It's impossible to know exactly what data Cambridge Analytica scraped from Facebook — but here's the kind of information apps could access in 2014

Avery Hartmans Mar 22, 2018, 7:19 PM

Facebook CEO Mark Zuckerberg. REUTERS/Shu Zhang

- **We don't know exactly which information Cambridge Analytica obtained from as many as 50 million Facebook users.**

- **But Facebook's old API, the software tool that gives third-party apps access to user data, provides insight into what would have been possible to obtain before Facebook changed its policies in 2014 to prevent that sort of data scraping.**

- **Back then, when users clicked "allow" before using a third-party app, they were often giving those developers access to a trove of their friends' personal information.**

- **That data included information like interests, likes, location, political affiliation, relationships, photos, and more.**

We'll most likely never know exactly what information Cambridge Analytica obtained from Facebook users — but by taking a look at Facebook's old way of doing business, we can hazard a guess.

Over the weekend, a whistleblower revealed that the British data company Cambridge Analytica illicitly obtained information from as many as 50 million Facebook profiles by abusing Facebook's data-sharing features.

The issue at play is Facebook's original application programming interface, or API, which allows third-party developers to use Facebook's platform and access some user data as long as those users give permission.

Facebook uses a different API now. But before 2014, Graph API v1.0, as it was called, allowed those developers to collect a stunning amount of information about you — and your friends.

The Cambridge Analytica scandal all started with a simple personality quiz.

A researcher named Aleksandr Kogan created an online personality quiz called "thisisyourdigitallife," which he paid about 270,000 people to take using Amazon's crowdsourcing platform, Mechanical Turk. Kogan forwarded the data he collected to Cambridge Analytica.

Handing over that data to Cambridge Analytica was against Facebook's policy. Facebook says Kogan violated his agreement to use the data for

academic purposes only, according to The Guardian.

But there was a larger issue at play: The quiz had also pulled data from the profiles of the 270,000 participants' friends, resulting in a trove of data from millions of users — as many as 50 million.

That's thanks to the first version of Facebook's API.

Consent from you — but not your friends

Jonathan Albright, the research director at the Tow Center for Digital Journalism at Columbia University, explained the user-privacy issues with Graph API v1.0 in a Medium post on Tuesday.

The main problem with the first version of the API, Albright said, was something called "extended profile properties." Those profile permissions allowed the apps to access not only *your* info — like gender, location, and

birthday — but your friends' information too, even if they weren't using the third-party app and hadn't authorized that access.

Essentially, the API asked for consent from you, but not from your friends.

Albright pointed to research from the Belgian university KU Leuven explaining what data was included in those extended profile properties — that is, exactly which data developers could access from your Facebook friends once you gave the app permission.

Here's the full list:

- **About me**
- **Actions**
- **Activities**

- **Birthday**
- **Check-ins**
- **Education history**
- **Events**
- **Games activity**
- **Groups**
- **Hometown**
- **Interests**
- **Likes**
- **Location**
- **Notes**
- **Online presence**
- **Photo and video tags**
- **Photos**
- **Questions**

- **Relationship details**
- **Relationships**
- **Religion**
- **Politics**
- **Status**
- **Subscriptions**
- **Website**
- **Work history**

That information, which your Facebook friends did not explicitly give a third-party app permission to use, was readily available anytime you hit "allow."

'Putting people first'

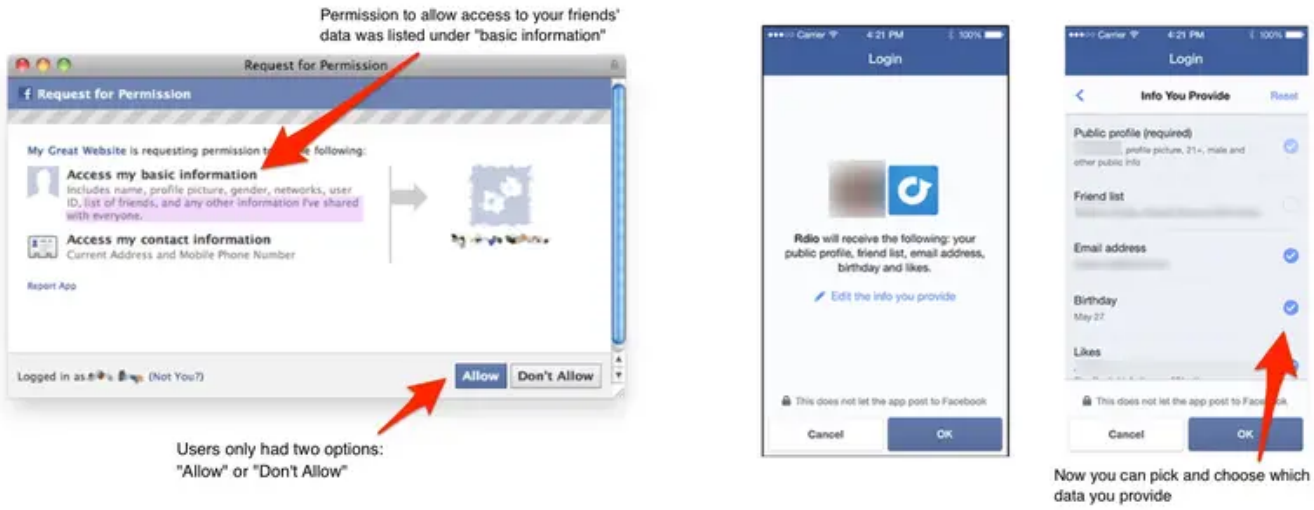
Graph API v1.0 launched in 2010 and lasted until 2014. It was shuttered and replaced with v2.0, still used today.

Even back then, Facebook users were concerned about how much information third-party apps had access to, and they voiced those concerns to Facebook. The company mentioned that when it announced the change to the API, describing it as "putting people first."

"We've heard from people that they are worried about sharing information with apps, and they want more control over their data," a Facebook press release said. "We are giving people more control over these experiences so they can be confident pressing the blue button."

The algorithm change allowed users to see what exactly they were permitting. Nathaniel Fruchter, Michael Specter, and Ben Yuan, researchers at MIT's Internet Policy Research Initiative, published a side-by-side comparison of the change.

V1 is on the left, and V2 is on the right:



Before the change, users couldn't selectively approve or deny specific permissions — they could hit "allow" or "don't allow."

To make matters worse, users weren't alerted to the extra permissions these apps were asking for, like those extended profile properties giving away their friends' information. As MIT points out, that was categorized as "basic information," which is most likely why so many people would grant permission.

'A chilling effect against free speech'

As of right now, there's no way to know for certain which information Cambridge Analytica had access to and eventually used. That's what various audits of Cambridge Analytica could reveal.

But the list of the types of data it *could* have taken includes 25 items.

The Guardian described Facebook's response to the breach of policy as "downplayed." A Facebook spokeswoman named Christine Chen told the publication that the data was "literally numbers" and that there was "no personally identifiable information included in this data."

Still, this much data in the hands of a third party could have far-reaching effects.

Fruchter, Specter, and Yuan summed it up best, saying that even though some of the data was public, what could be inferred from the full data set could be "very sensitive."

"If, for instance, this information were leaked, users' private data, such as political or religious affiliation, may quickly become public," they wrote. "Such a disclosure could cause people to lose jobs and insurance, and also cause a chilling effect against free speech and association."