its API policies.

In a call yesterday with reporters, Zuckerberg offered that "life is learning from mistakes" and that "we're an idealistic and optimistic company … we know now we didn't do enough to focus on preventing abuse and thinking through how people use these tools to do harm." The problem is that when a platform that holds the digital lives of two billion people learns from its mistakes and naively believes there aren't bad actors out there working hard to harvest its valuable data, then those two billion people lose their

irreplaceable privacy in the process and face greater exposure to identity theft, bullying and other ramifications.

Putting this all together, the real story yesterday was not that all two billion of Facebook's users may have had their public information harvested – that's old news well known to those that study data use and privacy. The story was why Facebook waited until April 2018 to finally confirm it and why for the past year it has refused to step up and condemn the activities it is now saying are incompatible with its corporate vision. If mass commercial harvesting is wrong, why did it condone it in 2012 and why didn't it forcefully denounce such behavior and take action to restrict and remediate it when asked repeatedly about it last year? In the end, it is nice to say that Facebook is learning from its mistakes, but in the real world there are real consequences to Facebook's actions and as a platform, its reach and influence over society is so great that one must ask whether Silicon Valley's mantra of moving fast and breaking things is not the right mindset in a world in which the things being "broken" are people's lives.