Original Article

SM+S
social media + society

Social Media + Society
July-September 2020: 1–11

DOI: 10.1177/2056305120940703
journals.sagepub.com/home/sms
SAGE

What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data

Moreno Mancosu[1] and Federico Vegetti[1,2]

Abstract

In reaction to the Cambridge Analytica scandal, Facebook has restricted the access to its Application Programming Interface (API). This new policy has damaged the possibility for independent researchers to study relevant topics in political and social behavior. Yet, much of the public information that the researchers may be interested in is still available on Facebook, and can be still systematically collected through web scraping techniques. The goal of this article is twofold. First, we discuss some ethical and legal issues that researchers should consider as they plan their collection and possible publication of Facebook data. In particular, we discuss what kind of information can be ethically gathered about the users (public information), how published data should look like to comply with privacy regulations (like the GDPR), and what consequences violating Facebook's terms of service may entail for the researcher. Second, we present a scraping routine for public Facebook posts, and discuss some technical adjustments that can be performed for the data to be ethically and legally acceptable. The code employs screen scraping to collect the list of reactions to a Facebook public post, and performs a one-way cryptographic hash function on the users' identifiers to pseudonymize their personal information, while still keeping them traceable within the data. This article contributes to the debate around freedom of internet research and the ethical concerns that might arise by scraping data from the social web.

Keywords

web scraping, social networks, research ethics, Facebook

Introduction

In early 2018, in the aftermath of the Cambridge Analytica (CA) scandal, Facebook drastically tightened the access to its Application Programming Interface (API). For almost 10 years, the API served as the main tool by which researchers collected behavioral and digital trace data from Facebook (for some recent examples, see Abdulla et al., 2018; Braun & Schwarzbözl, 2018; Larsson, 2016; Poell et al., 2016; Stier et al., 2017). By using the API, independent researchers and third parties were able to easily download public information about users' profiles, as well as comments and reactions to public posts, to study the impact of social media on society. Because the API represented the *only* means by which Facebook authorized third parties to collect data from their platform, its lockdown effectively cut off any possibility for independent researchers to conduct observational research on relevant topics in political and social behavior, such as the structure of information networks, the spread of real and fake news, and the dynamics of political engagement. This raised a general concern among scholars, and sparked a debate around the (potential) alternative ways to access data that are crucial to carry on social research on Facebook (Bruns, 2018; Freelon, 2018; Venturini & Rogers, 2019; Walker et al., 2019).

As a solution to this state of affairs, a group of American scholars, in partnership with Facebook itself and other non-profit organizations, founded a new entity called Social Science One, whose role is to collect and evaluate research proposals and possibly grant the access to Facebook data directly from the company (King & Persily, 2018). Although the proposal seems promising, it presents the main issue of leaving the last word on what can be researched to the company itself. Successful proposals should focus on questions

[1]University of Turin, Italy
[2]University of Milan, Italy

Corresponding Author:
Moreno Mancosu, University of Turin, Lungo Dora Siena, 100, Torino, 10153, Italy.
Email: moreno.mancosu@unito.it

that "provide valuable knowledge to inform product, programmatic, and policy decisions," or at, the very least, that are "orthogonal to company interests" (King & Persily, 2018, p. 12). According to Bruns (2019),

> this approach to finding research questions would presumably rule out research that seeks to address key current issues such as abuse, hate speech, or disinformation (which are certainly not "orthogonal" to these platforms' interests), if there was any likelihood that the research might find the platforms' own affordances or policies culpable in facilitating such phenomena. (p. 9)

In other words, this model of data access has the potential to limit substantially the scope of what can be found by researchers.

Scholars have also started discussing possible alternative methods to obtain Facebook data, in what some scholars have already called "the post-API era" (Freelon, 2018; see also Bruns, 2019). The set of techniques used by researchers and practitioners to extract data from the internet, called "web scraping," ranges from manually downloading the data by copy/pasting information from web pages, to fully automated routines of data extraction. APIs can dramatically simplify this process for the user, but they are not the only means by which data can be accessed.[1] Facebook data formerly collected from the API are, after all, still publicly available and visible to the users who visit Facebook's web pages. Hence, they can be harvested using so-called "screen scraping" methods—namely, techniques that allow to automatically download and parse the content on display of internet pages to obtain a usable dataset (Freelon, 2018).

However, scraping Facebook data comes with three main requirements that researchers must deal with. First, scraping social media data is a way of collecting human subjects' data, hence it must comply with the ethical standards accepted by the scientific community, such as preserving users' privacy and avoiding any possible harm that the connection between one's online data and his or her physical person might enable (Markham et al., 2012). Second, and partly related to the first point, the scraped data must comply with the highest number of (virtually every) possible legal regulations that protect individuals' data. In particular, starting from May 2018, the European Union (EU) has issued a new set of regulations for individual researchers and firms, the General Data Protection Regulation (GDPR), one of the tightest laws on data protection currently in force, which provides a set of constraints to which scientific researchers using human subject needs to adapt. Third, a data scraping procedure is expected to comply with the terms of service (TOS) of the platform from which the data are being collected. As we shall see, it is extremely difficult, if not impossible, to adhere to this latter requirement for researchers interested in analyzing Facebook data independently. After reviewing the ethical and legal issues arising from collecting and publishing Facebook data, we propose a scraping routine that produces a dataset that is usable for research and satisfies the first two of the three requirements discussed. As for the TOS, we discuss the potential consequences that researchers might incur in case of their violation.

Ethical and Legal Hurdles of Facebook Research

> Social media bear the possibility for researchers to collect large amounts of high-quality observational data about human interactions and behaviors. However, great possibilities come with a great responsibility toward the human subjects that are being studied, that is, the responsibility to make sure that they are treated *ethically*. But what is an ethical treatment when our subjects are Facebook users interacting freely on the platform? Literature on ethics in social media research is vast and growing, ranging from analyses of ethical guidelines to perceptions of the subjects themselves (see Fiesler & Proferes, 2018; Metcalf & Crawford, 2016; Townsend & Wallace, 2016; Williams et al., 2017; Zimmer, Kinder-Kurlanda, 2017). In general terms, ethical research is guided by the principles of respect for persons, beneficence and justice (Buchanan & Zimmer, 2016). In the specific case of social media research, where users are observed in their natural environment, the most important concern is to make sure that subjects do not incur into any *harm* due to their inclusion in the research. In Townsend and Wallace's (2016) words,[t]his risk of harm is most likely where a social media user's privacy and anonymity have been breached, and is also greater when dealing with more sensitive data which when revealed to new audiences might expose a social media user to the risk of embarrassment, reputational damage, or prosecution (to name a few examples). (p. 7)

Besides ethical concerns, social media scholars also need to make sure that their research activity does not breach any legal barrier. This is not much of a concern for researchers using more "traditional" approaches, like surveys or lab experiments, as subjects in these studies are asked to provide their *informed consent* prior to the collection and use of their data. However, one of the strengths of using digital trace data, whether from social media or other platforms, is exactly that subjects are observed performing their online activities in a spontaneous manner. In addition, the typical social media research project involves collecting data about *many* individuals, often hundreds of thousands. This makes it practically impossible to obtain informed consent from every one of them. In the EU, the use of personal data is regulated by the GDPR, which establishes some clear conditions under which scientific researchers can collect, store, and publish data about human subjects, with or without an informed consent. Furthermore, access to social media platforms is usually regulated by TOS agreements, which often pose limits to the amount of data that can be extracted (like in the case of Twitter) or to the extraction methods that can be used (like in the case of Facebook). These legal aspects must be kept in

Table 1. Ethical and Legal Concerns of Research Using Facebook Data.

	Ethical concerns	Legal concerns
Data preparation stage	Private/public nature of the information	Privacy & Data Protection laws (e.g., GDPR), TOS
Data reporting stage	Subjects' anonymity, Sensitivity of the information, Group privacy	Privacy & Data Protection laws (e.g., GDPR)

GDPR: General Data Protection Regulation; TOS: terms of service.

mind when designing a study using Facebook or other social media data.

For researchers who wish to work with Facebook data, ethical and legal hurdles can occur at two stages of the research process. The data *preparation* stage is when the data are collected, stored, cleaned, and analyzed. From the ethical standpoint, the only way that the researcher can cause harm to the subjects at this step is by observing them in a private environment, where they do not expect to be observed by external witnesses, thus violating their personal space. This may include their own timeline or the timeline of their connections, closed groups, as well as private messages. Moreover, this is the stage where the researcher has to deal with all the barriers that the platform has put in place to avoid third parties to collect their data, starting with the TOS. The data *reporting* stage is when the data are published. This can occur in different ways and for different purposes. Quantitative researchers may report the data in aggregate form, making it very unlikely that individual subjects are personally harmed. However, more and more academic journals require researchers to make their data available to the public for reproduction purposes, potentially disclosing the subjects' personal and/or sensitive information in case it is not concealed. Furthermore, the researcher may wish to quote a user's post to provide qualitative evidence, making it possible for third parties to search the user directly on the platform. Hence, at this stage, it is very important that the researcher deploys all the necessary means to make sure that the potentially sensitive information about the subjects, as well as their identity, remains anonymous. Luckily, doing this properly should also ensure that data protection laws, such as the GDPR, are respected. In the rest of the present section, we discuss these points more in detail.

Table 1 represents ethical and legal concerns of research using Facebook data.

Facebook Research and Ethical Concerns

We discuss here two examples of problematic collection and use of Facebook data that illustrate the ethical issues that the researcher may encounter. The first, and arguably more (in) famous, is the case involving CA. CA's data collection technique was based on a tool called "This Is Your Digital Life," one of the numerous Facebook applications that provide users information about their own alleged psychological profile based on their social media activity (Hern, 2018). Next to collecting psychological test data freely provided by the app users, the application was programmed to harvest the personal information they shared on their own profile, their private messages, as well as their Facebook friends' lists (Frenkel et al., 2018). Such lists were in turn fed into an algorithm which visited all their Facebook pages, and harvested all the public information that it could find (like, for instance, which Facebook pages they "liked"). Starting with about 270,000 app users, CA was eventually able to collect information of about 87 million users (Kang & Frenkel, 2018). This information was used to profile users based on their psychological characteristics, cultural tastes, as well as political and religious views, with the aim of fine-targeting them in political advertising campaigns.

The CA case is extremely problematic from several points of view. First, the information collected by the company about the initial 270,000 app users was not openly available to any Facebook visitor, but it was semi-private information shared by the users with their own closed circle of "friends." In other words, a user who was not friend of *all* the users who installed the app would not have been able to view all the data in the dataset. Hence, CA's data collection violated the users' privacy, observing them in a space where they would not "reasonably expect to be observed by strangers" (Townsend & Wallace, 2016, p. 10). Second, while the data collected by CA was never made public, it was used to build user profiles to inform micro-targeted political campaigns. Although individual users could not be personally identified, and thus were not exposed to potential harmful consequences *individually*, their sensitive information (cultural preferences, political views) was used to sort them into groups, which in turn were exposed to persuasion campaigns *collectively*. In this case, even though the individual privacy was preserved, their *group privacy* was violated (see Kammourieh et al., 2017).[2] So in sum, the case of CA illustrates an example of research that is ethically problematic both in the data preparation stage and in the data reporting stage.

Another informative example of ethical issues in Facebook research is the T3 dataset ("Tastes, Ties, and Time"), collected by a Harvard University team (see Lewis et al., 2008). The researchers publicly released profile data collected from the Facebook accounts of a cohort of college students from a US university, but although attempts have

Table 2. Privacy Issues with Facebook Data.

	Data are private	Data are public
Harms the users	Cambridge Analytica	T3 dataset
Does not harm the users	N.A.	?

been made to hide the identity of the institution and the students involved in the study, both the users and the university were rapidly identified, undermining the anonymity of the data (Kaufman, 2008). In a subsequent publication, Zimmer (2010) pointed out that

> the research team [argued] that their data collection methods were unproblematic since the "information was already on Facebook," [but] just because personal information is made available in some fashion on a social network, does not mean it is fair game for capture and release to all.

In other words, although these data *are* public in a sparse form on the Internet, the fact that the researcher is gathering them together and publishing them in a dataset may lead to possible identification of the users, with all the potential harmful consequences that this entails. Hence, the researcher should take all measures to ensure that public data collected from Facebook are treated in a way that all information potentially leading to recognizing or tracking the users is removed—at least as far as the dataset produced will be made publicly available, an increasingly common requirement in academic research. This requirement raises a number of technical questions when it comes with the need to produce analyzable data, as we shall discuss below.

These two examples can be organized in a taxonomy that clarifies the ethical constraints that researchers should expect when collecting and distributing data collected from Facebook. Table 2 summarizes the possibilities of data collection on Facebook discussed so far, based on the following two dimensions: the *public/private* nature of the information and the *sensitivity* of the information. The first dimension refers to the degree to which users can "reasonably expect to be observed by strangers" (Townsend & Wallace, 2016) in the context where they produce the information that the researcher wishes to collect. Information can be password protected (like in private messages) or stored in private Facebook groups (with different degrees of gatekeeping, such as the need for approval by an admin to enter the group, etc.). Furthermore, information can be accessible only to the Facebook "friends" of the user, to "friends of friends," and so on. Whereas, the distinction between public and private information is rather nuanced on Facebook, as a rule of thumb to determine what can be treated as really *public*, the researcher may ask the question "*would a Facebook user without friends be able to see this?*" In this case, elements like posts on public Facebook pages, together with the comments, likes, and reactions to such posts, pertain completely to the realm of the public debate on Facebook.

However, users' privacy settings can mislead the academic researcher. For instance, one could argue that a user's friends' list is not private information if the user decides to keep it public (or fails to set it up correctly in the privacy settings), and therefore, it should be rightfully collectable and distributable in a dataset form. However, as the T3 case suggests, friends' lists (as well as other personal information, see Zimmer, 2010) could potentially lead to the re-identification of the users, with the potential harm that this brings (e.g., reputational damage or off-line prosecution).[3]

While the data collection of CA can be located in the top-left cell (information which is formally private and harms the users), the T3 dataset situation is more nuanced, as the resulting dataset might harm the participants, although the data are formally public (top-right cell). Granted that in this domain there is virtually no information which is private and whose collection does *not* harm the subjects (bottom-left cell), the researcher's efforts should be devoted to produce and distribute datasets that collect only public information that is anonymized and otherwise treated in a way that its use cannot harm the users (bottom-right cell).

Facebook Research and Legal Concerns

Collecting and publishing social media data, if not done properly, might lead to two orders of legal concerns. The first is that publishing the data might lead to copyright infringement (see, for example, Markham et al., 2012). However, discussing this point is beyond the scope of this review. The second concern, at least in the EU, is the need to respect the GDPR. In general, the GDPR poses a number of conditions under which data "processing" may be considered *lawful*.[4] For the purposes of academic research, the easiest way to process the data in a way that complies with the regulation is by obtaining the informed consent of the subjects (Art. 6(1)(a)). However, the regulation also provides that data processing may be lawful if it is "necessary for the purposes of the legitimate interests pursued by the controller or by a third party" (Art. 6(1)(f)). The article further specifies that "legitimate interest" means that the data processing is "necessary for the performance of a task carried out in the public interest." Hence, since it is generally straightforward to defend academic research as pertaining the public interest, data collection, analysis, and publication for scientific purposes should be protected by the GDPR.

Furthermore, the regulation requires that particular care be put in case the researcher deals with sensitive data. In general, the GDPR forbids the processing of

> personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation. (Art. 9(1))

These include many hot topics in social science research. However, another paragraph in the same article specifies that the rule does not apply for processing performed for scientific research, assuming that the researcher "respect[s] the essence of the right to data protection and provide[s] for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject" (Art. 9(1)(j)). This derogation is further specified in Art. 89, where two examples of effective "measures" are made: "pseudonymization" and "further processing which does not permit or no longer permits the identification of data subjects" (Art. 98(1)). In our understanding, the second example refers to reporting data in an aggregate manner, with summary statistics. What we find more important to the purposes of collecting, analyzing, and reporting data about public activity of Facebook users, is the reference to *pseudonymization*.

Pseudonymization is a de-identification technique that substitutes all the identifying information of a single individual into pseudonyms, making it impossible to directly identifying the user, but maintaining the possibility to analyze the data and to track the users in their activities among different publicly visible pages. For instance, the pseudonymization of the personal information about the user "John Smith," who performed a certain activity on a public Facebook page and whose Facebook address is "https://www.fb.com/john.smith.2092," would pass through a one-way encryption algorithm that transforms the URL of the user into a anonymous hash that, however, remains the same in all instances the same user appears in the data. In general, pseudonymization may represent a winning compromise between the need for social researchers to obtain social media data and the need to respect the legally sanctioned right to privacy of the users, which includes their protection from the possible negative consequences that may come from exposing sensitive information about them. In fact, such a technique for data de-identification is often used in some forms by researchers dealing with very sensitive information, like genetic or health data, to protect the subjects anonymity and at the same time be able to link together different data sources about the same individuals (see, for example, Aamot et al., 2013; Elger et al., 2010; Erlich & Narayanan, 2014).

Facebook Research and TOS Compliance

A central legal concern is represented by the compliance (or lack thereof) of scraping activities with TOS of the platform that the researcher wishes to scrape. Freelon (2018) even mentions this as one of the two challenges that computational researchers need to face in the post-API age (the other one being learning how to implement scraping techniques). TOSs are contractual restrictions that can be used by social media platforms to forbid companies and individuals from scraping information from their sites. By accessing a platform users typically must accept the TOS, and by accepting the TOS users are bound in the activities they can perform on the platform—researchers are actually *users* like everyone else. Facebook, in the Automated Data Collection terms, defines automated data collection as "the collection of data from Facebook through automated means, such as through harvesting bots, robots, spiders, or scrapers" and clearly states that people cannot "engage in Automated Data Collection without Facebook's express written permission" (Facebook, 2010). What does this imply in practice for scholars who wish to scrape public Facebook data for research purposes?

As Halavais (2019) maintains, there are three ways by which TOSs can be enforced to restrict automated data collection: they may be enforced directly by the state, they may be enforced by universities' institutional ethic panels, and they may be embodied in the technical infrastructure of the platforms themselves. State enforcement of TOS is arguably the most concerning risk for researchers, hence we limit our discussion to the first point. As Freelon (2018) points out, "even the remote prospect of criminal prosecution for violating TOS creates a chilling effect strong enough to deter most researchers" (p. 667). In the United States, criminal prosecution is possible based on the Computer Fraud and Abuse Act (CFAA), a bill enacted in 1986 to handle computer crimes. According to the CFAA, to access a computer "without authorization" or in a way that "exceeds authorized access"[5] constitutes a crime which can be harshly punished. An episode that comes to mind is the case of Aaron Swartz, who was arrested in 2011 under the accusation of violating the CFAA for downloading automatically a large number of academic articles from JSTOR. If convicted, Swartz could have been charged with 35 years of prison and a US$1 million fine. He eventually committed suicide before trial in 2013.

While fear of prosecution is a powerful deterrent, increasing evidence suggests that violating the TOS to scrape public information from social media platforms might be *de facto* safe for researchers. As some have pointed out (Bruns, 2018, 2019; Halavais, 2019), research on social media serves the public interest, and threatening to apply the CFAA to prevent academics and journalists from collecting public information from Facebook for research purposes constitutes a violation of the First Amendment (Knight First Amendment Institute

at Columbia University, 2018; Sandvig, 2017) and of the human right to free research (United Nations, 1976). This might sound as a purely theoretical point, however, there is evidence that this line of arguing informs legal decisions as well. A prominent example is the *Sandvig v. Sessions* case, where a court in Washington, D.C. ruled that scraping publicly available information is not a computer crime even when the TOS explicitly forbids it (Williams, 2018). The key point is the public nature of the information that researchers wish to scrape. As the court stated,

> [s]craping is merely a technological advance that makes information collection easier; it is not meaningfully different from using a tape recorder instead of taking written notes, or using the panorama function on a smartphone instead of taking a series of photos from different positions. (Sandvig v. Sessions, 2018, p. 15)

More recently, in the *HiQ Labs v. LinkedIn* case, the Ninth Circuit Court of Appeals ruled that scraping publicly available data does not constitute "unauthorized" access to a computer, even when the owner (in our case, the owner of the servers where the data are stored) has sent a cease-and-desist letter to the visitor of the website. Even here, the argument is that when some information is available to the public, scraping it does not violate the CFAA. This is surely the case for posts from public Facebook pages.[6]

While the CFAA and its applicability in case of violation of the TOS is relevant for US-based researchers and journalists, states can only enforce laws within their borders, hence, the risk of criminal prosecution for violating the TOS is substantially smaller for researchers based in other parts of the globe. Social media platforms can still claim that the user performing the scraping has breached a contract, and possibly claim for damages. However, the latter option is generally not viable if the user did not cause any demonstrable loss to the company, hence the biggest risk for researchers remains that the company closes their account (see Beurskens, 2013, who also points out that some information, like in our case public Facebook posts, does not require a user account to be accessed, hence users are not required to subscribe to any TOS document in order to access it). Moreover, even with respect to contract violation, the magnitude of the risk is all but clear. Recent studies have been pointing out that TOS documents of many online platforms or services contain several *unfair* or *potentially unfair* terms (see Lippi et al., 2019; Loos & Luzak, 2016; Micklitz et al., 2017), making them hard to enforce in practice.[7]

Finally, there is evidence of software collecting Facebook data automatically which still works undisturbed after the closing of the API. Three examples are *NCapture*, a browser extension that allows the user to download Facebook posts and comments to be analyzed with the software *NVivo, Power BI Desktop*, an application by Microsoft that allows users to download information about posts from a public Facebook page (Microsoft, 2019), and *Facepager*, an open-source application to fetch public data from JSON-based APIs (Jünger & Keyling, 2019). However, we are not aware of the legal status that these packages hold with respect to Facebook: this makes it difficult to use such examples as an argument in favor of scraping *tout court*, especially when performed without an explicit permission. Nevertheless, we believe that the very presence of these cases suggests that automatically collecting information from public Facebook pages is generally tolerated. So in sum, while violating the TOS remains a crucial point of concern for researchers who wish to scrape Facebook public data, the chance to undergo substantial negative consequences appears to be rather limited, although the situation in this respect is still fluid.

Screen Scraping Data From Facebook

The Logic Behind Screen Scraping

Screen scraping is a technique based on automated browsing that allows to simulate a user's behavior and to collect the data visualized on the screen. At the present moment, given the demise of the API, screen scraping appears to be the only method that can technically allow researchers to systematically collect large amounts of data from Facebook. As an attempt to provide a tool that respects ethical and legal constraints as much as possible, and at the same time allows the researchers to obtain the data that she/he needs, we have developed a pseudocode with a routine for using screen scraping techniques to extract reactions from a single public post of a Facebook page (see Figure 1).

Automatic browsing is a technique that allows a user to program a web browser to simulate a user's actions, such as clicking on buttons, links, or writing text in a specific web page. Facebook is a website designed to have a very stable page structure: for instance, the HTML structure of a web page showing comments to a certain post will always be the same, with the noticeable exception of the data contained in it. It is, thus, quite easy to design a web browser session aimed at visualizing all the comments or reactions to a certain post. Once the information has been visualized as a whole in the browser, the routine parses the substantive information contained in it (in the code example, the reactions), and produces a dataset that is usable for the analysis.

The example routine downloads, for a single example post, the unique code that identifies the users reacting to it, and the reaction type. It first creates a Google Chrome profile with the credentials of an existing Facebook user (such as user name and password) pre-stored in it, then it instructs the browser to go to the target page, open the reaction sub-page and press automatically the "*See more reactions*" button until all the reactions are visualized. After that, the routine automatically downloads and parses the data on the page, providing a dataset containing information of those who have reacted and their reaction (whether they are "*like*," "*wow*,"

```
Get Chrome profile
Open the session and navigate to the chosen page (i.e. the reactions to a public post)
Store in var the array of HTML elements containing the reactions to the post organized by reaction type
Create an empty array called link_list
Set index to 1
While index is less than or equal to the length of var
        Extract HTML element in var[index]
        Extract reaction type in var[index]
        Repeat
                Find "See more" button
                if "See more" button is present
                        Click on "See more" button
                        Session timeout
                else
                        Break function
        Store in var the updated array of HTML elements containing the reactions to the post organized by reaction type
        Parse the vector of used IDs stored in var[index]
        Store the parsed vector into link_list[index]
        Update index as index + 1
Append all vectors in the array link_list in a table with 2 columns: user_id and reaction
Apply encryption to the column user_id
Print the table
```

Figure 1. Pseudocode for scraping reactions to a public Facebook post.

"*laugh*," and so on). This routine can be easily coded up to work on open-source software like *R*, by using external libraries designed for automatic browsing (such as *RSelenium*). An example code which implements the routine in *R* can be found in the Supplemental Appendix.

How Can Screen Scraping Keep Users' Privacy Safe?

This approach is a simple replacement of what was possible to do until 6 Februrary 2018 with other software utilities (such as *RFacebook*, see Barberà, 2017) through the Facebook API, or what is still doable with browser add-ons like *NCapture*. However, as discussed before, simply collecting the same data with a new method might result in a form of data processing that does not comply with recent privacy regulations, like the GDPR. For this reason, an ideal routine would automatically pseudonymize the data by operating a one-way encryption of the unique identifier of the users reacting to the post (such as, for instance, a MD5 hash conversion). This way, although it becomes much more difficult to find the real identity of a person acting on a post, the pseudonymized identifier of the same user is always the same across the data. Hence, it is possible to track the same people as they act on different pages, while at the same time concealing their real identity. This combines the opportunity to satisfy the research needs of social scientists studying behavior on Facebook with the need to keep users anonymous pointed out by ethical scholars and requested by European laws.

However, to pseudonymize user identifiers is not enough to make sure that subjects cannot be personally identified. By knowing the ID or the URL of the post to which the reactions are observed, it is possible to visit the page and compare the dataset containing the scraped data (which includes the pseudonymized user identifiers and the reaction type they expressed to the post) with the actual names of the users who reacted to the post. Hence, some additional measures are necessary to reduce the risk of user re-identification. Note that this measures could be taken at the data preparation stage, by incorporating them in the scraping routine, or at the data reporting stage, by editing the dataset containing the scraped information. Hence, we will focus this discussion on what the data should look like when it is published. First, published data should not include a list of users reacting to a post when the reactions of any type are too few. It is difficult to make a concrete case of what "too few" is, however, we feel that this number should be no lower than 20. After all, one of the reasons why researchers are increasingly turning to studying behavior on Facebook is the real-world significance of the phenomena occurring on the platform. While it might be interesting to study small communities acting on public pages, it is difficult to maintain that observing "*wow*" reactions to a post where there are less than 20 of them can help in revealing any pattern beyond random noise.[8] Second, as the screen scraping procedure arranges the observations in the dataset in the same order as they appear on the page, it is important that the order of the observations is randomized in the published data, so the first observation appearing as reacting to a post is not the same as the first Facebook user reacting to the post in the publicly visible page.[9]

However, there are potential other ways to track down users, even when their identifiers are pseudonymized and the small groups are removed. For instance, one could compare patterns of reactions to different posts to match a pseudonymized user in the data with an actual user across Facebook. Given the type of data that we are dealing with (information publicly available on the Internet), it may just not be feasible to produce a dataset that is completely re-identification-proof. However, this problem arises only in the data reporting stage. These data that are used for the analysis, if stored in a secured machine, with satisfactory security standards, may be as detailed as the researcher needs. However, publishing the data, unless this is done in a completely aggregate form, may be problematic. A radical solution would be to pseudonymize *all* the qualitative information in the data, including the post and page IDs, or the text of the comments (in case one wishes to collect those). Alternatively, the researcher may agree to disclose the data to specific third parties only upon a signed commitment of non-dissemination. An important future task for scholars interested in working with Facebook data would, therefore, be to find a way to maximize research transparency and subjects' privacy at the same time.

Limitations of Screen Scraping

Screen scraping procedures, despite being the only ones that can be used, *rebus sic stantibus*, to collect large amounts of data from Facebook, are still problematic in terms of TOS compliance. As discussed earlier, Facebook TOS (and in particular, the Automated Data Collection Terms, see Facebook, 2010) defines automated data collection as the "collection of data from Facebook through automated means, such as through harvesting bots, robots, spiders, or scrapers" and forbids anyone to "engage in Automated Data Collection without Facebook's express written permission." Hence, currently, performing screen scraping on Facebook constitutes a violation of the TOS, potentially making the researcher susceptible to different actions from Facebook, from "immediate ban" to "injunctive relief." Based on the considerations discussed earlier, and based on the fact that other software packages to collect Facebook data are fully operational, we do not believe that researchers collecting public Facebook data for scientific purposes are at risk of too negative repercussions. However, it is important to be aware of the potential consequences that may occur.

Aside from TOS-related concerns, there are some technical limitations that may affect the researcher's ability to use

screen scraping and the quality of the data collected. First, Facebook (as well as other social media platforms) may react to the presence of third parties wishing to scrape their data by adopting technical measures that can undermine the functioning of scraping tools. For instance, the platform could act on the HTML code of the pages by altering their structure or (more likely) by changing the HTML tags or classes of the elements that the software needs to identify to proceed with the scraping (such as the "*See more*" button or the list of reactions itself). Of course, scraping tools can be refined to adapt to the new pages, however, this needs to be done every time the platform changes the HTML of its pages, which may happen frequently if the platform is particularly determined to deter scrapers. This can bring the researcher into a never-ending loop of trial, error, and revision, which can be extremely time and resource consuming especially for less experienced scholars. Second, researchers using screen scraping to collect data from Facebook must be aware that what appears on their own screen (the scraper uses the researcher's own account to access Facebook) might not be the same as what other users view—not necessarily because some content is hidden to them, but because of the personalization of the content. As far as we are aware, this problem is not likely to affect what users can view on public pages, however this is a possibility that researchers should acknowledge as they report on the data collection and qualify their sample.

Concluding Remarks

According to Freelon (2018), the post-API era in digital communication studies has begun. After the CA scandal burst, Facebook decided to close the API, the sole legitimate tool that allowed third parties to access and download information about the user activities on the platform. Moreover, it seems that other platforms are going to follow this lead (see Roth & Johnson, 2018), making it more difficult, or even impossible, to collect behavioral data in a safe and straightforward way. The paradox, however, is that the information has not disappeared from the web, but instead it is still publicly available, and anyone who has a Facebook account is able to access the identity, activities, and even the personal/sensitive data of an unimaginable amount of users. We have argued that, although randomly surfing this information is acceptable, gathering the same information into a dataset and eventually publishing it might lead to ethical and legal concerns. We have discussed previous examples (the CA and the T3 data collections), stating that, to be sufficiently ethically and legally acceptable, a publishable dataset of Facebook data must not contain private information about the users, and must remove all the potential identifying information (by pseudonymizing their user identifiers). Being the API currently unavailable for scientific purposes, this article proposes, borrowing the idea from Freelon (2018), a screen scraping approach to data collection and an example code for it in *R* (see Supplemental Appendix). We have discussed that, while such a procedure can be reasonably defined as both ethically and legally (GDPR) acceptable, it can still expose the researcher to legal risks due to the violation of Facebook TOS. At the moment, the situation being very fluid, it is difficult to clearly evaluate the impact that using such a technique might have legally.

Although sub-optimal with respect to the API approach, the screen scraping procedure proposed here might be easily expanded by developing, for instance, complete packages that would replace old libraries like *RFacebook* (Barberà, 2017) or the *Facebook Python SDK* (Dodoo, 2018). Although this approach does not seem to be particularly dangerous (software like *NCapture*, heavily based on screen scraping, have worked without issues before and after the closing the API), Facebook's unpredictable reactions, together with the fact that using such packages could lead the researcher to be charged with TOS violation, prompt us to suggest caution to scientific researchers who wish to engage with this type of data collection.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Moreno Mancosu https://orcid.org/0000-0002-3017-4066

Supplemental Material

Supplemental material for this article is available online.

Notes

1. It should also be noted that data collected through the Application Programming Interfaces (APIs) are likely to be biased by the standardization procedures operated by the platforms, and so, they are far from the "raw" observational data that researchers may collect through other means (see Puschmann, 2019; Puschmann & Burgess, 2013; Venturini & Rogers, 2019).
2. It must be noted that the extent to which a violation of group privacy is actually harmful for the subjects is not always straightforward. However, as with individual privacy, the lack of *actual* harmful consequences does not rule out the presence of *potential* harmful consequences. For instance, as Kammourieh et al. (2017) maintain,

   > [i]n some countries, group privacy violations mainly result in unwanted targeted ads and other inconveniences in customer experience. While these violations can and should warrant attention, the consequences and effects of group privacy violations for vulnerable groups, particularly those in fragile

contexts and/or areas of limited statehood, can be potentially life-threatening. (p. 48)

3. Of course, the degree of sensitivity of the information is even harder to pin down based on objective criteria than its public/private nature, and this evaluation is often left to the researcher. However, as we will see in the following section, there are criteria that can help the researcher minimize the risk to harm the users when collecting their social media data.
4. According to the definition in Art. 4(2),

   "processing" means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction

   —in other words, every contact to the data at any stage of the research.
5. Exceeding the authorized access means "to access a computer with authorization and to use such access to obtain or alter information in the computer that the accessor is not entitled so to obtain or alter" (see Jarrett & Bailie, 2015, p. 5).
6. The court still holds that scraping a website may violate the common law tort of trespass to chattels, or may be subject of civil causes for "copyright infringement, misappropriation, unjust enrichment, conversion, breach of contract, or breach of privacy" (HiQ Labs v. LinkedIn, 2019, pp. 34–35).
7. Note that the cited studies focus on the European Union (EU) law, specifically the Unfair Contract Terms Directive (UCTD).
8. To minimize the data that are eliminated, reactions can be grouped in some way, like for example, "positive" and "negative" reactions.
9. This measure can be more easily incorporated in the scraping routine, see the *R* code in the Supplemental Appendix.

References

Aamot, H., Kohl, C. D., Richter, D., & Knaup-Gregori, P. (2013). Pseudonymization of patient identifiers for translational research. *BMC Medical Informatics and Decision Making*, *13*, Article 75.

Abdulla, R., Poell, T., Rieder, B., Woltering, R., & Zack, L. (2018). Facebook polls as proto-democratic instruments in the Egyptian revolution: The "We Are All Khaled Said" Facebook page. *Global Media and Communication*, *14*(1), 141–160.

Barberà, P. (2017). Package "RFacebook." https://tinyurl.com/y9fnv9kx

Beurskens, M. (2013). Legal questions of Twitter research. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and society* (pp. 123–133). Peter Lang.

Braun, D., & Schwarzbözl, T. (2018). Put in the spotlight or largely ignored? Emphasis on the Spitzenkandidaten by political parties in their online campaigns for European elections. *Journal of European Public Policy*, *26*(3), 428–445.

Bruns, A. (2018). Facebook shuts the gate after the horse has bolted, and hurts real research in the process. *Internet Policy Review*. https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786

Bruns, A. (2019). After the "APIcalypse": Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566.

Buchanan, E., & Zimmer, M. (2016). Internet research ethics. http://plato.stanford.edu/entries/ethics-internet-research/

Dodoo, M. (2018). Facebook SDK for python. https://tinyurl.com/jepyg4c

Elger, B. S., Iavindrasana, J., Lo Iacono, L., Müller, H., Roduit, N., Summers, P., & Wright, J. (2010). Strategies for health data exchange for secondary, cross-institutional clinical research. *Computer Methods and Programs in Biomedicine*, *99*(3), 230–251.

Erlich, Y., & Narayanan, A. (2014). Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics*, *15*(6), 409–421.

Facebook. (2010). Automated data collection terms. https://tinyurl.com/cudjd2l

Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media + Society*, *4*(1), 2056305118763366.

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668.

Frenkel, S., Rosenberg, M., & Confessore, N. (2018, April 11). Facebook data collected by quiz app included private messages. *The New York Times*. https://nyti.ms/2Hpdqkx

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society*, *22*(11), 1567–1581.

Hern, A. (2018). How to check whether Facebook shared your data with Cambridge Analytica. *The Guardian*. https://tinyurl.com/ybvttn35

HiQ Labs v. LinkedIn. HiQ Labs v. LinkedIn Opinion., No.17–16783 (Ninth Circuit Court of Appeals September 9, 2019).

Jarrett, H. M., & Bailie, M. W. (2015). Prosecuting computer crimes. https://tinyurl.com/y88bpgl6

Jünger, J., & Keyling, T. (2019). Facepager. An application for generic data retrieval through APIs. *Source code and releases*. https://github.com/strohne/Facepager/

Kammourieh, L., Baar, T., Berens, J., Letouzé, E., Manske, J., Palmer, J., Sangokoya, D., & Vinck, P. (2017). Group privacy in the age of big data. In L. Taylor, L. Floridi, & B. van der Sloot (Eds.), *Group privacy: New challenges of data technologies* (pp. 37–66). Springer.

Kang, C., & Frenkel, S. (2018). Facebook says Cambridge Analytica harvested data of up to 87 million users. *The New York Times*. https://tinyurl.com/y85kfv2z

Kaufman, J. (2008). Michael—We did not consult . . . . *michaelzimmer.org*. https://tinyurl.com/y8x3vbz4

King, G., & Persily, N. (2018). A new model for industry-academic partnerships [Working paper]. http://j.mp/2q1IQpH

Knight First Amendment Institute at Columbia University. (2018). Knight institute calls on Facebook to lift restrictions on digital journalism and research. https://tinyurl.com/s9bz624

Larsson, A. O. (2016). Online, all the time? A quantitative assessment of the permanent campaign on Facebook. *New Media & Society*, *18*(2), 274–292.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, *30*(4), 330–342.

Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., & Torroni, P. (2019). CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, *27*(2), 117–139.

Loos, M., & Luzak, J. (2016). Wanted: A bigger stick. On unfair terms in consumer contracts with online service providers. *Journal of Consumer Policy*, *39*(1), 63–90.

Markham, A., & Buchanan, E. AoIR Ethics Working Committee. (2012). *Ethical decision-making and Internet research: Version 2.0.* Association of Internet Researchers.

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society*, *3*(1).

Micklitz, H.-W., Palka, P., & Panagis, Y. (2017). The empire strikes back: Digital control of unfair terms of online services. *Journal of Consumer Policy*, *40*(3), 367–388.

Microsoft. (2019). Tutorial: Facebook analytics using Power BI Desktop. https://tinyurl.com/umckm9r

Poell, T., Abdulla, R., Rieder, B., Woltering, R., & Zack, L. (2016). Protest leadership in the age of social media. *Information, Communication & Society*, *19*(7), 994–1014.

Puschmann, C. (2019). An end to the wild west of social media research: A response to Axel Bruns. *Information, Communication & Society*, *22*(11), 1582–1589.

Puschmann, C., & Burgess, J. (2013). The politics of Twitter data. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and society* (pp. 43–54). Peter Lang.

Roth, Y., & Johnson, R. (2018). New developer requirements to protect our platform. *Yoel Roth blog post*. https://tinyurl.com/y74uuqsm

Sandvig, C. (2017). Heading to the courthouse for Sandvig v. Sessions. *"Social Media Collective" blog post*. https://tinyurl.com/yacd63t2

Sandvig v. Sessions. Sandvig v. Sessions Memorandum Opinion, No.16-1368 (JDB) (District of Columbia March 30, 2018).

Stier, S., Posch, L., Bleier, A., & Strohmaier, M. (2017). When populists become popular: Comparing Facebook use by the right-wing movement Pegida and German political parties. *Information, Communication & Society*, *20*(9), 1365–1388.

Townsend, L., & Wallace, C. (2016). *Social media research: A guide to ethics*. University of Aberdeen.

United Nations. (1976). International covenant on economic, social and cultural rights. https://tinyurl.com/yxaym3zj

Venturini, T., & Rogers, R. (2019). "API-based research" or how can digital sociology and journalism studies learn from the Facebook and Cambridge Analytica data breach. *Digital Journalism*, *7*(4), 532–540.

Walker, S., Mercea, D., & Bastos, M. (2019). The disinformation landscape and the lockdown of social platforms. *Information, Communication & Society*, *22*(11), 1531–1543.

Williams, J. (2018). D.C. court: Accessing public information is not a computer crime. *Electronic Frontier Foundation*. https://www.eff.org/deeplinks/2018/04/dc-court-accessing-public-information-not-computer-crime

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, *51*, 1149–1168.

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313–325.

Zimmer, M., & Kinder-Kurlanda, K. (2017). *Internet research ethics for the social age: New challenges, cases, and contexts*. Peter Lang Publishing, Incorporated.

Author Biographies

Moreno Mancosu is assistant professor at the University of Turin (Italy). His research interests are mainly in electoral behavior, political communication, and quantitative methods. His articles appeared in Political Psychology, Political Communication, Communication research, and European Journal of Political Research.

Federico Vegetti is a postdoc research fellow at the University of Turin (Italy). His research interests include political perceptions and behavior, as well as social media and communication.