Facebook data harvesting—what you need to know

April 4 2018, by Gráinne Maedhbh Nic Lochlainn

Credit: Tracy Le Blanc from Pexels

Facebook makes most of its money from advertising, and – as the Cambridge Analytica scandal continues to haunt Mark Zuckerberg's company – users are demanding to know how their data is being wrangled and harvested.

But while concern about Facebook user privacy has spiked, it's been clear since Facebook's inception that its business is based on widespread surveillance of people, whose data is the product.

Some have portrayed the revelations of the Cambridge Analytica scandal – in which data was allegedly harvested from 50m Facebook profiles – as an "existential crisis", while others have highlighted potential implications for academic research.

In short, Facebook's data harvesting methods have become a subject of sudden and widespread concern.

What is data harvesting?

Harvesting data, as its agricultural name suggests, is similar to gathering crops because it involves collection and storage with the expectation of future reward.

Data can be harvested in different ways, ranging from simple copy-and-pasting to more complicated programming. The chosen method is often constrained by the site being harvested. At simple search levels, many sites combat automated harvesting with Google CAPTCHAs and reCAPTCHAs, which help sites differentiate between humans and bots.

If you've ever copy-and-pasted text from Facebook or saved an image from Twitter, you've harvested social media data. The action of "screenshotting" is permitted on most sites because users can usually only access information that is either public or visible to them because they have logged in. Also, it would be impossible to completely eradicate the simplest data harvesting methods, such as making notes and taking photographs.

Facebook and other social networks are more concerned with restricting automated data harvesting, due to demands on web servers and to control who has access to what data (and why). Personal information and behaviour on social media have commercial, political and research value.

Social networks decide their own usage policies, balancing commercial interests with third parties and regulatory user privacy concerns – often described in company documents as juggling the optimisation of "customer behaviour" and adhering to "community standards".

How is data harvested?

Application Programming Interfaces (APIs) are used by Facebook, Twitter, Instagram and other sites to restrict would-be harvesters' access. APIs work as a software go-between that allows a researcher or app developer's computer to "talk" to a social network in a controlled way.

Read more: How Cambridge Analytica's Facebook targeting model really worked – according to the person who built it

One of the main conditions involves restrictions on how collected data can be used and shared, which can be pursued aggressively. In 2010, computer programmer Pete Warden harvested data from 210m public Facebook profiles for research purposes. But he failed to seek permission from Facebook first, thereby violating its terms of service. He later faced the threat of legal action from Facebook and was forced to delete the data – in an echo of academic researcher Aleksandr Kogan's alleged part in the Cambridge Analytica scandal.

Kogan's app, dubbed "thisisyourdigitallife", developed in 2014 through his company Global Science Research (GSR) – separate from his university work – was a personality test that 270,000 users logged into, accepting that it would have access to some of their personal information and some of their friends' data too. It also meant that those friends had not consented to their data being used in this way.

Facebook routinely updates its API and in 2014 the company confirmed it would stop allowing third-party apps to have access to data on the friends of app users. This disabled the data collection method allegedly used by Kogan.

There are a few different ways developers – who are required to agree to Facebook's policies – can harvest data using the company's API and they all assume at least basic computer programming skills. One of the easiest ways to do this is to access the API using a specialist software toolbox – Python and R have tools designed specifically for this purpose. In my research, I use the Rfacebook package to harvest Facebook data.

A key distinction between my app and others is that I'm not interacting with users, because my app isn't live. My app is essentially an automated way to copy-and-paste information from public Facebook groups. I use the Facebook API to research how public community group pages have been used to protest austerity in Ireland.

Because I'm harvesting public data from public pages, I'm not asking users to login and there's no front-end interface on Facebook, although this can be done using Facebook's API toolkits to expand the amount of data that can be accessed. It's a method that raises a number of questions about functionality, user information and access permissions.

Facebook's API can be used to harvest all sorts of publicly available information, like some of The Conversation UK's recent posts or posts in public groups.

But attempts to move beyond public information to harvest data of Facebook users who haven't logged in to the app – such as Zuckerberg, for example – return errors. Facebook "likes" can't be harvested because Zuckerberg isn't a user of my app and he hasn't granted it permission to access his data.

Under Facebook's latest API updates, app permissions are required to harvest any information beyond public profile properties. This means that users have to login to an app and authorise access to any other information to allow developers to harvest the data.

Legimate research under threat?

While ad-stuffed companies clearly have an interest in "leveraging" data, academics – in recent weeks – have drawn attention to researchers who harvest Facebook data. The practice has become relatively mainstream in social sciences research.

The extent to which future research could be restricted by changes to Facebook's API is a pressing one. But it's worth noting that, once data has been harvested, Facebook – which can legally pursue people who "violate" its terms of service to try to force them to delete data – has limited control over where data that ends up.

For researchers who are fretting about how the Cambridge Analytica scandal will affect their work, it's worth keeping an eye on what changes Facebook implements in its next API update. It may provide a better understanding of the type of research that can be permitted from the use of harvested Facebook data – and what may be permanently excluded.

This article was originally published on The Conversation. Read the original article.

Provided by The Conversation

Citation: Facebook data harvesting—what you need to know (2018, April 4) retrieved 28 April 2024 from https://phys.org/news/2018-04-facebook-harvestingwhat.html