Information, Communication & Society

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/rics20

User-centric approaches for collecting Facebook data in the 'post-API age': experiences from two studies and recommendations for future research

Johannes Breuer, Zoltán Kmetty, Mario Haim & Sebastian Stier

To cite this article: Johannes Breuer, Zoltán Kmetty, Mario Haim & Sebastian Stier (2023) User-centric approaches for collecting Facebook data in the 'post-API age': experiences from two studies and recommendations for future research, Information, Communication & Society, 26:14, 2649-2668, DOI: 10.1080/1369118X.2022.2097015

To link to this article: https://doi.org/10.1080/1369118X.2022.2097015

Published online: 08 Jul 2022.

INFORMATION, COMMUNICATION & SOCIETY
2023, VOL. 26, NO. 14, 2649–2668
https://doi.org/10.1080/1369118X.2022.2097015

Routledge
Taylor & Francis Group

User-centric approaches for collecting Facebook data in the 'post-API age': experiences from two studies and recommendations for future research

Johannes Breuer [a,b], Zoltán Kmetty [c,d], Mario Haim [e] and Sebastian Stier [f]

[a]Survey Data Curation, GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany; [b]Research Data & Methods, Center for Advanced Internet Studies, Bochum, Germany; [c]Faculty of Social Sciences, Eötvös Loránd University, Budapest, Hungary; [d]CSS-Recens Research Group, Centre for Social Sciences of the Hungarian Academy of Sciences, Budapest, Hungary; [e]Department of Media and Communication, LMU Munich, Munich, Germany; [f]Computational Social Science, GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany

ABSTRACT

Although other social media platforms have seen a steeper increase in users recently, Facebook is still the social networking site with the largest number of users worldwide. A large number of studies from the social and behavioral sciences have investigated the antecedents, types, and consequences of its use. In addition or as an alternative to self-reports from users, many studies have used data from the platform itself, usually collected via its Application Programming Interfaces (APIs). However, with the drastic reduction of data access via the Facebook APIs following the Cambridge Analytica scandal, this data source has essentially become unavailable to academic researchers. Hence, there is a need for different modes of data access for what Freelon (2018) has called the 'post-API age'. One promising approach is to directly collaborate with platform users to ask them to share (parts of) their personal Facebook data with researchers. This paper presents experiences from two studies employing such approaches. The first used a browser plugin to unobtrusively observe Facebook use while users are active. The second asked participants to export and share parts of their personal Facebook data archive. While both approaches yield promising insights suitable to extend or replace self-reports, both also entail specific limitations. We discuss and compare the unique advantages and limitations of both approaches and provide a list of recommendations for future research.

ARTICLE HISTORY
Received 21 December 2020
Accepted 13 June 2022

KEYWORDS
Facebook; digital trace data; social media; data donation; API

Introduction

Despite the increasing popularity of other social media platforms (e.g., TikTok) and a plateau in the growth of user bases in many countries, Facebook is still the largest social

CONTACT Johannes Breuer johannes.breuer@gesis.org Survey Data Curation, GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany Research Data & Methods, Center for Advanced Internet Studies, Bochum, Germany

networking site (SNS) worldwide with over 2.9 billion monthly active users in the third quarter of 2021 (Statista, 2021). There is a large body of research showing that Facebook is widely used, for example, to keep in touch with friends but also to keep up with news. While large parts of this research are based on self-report data, numerous studies have also collected data from the platform itself. Thus, Facebook has not only been a popular topic for research in the social sciences but also a research tool to collect data (Kosinski et al., 2015).

Most of the studies that used Facebook as a research tool made use of the Application Programming Interfaces (APIs) offered by Facebook. However, this research came to a halt when Facebook drastically reduced the access options for its APIs in the wake of the Cambridge Analytica scandal in early 2018. In a widely recognized commentary, Bruns (2019) called the aftermath of the Cambridge Analytica scandal the 'APIcalypse' for academic research. In a similar vein, Freelon (2018) wrote that research is entering a 'post-API age.' Conversely, in his response to Bruns (2019), Puschmann (2019) also sees some positive aspects in the restrictions of API access as they bring an end to what he calls 'the wild west of social media research.' Regardless of how they judge the (re)actions by Facebook (and other social media platforms that have put measures into place to more strictly regulate access to their APIs), what all researchers agree on is the need to develop and discuss alternative models of access to social media data.

Breuer et al. (2020) distinguish between three options for collecting digital trace data (including social media data). Researchers can collect data themselves, cooperate directly with the companies that produce or hold these data, or purchase data from a third party (data resellers, social media monitoring or market research companies). The most commonly used method, so far, has been collecting data via platform APIs, despite the substantial downsides associated with this mode of data collection.

In addition to these three ways of accessing social media data, a fourth option that has recently gained increased attention in academic research is what Halavais (2019) has described as 'partnering with users to collect big data' (p. 8). There are different ways in which researchers can employ such user-centric approaches for gathering digital trace data. The two main approaches are 1) using browser plugins or (standalone) software applications, 2) making use of data export functions and having users manually share their exported data with researchers. Boeschoten et al. (2020) call this data download packages (DDPs). In contrast to previous API-based approaches, these options grant researchers independence from platforms and also largely facilitate the collection of informed consent (Breuer et al., 2021).

These two data donation approaches have been receiving increased attention among researchers who aim to use social media data, now that the risks associated with data access via platform APIs have become more obvious. However, only a few studies have used them empirically so far, and, at least to our knowledge, no study has systematically compared the two in terms of what data they collect, the kinds of studies they can be useful for, and their ethical and privacy implications. In this paper, we present experiences from two studies following the model of partnering with users to collect social media data (one from Germany, one from Hungary) that made use of such approaches. After briefly discussing previous work in the area of partnering with users for the collection of Facebook data, we present and compare the methods used in our studies. Based

on the comparison of our approaches, we provide five recommendations for future research that seeks to collaborate with users for collecting Facebook data.

Approaches and tools for user-centric Facebook data collection

While the seminal paper by Halavais has been published in 2019 as a response to social media platforms drastically reducing or entirely shutting off API access, the idea of recruiting platform users to collect data is not entirely new. Several research groups have developed and employed methods and tools for such purposes. Christner et al. (2022) provide an exhaustive critical review of what they call 'user-centric tracking tools' (p. 3). They evaluate them along the criteria of 'types of information, technical complexity, privacy implementation, user experience, and availability' (p. 1).[1] While their focus is on web tracking, several of the tools and methods that they review can also be used to collect data on/from social media platforms, including Facebook. In addition, the evaluation criteria they apply are largely also valid for dedicated (user-centric) Facebook data collection tools and we use similar criteria for the comparison of the two specific approaches we used in our own projects.

While we cannot provide a systematic comparison of existing tools here as this would be beyond the scope of this paper, we discuss some relevant examples in this section to a) identify some of the limitations of existing solutions and b) establish a background which allows for an informed evaluation of the specific advantages and disadvantages of the approaches we present in detail in this paper. In this section, we will focus on tools and approaches that can a) be used to collect Facebook data, b) are 'user-centric' in the sense that they require an active opt-in from the user, and c) do not require API access.

Browser plugins or other software applications have been developed and used by for user-centric tracking by several research groups over the last decade. Importantly, however, popular tools that have been used in quite a few publications, such as the *myPersonality* app (see, e.g., Kosinski et al., 2013; Quercia et al., 2012) or the *Digital Footprints* application (Bechmann & Vahlstrup, 2015) have relied on the Facebook API and have, hence, become unusable. Some researchers have also created bespoke browser plugins (e.g., Skeggs & Yuill, 2016; Wang et al., 2014) or Facebook apps (e.g., Vaidhya et al., 2017) for their studies, the latter of which also made use of the API.

Two key limitations of the previously used Facebook data collection tools are that they a) have relied on the Facebook APIs and/or b) have been created for the specific purpose (s) of a study. While the former poses the risk of these tools becoming unusable if APIs are altered or access becomes (more) restricted, the latter means that the applications cannot be used easily for other projects. One alternative to the use of browser plugins and other dedicated software applications is to make use of the data export functionalities that most social media platforms, including Facebook, offer their users nowadays.

The use of such data download packages (DDPs) is a much more recent phenomenon compared to apps and browser plugins. One reason for their relative novelty is that the data export functionalities for users have been implemented (or at least extended) by most social media platforms in response to the introduction of the General Data Protection Regulation (GDPR) in the European Union in May 2018. Accordingly, there are only a few studies so far that have employed this method. Two studies that used DDP have

recruited college student samples to study patterns of problematic use (Marino et al., 2017) and the exposure to news and politics on Facebook (Thorson et al., 2021). The paper by Boeschoten et al. (2020) mentioned before, however, presents a general 'framework for digital trace data collection through data donation' and discuss the strengths as well as the limitations of this methodology. Based on this conceptual work, in another recent preprint, Araujo et al. (2021) introduce 'OSD2F: An Open-Source Data Donation Framework' as a novel technical solution for data donation studies. Notably, these important contributions are more conceptual as they describe data donation as a methodological framework (Boeschoten et al., 2020) or present the specifications and implementation of a specific technical data donation solution (Araujo et al., 2021). Our paper adds to this line of work by comparing two different data donation methods (browser plugin and data download packages) based on experiences and data from two pilot studies. Another feature of our study is that we specifically focus on Facebook for which the options for collecting user-centric data via its APIs have become extremely limited. Finally, while there are a few empirical studies that have made use of data donation approaches to study Facebook use (Marino et al., 2017; Thorson et al., 2021), our studies did not use data from university student samples and were conducted in two different countries.

Study 1

Study 1: Methods

The data in Study 1 came from a non-probability web tracking panel maintained by a German market research company. The panel consists of around 2,000 participants per month (as there are dropouts the company recruits new participants on a regular basis to keep the overall sample size at approx. 2000). The panelists have agreed to use software that tracks their web browsing activities on their desktop computers and/or mobile devices. For a research project dealing with methodological questions related to combining different types of data and substantive questions related to media use and political interest and orientation, we acquired access to the web tracking data for 12 months (June 2018 to May 2019). Over the course of these 12 months, we invited the panelists to three online surveys, each containing different questions related to the methodological and substantive interests of the project. The results we present here are based on the second online survey which was conducted in March 2019.

As part of this survey, participants were asked whether they have a personal Facebook account. Those who indicated that they did were asked whether they would be willing to install and use a plugin for their desktop browser that unobtrusively captures public posts from their personal Facebook feeds. As the plugin only collects public posts from the users' Facebook feeds, these primarily contain posts from pages (e.g., of news outlets, brands, celebrities, or other institutions and individuals that people like/follow) and public groups. However, the plugin may also collect posts from the users' Facebook friends in case their profiles or specific posts are public. The plugin was available for Firefox and Chrome on desktop computers and respondents could install it through the respective official plugin stores. In addition to the posts themselves, the plugin captured the numbers of likes, shares, and comments with which a public post was presented to the respective user as well as some additional metadata. Besides data about the individual

Figure 1. Exemplary data from the browser plugin.

post, the plugin also provides some data about the user's Facebook session. Figure 1 shows exemplary data from the plugin for one Facebook post by a German newspaper. A more detailed description of the plugin can be found in the paper by Haim and Nienierza (2019).

For installing the plugin, participants received an incentive of five Euros (in addition to the incentives they received from the market research company for participating in the web tracking and completing the online survey). In the online survey, respondents were presented with a brief informed consent that provided a description of the browser plugin as well as the purpose of this data collection. For more detailed information, respondents could access an extended data privacy information page via a link provided in the short description within the survey. This procedure as well as the information provided in the informed consent and the extended privacy information were adapted (with permission) from a study by Sloan et al. (2020) and are described in detail in another paper (Haim et al., 2021). In order to connect the survey data with the Facebook data, participants were asked to generate a six-digit code following a certain pattern that they also had to enter when installing the plugin. During plugin installation, respondents were, once again, informed about the research purposes and asked to give their informed consent to the data collection by the plugin. Once they had installed the plugin, participants could also login to a project website to see and even delete any collected data. Notably, none of the participants made use of this option.

Study 1: Data

In the month in which the online survey was conducted (March 2019), there were 1931 active web tracking panelists. Of those, 1,240 (64.2%) completed the online survey. The age of those who completed the survey ranged from 16 to 68 (M = 45.4, SD = 12.8) and 49% of the sample was female.

In our sample, 79.4% (n = 985) reported having a personal Facebook account. Of those, n = 785 received the question whether they would be willing to install the browser

plugin that collects their Facebook data. The reason that not every participant received this question was a technical issue.[2] When we discovered this issue (we used and regularly tested the plugin ourselves over the course of the data collection phase), we had to pause recruitment for the Facebook data collection for several days until the cause was identified and fixed in the plugin. Once this problem was solved, the question asking to install the plugin was included again in the online survey. Of the n = 785 participants who received the respective question, 59% (n = 463) indicated that they were willing to install the plugin.[3]

Although n = 463 participants consented to installing the plugin, in the data gathered with the plugin, we could only identify n = 301 of the participants (152 of those installed the plugin for the Chrome browser and 109 for the Firefox browser; the remaining 40 participants never visited Facebook after plugin installation, which is where browser detection took place). There are several possible reasons for the dropout at this stage. Part of it is likely due to participants being unable to reproduce their participant code when installing the plugin (although they were asked to write it down in the survey and the instructions for generating it were presented again when people activated the plugin). Others may have changed their mind or simply forgotten to install the plugin. The flowchart in Figure 2 summarizes the different stages of dropout.

Despite the dropout at various stages of the recruitment process, a total of N = 50,698 sessions (i.e., Facebook feeds) were recorded, including a total of N = 793,907 public posts. On average, a median of 32 sessions (M = 194; SD = 493) including a median total of 388 public posts (M = 3,077; SD = 8,424) per session were recorded per participant. The collected data also show some fluctuation in the daily use of the plugin (see Figure 3). Participants who had already installed the plugin before the technical issue was discovered (see above) were asked to update it. This quite likely affected some respondents' willingness to (continue to) use the plugin.

Study 2

Study 2: Methods

The second study made use of what Boeschoten et al. (2020) call data download packages (DDPs) by asking Hungarian participants to export their own Facebook data and share it with the researchers. N = 150 respondents took part in this study.[4]

The sample was a non-probability quota sample, with quotas for age and gender. The participants had to be regular Facebook users (i.e., use the platform at least on a weekly basis). All the respondents were Hungarian, living in the Eastern part of the country, mostly in one major city. The fieldwork was done by a professional market research company from April to September 2019. Study participants received an incentive of ten Euros. The market research company used their existing participant pool for recruitment. In addition, they also recruited people through student cooperatives (i.e., special organizations that outsource university students for different jobs).

Participants who agreed to take part in the study were sent a project description and were invited to come to the office of the market research company. As exporting one's personal data from Facebook can take a while, participants were asked to initiate the data export process one day before the scheduled meeting. To facilitate the process,

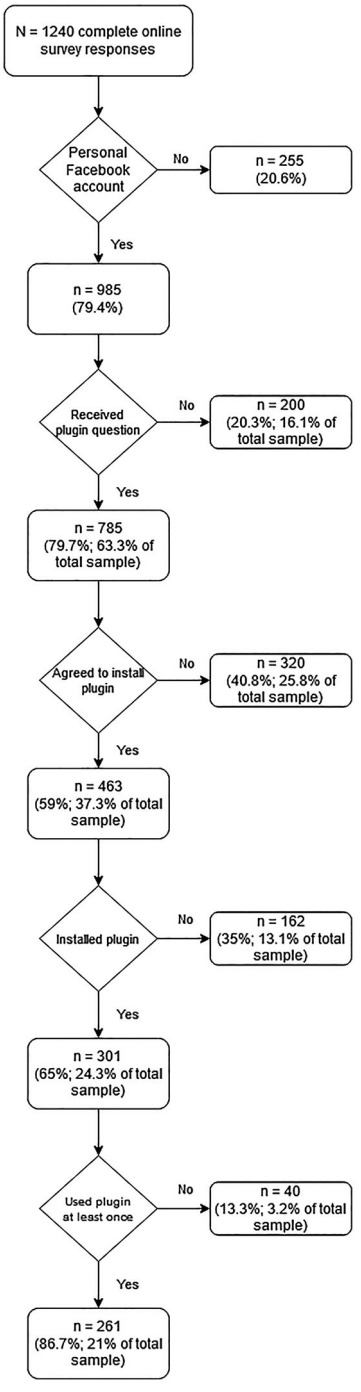

Figure 2. Flowchart illustrating dropout stages in Study 1.

participants were provided with a detailed description of how to access their Facebook data and how to start the export. Before starting the study, participants were asked to sign an informed consent form. In an initial testing phase, we identified several problems

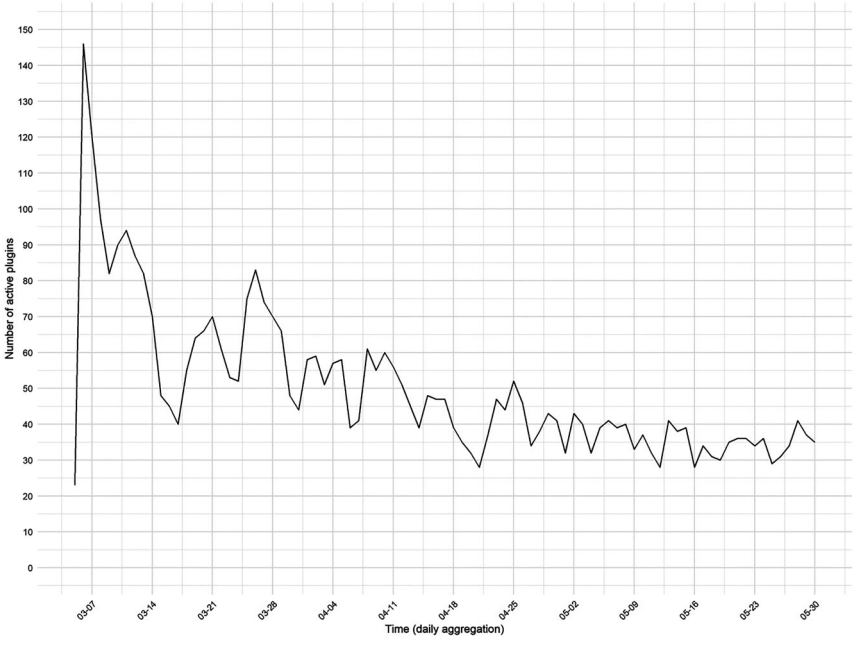

Figure 3. Fluctuations in the Facebook-plugin use.

with character encodings and different program versions during data pre-processing (see below). To avoid these problems, the market research company used one dedicated laptop for this study. Once in the office, participants were asked to log into their Facebook account using the designated laptop and then downloaded their prepared Facebook data profile archive in JSON format.

Some of the data that users can export from Facebook, such as their photos or private messages, are quite sensitive, and participants may, understandably, not be willing to share those. To take this into account, before exporting their data, participants were able to choose which parts of their data types they wanted to share. This process is illustrated in Figure 4.

Before the data collection, we discussed with university students who participated in the preparation of the project which type of data they would be willing to share and which type of data they would not share with researchers. Based on these discussions, we decided to omit private messages and search histories from the list of requested data. We also decided to exclude some data on activities on Facebook that many users rarely or even never engage with. A typical example for this is the Facebook marketplace activity. For reasons of privacy and to avoid potential data storage issues, we also did not ask participants to share audio-visual content (photos and videos). Despite the exclusion of these types of data, the data collected in this study still cover a wide range of Facebook activities: posts, comments, likes and reactions, liked pages, friends, profile information, and data about ads. The data span the whole time of participants' Facebook use. Similar to what happened in Study 1 when the structure of the Facebook feed changed, we also had to pause data collection in Study 2 due to technical changes on the side of Facebook.

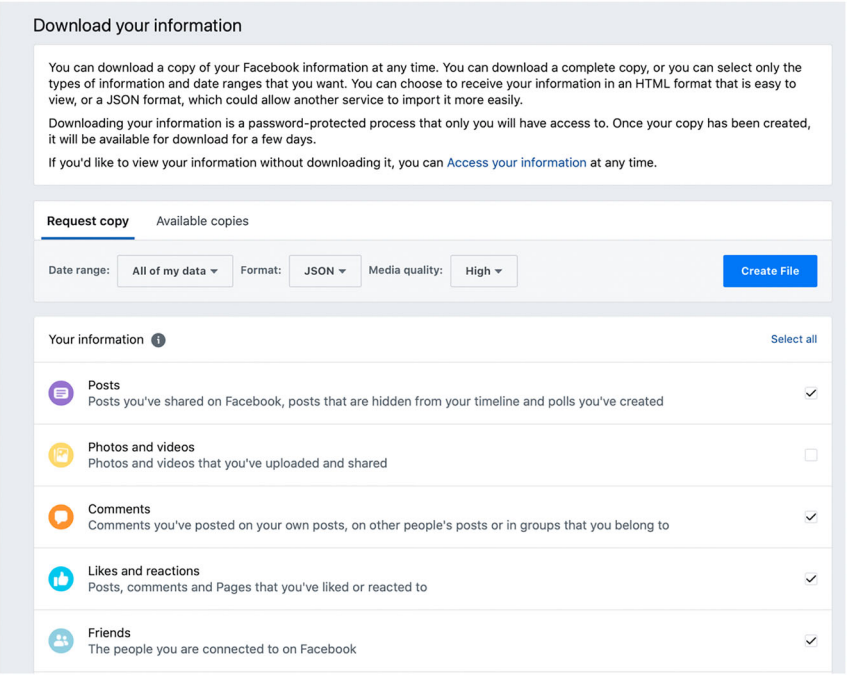

Figure 4. Screenshot from the Facebook archive download page with the list of options for the DDPs.

Facebook suspended its archive download function for several weeks in August 2019, so we had to pause the data collection for this period.

Importantly, preserving participants' privacy was a key issue in the data collection process. To ensure this, the raw data were anonymized right after the export. An R script that replaced all person names with hashed IDs was run by the interviewer while the respondent was present (this is similar to the approach employed by Mancosu & Vegetti, 2020). Thus, not only the name of the participant but also the names of all their Facebook friends (who were in their friend list) were hashed. The same hashing method was used in the entire process, which allows us to link the participants to their respective Facebook contacts. Figure 5 illustrates this anonymization process with an example.

Figure 5. Example of the anonymization process (with fake names).

In addition to sharing their Facebook data, participants were asked to fill out an online questionnaire. This questionnaire included various questions about politics, media use, self-representation, mental health, spare-time activities, and music preferences. To link these two data sources, the same ID code was used for the online questionnaire and storing the Facebook data. The raw Facebook data were deleted after the anonymization process. The market research company only shared the anonymized data with the principal investigator of the study. Such 'local processing' is also proposed by Boeschoten et al. (2020) as a suitable way of increasing data privacy.

Study 2: Data

As this was a first exploratory study with a limited budget, we were only able to recruit a relatively small non-probability convenience sample. The age of those who completed the survey ranged from 18 to 71 (M = 30.3, SD = 13.4) and 75.3% of the sample was female.

Despite the relatively modest sample size of N = 150, the resulting data set is quite large and complex. While discussing all features included in this data set is beyond the scope of this paper, we present a few key dimensions in the following.

Friends: The data regarding participants' Facebook friends include lists of all their friends, the timestamp for the start of the friendship, the hashed name of the friend, and the estimated gender of the friend (we explain and discuss this in the section on 'Legal, privacy and ethical implications'). We also have information about rejected and pending friend requests as well as removed friends. In total, this data set contains 115,955 friend links, with 78,569 unique friends. On average, participants had 773 friends (SD = 513), with the highest number being 2,840.

Pages: The data include lists of all pages followed by participants plus a timestamp, indicating when they started following that page. Overall, the 150 participants in our study followed 83,232 pages. The number of unique pages is 52,700. On average, participants followed 562 pages (SD = 1,569). The highest number of pages followed by a participant in our sample was 17,724.

Reactions: This data set contains all the reactions of the users, with a timestamp and the type of the reaction (the most frequent one being likes). It also contains the target of the reaction (i.e., if it was a friend or a page). Reactions to posts by friends are, by far, more common (76.5%) than reactions to content from pages. Importantly, the information about the content to which the user reacted that we have is on the metadata level. Our data set does not include the actual content, only information about the type of content. The content is categorized into 11 types plus an 'other' option: Post, Picture, Comment, Video, Link, Album, Life Event, GIF, Activity, Note, Watch. Most of the reactions in our data were reactions to posts (45.0%) and pictures (43.8%). The total number of reactions in our data set is 1,802,430.

Other Facebook activities: Besides reacting to content, there are many other types of activities that people can engage in on Facebook. The most common ones are posting and commenting, but users can also upload photos, videos, share links, play games, start votes, attend events, etc. We have created 20 categories to map all these activities (see Figure 6).

Overall, records of 346,407 activities are included in our data set. The most frequent activity is posting. For the activity of posting, we have the timestamp of the activity as well

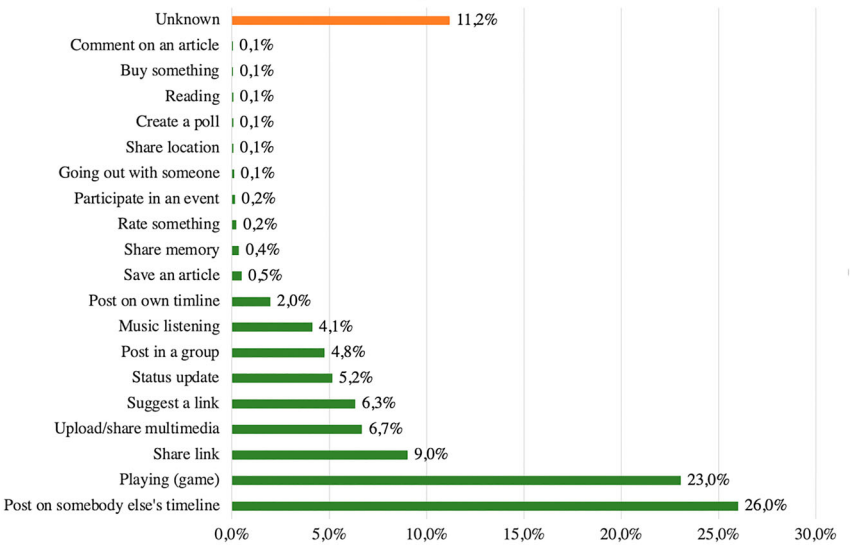

Figure 6. Types of other Facebook activities.

as the actual content of the post, the (masked) name of the friends involved in the activity, the links shared by the users, and the number of comments on the post (if applicable). In addition, the data also include further information, such as the name of the event or group for/in which the user was posting/uploading/sharing.

Ad interests: Facebook categorizes every user for advertising purposes. This is an algorithmic classification of the users based on their own likes, activities, used keywords, and their friends' preferences (DeVito, 2017). The algorithm is a black-box, so we can only observe the results of the categorization. There is no timestamp in this data set, just the category names per user. The 150 participants were linked to 105,642 interest categories (18,689 unique ones). On average, participants in our sample were linked to 704 categories (SD = 479), with the highest value being 3,727. A wide range of categories appeared in the classification, from general ones, such as sports or food, to more specific ones, such as snowboarding and hamburgers. Notably, the interest categorization might be biased in smaller countries as country-specific categories are missing. An example for our data is 'Roma-music', a popular specific music genre in Hungary that does not appear in the interest categorization (Kmetty & Németh, 2022).

Comparison of the two approaches

Based on our experiences from the two pilot studies, we compare the two approaches with regard to two key dimensions: 1) the types of data they generate (and what those can be used for), 2) their legal, privacy, and ethical implications. We consider their specific advantages as well as the challenges and limitations associated with the two approaches across these different domains to aid researchers in taking into account all relevant aspects for making an informed choice which method to use for collecting Facebook data.

Types of collected data

The browser plugin used in Study 1 collects all public posts from a participant's Facebook news feed. These data are very well-suited for studying exposure to news or other forms of content that are typically shared publicly on Facebook, e.g., by political parties, companies, or governmental institutions that users follow/like.

Although the data it collects can be used to study a variety of topics, a clear limitation of using the browser plugin is that it can only capture desktop use. For many Facebook users the mobile app may be the preferred way of access. In the online survey for Study 1, Facebook users were asked to estimate what percentage of their use occurs via the browser on their desktop computer(s). On average, they indicated that 46.3% of their Facebook use took place via a desktop computer. Notably, n = 133 respondents indicated that they never access their Facebook account via the browser on their desktop computer(s), while n = 127 participants reported that all of their Facebook use (i.e., 100%) happens via their desktop browser. Notably, this distinction between devices used to access the platform becomes irrelevant or at least much less relevant in case of API-based data access. However, unlike for Twitter, user-centric API access has become extremely limited for Facebook.[5]

Unlike the browser plugin, the DDP approach can be used to collect data on both desktop as well as mobile platform use. In general, the DDP approach employed in Study 2 can be used to collect a large variety of different types of data. Hence, even with relatively small sample sizes, the data generated with this method are quite detailed and rich. With these data it is possible to track the whole Facebook history of participants. This allows a detailed temporal analysis of behavior change on the individual as well as the group level. Such data can, e.g., allow tracking the effect of personal life events (e.g., starting university) or significant events and situations (e.g., elections, COVID-19 pandemic). Another advantage is the access to both private and public posts. A limitation of the DDP approach is that it cannot provide data on what users actually saw in their news feeds, meaning that it cannot be used to study exposure to news or other content.

Legal, privacy, and ethical implications

The browser plugin used in Study 1 requires active informed consent from the users and offers them control over what data they share as they can uninstall the plugin, temporarily turn it off, use a browser for which they have not installed the plugin, or selectively delete any of the collected data. Compared to DDPs that may, e.g., include private messages or posts and comments in which a user's friends are mentioned/tagged, collecting data only on public posts from the users' feeds is less sensitive in terms of privacy. Nevertheless, it is possible that the plugin also collects posts from a user's personal contacts if they post publicly. Also, as the minimum age for having a Facebook account is 13, it may well be that data collected with the plugin through consenting users may contain posts from non-consenting minors. Hence, extra care with regard to data privacy is warranted when researchers work with such data; especially if they want to work with the text data from the posts.

While author names for public posts can be excluded, mentioned usernames within the post text cannot be easily removed automatically. One possible way to address this

is working with whitelists, so that only posts from specified sources are kept in the data. In another study that used the browser plugin to look at online news use (Authors, 2021), the whitelisting approach was used to make sure that only posts from a specified list of news sources were included in the analyses. Another option could be to identify user links in the textual posts (which are recognizable as Facebook highlights them) and remove them/replace them by some placeholder or hash. In general, the full posts are potentially sensitive data. Hence, in order to maximize privacy, the full text data should ideally be deleted once the derived or aggregate variables required for analyses are created.[6]

Importantly, these issues do not only have ethical but also legal implications. Public Facebook posts by private users are very likely to contain real names and other personal data that have to be treated with special care according to the GDPR in Europe (but likely also privacy and data protection laws in other jurisdictions). Another relevant legal aspect is that the browser plugin used in Study 1 employs screen scraping and Facebook's ToS do not permit large-scale database creation through scraping. While there is an ongoing debate about the legal status of web scraping for academic research purposes and the legal and ethical ramifications of violating platform ToS, a prominent court case from the United States (Sandvig v. Barr, 2020) indicates that academics 'should not incur criminal and civil liability if they do not abide by a site's posted terms of use' (Thomas, 2021) in their research. For the European context, a legal opinion included in a recent publication on big data in the social sciences by the German Data Forum (RatSWD, 2020) also states that web scraping for the purpose of academic research in the public interest should be possible and be treated differently than web scraping for commercial purposes from a legal perspective. Nevertheless, the recent actions that Facebook took against researchers from New York University (Vincent, 2021) and the German non-profit Algorithm Watch (Brandom, 2021) show that there is a risk associated with the use of scraping approaches.

From an ethical perspective, a positive attribute of the DDP approach used in the second study is the transparency for the participants. While exporting and sharing the data requires some effort from the participants, and, thus, increases respondent burden compared to using the browser plugin, this method allows users to see exactly what types of data they will share with the researchers. Nevertheless, the data donation approach employed in Study 2 also raises a set of legal, privacy, and ethical questions that need to be considered and addressed. One of the major challenges is to ensure the anonymity of users and their peers as the data contains various pieces of information that can make participants identifiable. Also, if, for example, a participant's friend posted on their Facebook feed, this post is available in the data set. A way of increasing privacy in this regard can be to delete the timeline posts from other people (or to exclude them from the DDP).[7]

In order to ensure privacy for the participants and their contacts, the raw data were anonymized directly after the export in the pilot study. All the names of the users and their friends have been replaced with hashed IDs. A key disadvantage of this process is that masking the names results in the loss of information about the composition of the participants' networks. One piece of information that may be relevant in this regard is the gender of participants' friends. To address this problem in the pilot study, prior to hashing, the gender of the user's friend was detected via their names.[8] Notably, while their names were hashed, the detection of the gender of participants' Facebook friends

in the second pilot study constitutes a use of data from users who have not actively consented to their data being used. To maximize the protection of non-consenting users' data, beyond hashing names, the strictest protection mechanisms would be to only ask users to export the data that are truly needed for the study and avoid those that are likely to contain data from others. In general, with the DDP approach employed in Study 2, due to the amount and detail of information about participants' Facebook activities it cannot be ruled out that these can be used to identify individuals, even when names are removed or hashed. This disclosure risk associated with these data means that they cannot easily be shared. Van Atteveldt et al. (2021) present several suggestions for ways in which such data can potentially be shared. One option is to only share aggregated data. While this decreases the reuse value of the, it can be used to generate what King (1995) has called replication data sets.

Recommendations & conclusion

Based on the experiences from our pilot studies and the systematic comparison of the two approaches, we provide five recommendations for researchers who want to collect Facebook data (or social media data more generally) via users themselves. To facilitate making an informed choice between the two data donation approaches we presented in this paper, Table 1 sums up the main characteristics of the two methods as described in the previous sections.

Table 1. Comparison of key features of the two data donation approaches for collecting Facebook data used in our studies.

	BROWSER PLUGIN	DATA DOWNLOAD PACKAGE
Captured Data		
Desktop use	✓	✓
Mobile use	X	✓
Exposure to content in news feed	✓	X
User interaction (e.g., likes, comments)	X	✓
Contacts/networks	X	✓
User (profile) information	X	✓
Time period covered	Current use (while the plugin is installed and used)	Past use (possible to capture the whole usage history)
Legal & Ethical Aspects		
Compliant with platform ToS	X[1]	✓
Possible to obtain informed consent	✓	✓
Transparency for users	Medium	High
Privacy		
Control over what is collected/shared	High	Medium
Disclosure risk for participants (based on raw data)	Low	High
Risk of inclusion of data from other private users	Low	High
Deployment & Maintenance		
Maintenance effort for researchers	High	Low
Respondent burden	Medium	High
Amount of user support required	Medium to High	Medium to High

Note. [1] However, a recent court decision in the United States indicated that academic researchers should not incur criminal or civil liability when they violate a platform's ToS (Sandvig v. Barr, 2020) for their research. Similarly, legal opinions from Germany suggest that non-commercial academic researchers can create databases for their research even if this is not in accordance with platform ToS as long as privacy rights are respected (see, e.g., RatSWD, 2020 for an expert opinion on the subject).

1. Focus on research interests and available resources to make an informed choice regarding the appropriate data collection method and required data

Ultimately, the research interest should guide the choice of an appropriate data collection method. As the two approaches presented in this paper generate different types of data, researchers should choose the one that produces the kind of data they need to answer their research questions. For example, if one is interested in the determinants or effects of exposure to (specific types of) content on Facebook, using the browser plugin is the more suitable approach. If, however, one wants to study the activities, interactions, or networks of Facebook users, the DDP method is the preferable solution.

In addition to the kind of data needed to answer the specific research questions and the advantages and disadvantages of the data collection options, the available resources for any study/project should be considered: How much time can be spent on the data collection? What expertise/skills are available in the team to handle anonymization and/or provide user support? How much money is available for incentivizing participants (or paying other costs, such as commissioning a market research company for the data collection)?

2. Be aware of potential biases in the sample

Asking Facebook users to share their data or to use a browser plugin that collects data on their Facebook use likely introduces biases into the sampling procedure (Boeschoten et al., 2020; Jürgens et al., 2020; Sen et al., 2021). The most obvious one is a self-selection bias as participants have to actively opt-in and also become active in other ways (installing the plugin, exporting their archive, etc). In addition, to motivate participants it is typically necessary to provide them with incentives which, depending on the type and amount of compensation, can further bias the sampling. There may also be other sources of bias. For example, more intense users of Facebook may be more likely to participate, while people with higher privacy concerns may be less likely to do so.

3. Identify and deal with disclosure risks by implementing data privacy measures and data minimization

Both presented methodological approaches produce potentially sensitive data containing personal information, such as names of users and their contacts. While the browser plugin used in Study 1 only collects public posts from the users' Facebook feeds, they may also contain posts from Facebook friends with public profiles (or if they set the visibility for individual posts to public). This example also illustrates that not only the privacy of users who participate in the study but also that of their Facebook friends needs to be considered and protected. For data from DDP, the procedure of hashing usernames employed in Study 2 is a suitable method of increasing the privacy of both the participants and their Facebook friends. For the browser plugin data, a safe solution for removing public posts from users' Facebook friends from the data can be the use of whitelists that contain the names of specific

pages or groups you are interested in. Another approach that preserves the privacy of participant's Facebook friends is to automatically delete their data, if this is possible for the data types that are collected.

In general, full texts (e.g., of posts or comments) are a disclosure risk. As these are very difficult to anonymize, they need to be treated with extra care. One way to achieve this is to delete the raw text data once the derived variables (e.g., word counts or the detection of specific named entities) required for further analyses have been created (which is also in line with the procedure suggested by Sloan et al., 2020 for handling linked survey and Twitter data). This reduces the replicability of analyses but is a substantial gain for data privacy. While social media texts generated by users are a disclosure risk across platforms, this issue is particularly pronounced for Facebook data. A main reason for this is that, unlike tweets, posts and comments by users on Facebook are typically not world-readable, unless the users have public profiles or post on public pages or in public groups.

In Study 2, audio-visual content was deliberately excluded. However, if researchers also want to use photos or videos generated by the users for their analyses, the issue of disclosure risk and privacy protection becomes even more pronounced and requires further technical solutions (e.g., cropping or blurring of images). Again, the use of extracted information (e.g., through object detection or classification via computer vision) instead of the raw data can be a viable approach here. In general, it is advisable to ensure that only the data that are truly needed to answer the research questions of a study/project are collected. The guiding principle should be data minimization.

4. Establish transparency for your respondents

A general benefit of collaborating with users to collect Facebook data is that it can be more transparent for the users what data are collected and how they are used. When partnering with the users, researchers need to get informed consent from participants for collecting their data. Participants should be informed about what data is collected, for what purpose(s) the data is used, how it is stored, and who will have access to it. Of course, the informed consent needs to adhere to relevant legal regulations (e.g., the GDPR in the European Union) and follow ethical standards (as defined, e.g., by Institutional Review Boards or scientific associations). In practice, it is also necessary to design the informed consent in a way that the necessary information is provided without potentially overwhelming participants with technical details. Sloan et al. (2020) and Breuer et al. (2021) provide a good template for an informed consent for the collection and linking of social media data that can also be used/adapted for different types of Facebook data. What can further increase transparency and also serve as a way of increasing privacy protection is to offer participants the option to see and selectively delete their data. The browser plugin we used in Study 1 provides this option. In Study 2, participants could decide which parts of the data from their personal Facebook archive they wanted to share with the researchers, and, after the pre-processing of their raw data, they had the option to delete any data they wanted (similar to Study 1, no participant made use of the latter option).

5. Implement strategies for minimizing dropout and dealing with potential contingencies

As the experiences from Study 1 have shown, there are many stages at which dropout can occur and the dropout at these different stages can have various reasons. While some sources of dropout are foreseeable, such as people's unwillingness to share their data, others may not always be expected. A good example for this is the data loss due to technical issues caused by changes in the structure of the Facebook newsfeed in Study 1. Another challenge is that, while sources of dropout may be known, their extent can be quite difficult to predict. Although you may have an estimate of the share of Facebook users in your sample if you have data (from other studies) on the share of Facebook users in the country/countries where your data comes from, you may not be able to come up with a good estimate of the share of participants who are willing to share their data with you or the share of people who are willing to share their data but for whom the sharing does not work (e.g., due to technical problems). Such considerations are even more important if you want to achieve a specific sample size (e.g., because you determined a required sample size for a specific model via an a-priori power analysis). For these reasons, it is important to implement measures to reduce dropout risks. Such strategies include the use of an appropriate informed consent, the provision of adequate incentives (while also keeping in mind potential biasing effects of those), offering comprehensible instructions and support to participants (if there are technical issues), as well as reducing the effort required for participants (e.g., easy availability and use of the browser plugin). You should also develop contingency plans. Before you start your data collection, you should come up with answers to questions, such as what happens if the willingness to share Facebook data is (much) lower than expected, if people decide to withhold or delete large parts of their data, or if there are technical issues.

Taken together and considered in combination with the experiences from our two studies, these five recommendations can hopefully provide useful guidance for researchers who want to collect Facebook data in future studies.

Overall, despite the challenges associated with the approaches presented in this paper, we agree with the assessment by Halavais (2019) that partnering with the users is a promising approach for collecting Facebook (and other social media) data in the 'post-API age' (Freelon, 2018). Recent work, such as that by Araujo et al. (2021) shows that many researchers see the potential in user-centric approaches and have begun to develop solutions for implementing and improving them. While there has been some previous work in this area, our in-depth comparison of two different methods (one using a browser plugin, the other one a DDP) that is based on actual data from two pilot studies with diverse samples highlights the advantages as well as some of the key challenges associated with such approaches. We hope that the experiences and recommendations we present in this paper can aid researchers in adopting and refining methods for the user-centric collection of Facebook and other social media data.

Notes

1. Notably, their review also includes the tool we used in our first study (described in detail in the paper by Haim & Nienierza, 2019).

2. Shortly after the data collection started, Facebook slightly changed the structure of the news feed. This caused an issue with the scrolling listener of the plugin: When users clicked on a picture from their feed to enlarge it, the scrolling listener interpreted the feed to change and, hence, searched for new posts, but ended up in a conflicting state where the enlarged picture was interpreted as a feed that did not contain any posts previously identified. For the users this caused the browser tab to remain occupied which essentially turned it unusable.
3. Both the share of Facebook users (79.4%) and consent to install the plugin (59%) were quite high in our study. What has to be taken into account here, however, is the nature of our sample. First, the study was conducted in Germany where Facebook is still widely popular. Second, the starting point for the sampling process for this study was a panel of people who agreed to having their internet use tracked by a market research company. These participants are likely to be more willing to consent to having their Facebook use tracked than members of the general population.
4. 150 was the target number set by the researchers and the fieldwork company stopped data collection once this number of respondents was reached.
5. This key difference between the Twitter and Facebook APIs illustrates in what way(s) data access decisions by companies have major implications for the kinds of research that can be conducted.
6. Of course, this comes at the cost of a potentially reduced reproducibility and reusability of the analyses and data.
7. Notably, in the data from our pilot study, the majority of timeline posts from other users were birthday greetings which are quite unlikely to contain information that could be of interest for scientific studies.
8. For the detection, we compared the names of the user's friends with a list of Hungarian first names. In some cases, we did not find a match, because some people have nicknames, foreign names, or both a male and a female first name. However, on average, 90 percent of the names in the users' networks were classified before the hashing. Although this process hashed the names of the participants and their friends, it could not hash the names of those users who are not Facebook friends with the participants (possibly because the friendship was deleted after the post).

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Johannes Breuer http://orcid.org/0000-0001-5906-7873
Zoltán Kmetty http://orcid.org/0000-0002-6775-8938
Mario Haim http://orcid.org/0000-0002-0643-2299
Sebastian Stier http://orcid.org/0000-0002-1217-5778

Notes on contributors

Johannes Breuer is a senior researcher at the department Survey Data Curation of GESIS – Leibniz Institute for the Social Sciences in Cologne, Germany, and co-leads the team Research Data & Methods at the Center for Advanced Internet Studies (CAIS) in Bochum, Germany. His research focuses on the use of digital behavioral, the use and effects of digital media, computational methods, and open science.

Zoltán Kmetty, Ph.D., is a senior research fellow at the Centre for Social Sciences, CSS-RECENS research group; and an associate professor at the Eötvös Loránd University Faculty of Social Sciences,

Sociology department. He has diverse research interests, including political sociology, network studies, and suicide research. He is an expert in methodology, survey design, and quantitative analysis.

Mario Haim is a full professor for communication science at the Department of Media and Communication at LMU Munich. His research focuses on political communication, computational journalism, news use within algorithmically curated media environments, and computational social science.

Sebastian Stier is a Senior Researcher at the Department Computational Social Science of GESIS – Leibniz Institute for the Social Sciences in Cologne. His research focuses on political communication, comparative politics, political behavior, and the use of digital behavioral data and computational methods in the social sciences.

References

Araujo, T., Ausloos, J., van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., van de Velde, B., de Vreese, C., & Welbers, K. (2021). *OSD2F: An open-source data donation framework*. *SocArXiv*, https://doi.org/10.31235/osf.io/xjk6t

Bechmann, A., & Vahlstrup, P. B. (2015). Studying Facebook and Instagram data: The digital footprints software. *First Monday*, *20*(12). https://doi.org/10.5210/fm.v20i12.5968

Boeschoten, L., Ausloos, J., Moeller, J., Araujo, T., & Oberski, D. L. (2020). Digital trace data collection through data donation. arXiv preprint arXiv:2011.09851

Brandom, R. (2021, August). Facebook shut down German research on Instagram algorithm, researchers say. *The Verge*. https://www.theverge.com/2021/8/13/22623354/facebook-instagram-algorithm-watch-research-legal-threat

Breuer, J., Al Baghal, T., Sloan, L., Bishop, L., Kondyli, D., & Linardis, A. 2021). Informed consent for linking survey and social media data—Differences between platforms and data types. *IASSIST Quarterly*, *45*(1), 1–27. https://doi.org/10.29173/iq988

Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*, *22*(11), 2058–2080. https://doi.org/10.1177/1461444820924622

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Christner, C., Urman, A., Adam, S., & Maier, M. (2022). Automated tracking approaches for studying online media Use: A critical review and recommendations. *Communication Methods and Measures*, *16*(2), 79–95. https://doi.org/10.1080/19312458.2021.1907841

DeVito, M. A. (2017). From editors to algorithms. *Digital Journalism*, *5*(6), 753–773. https://doi.org/10.1080/21670811.2016.1178592

Freelon, D. (2018). Computational research in the post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Haim, M., Breuer, J., & Stier, S. (2021). Do news actually "Find Me"? Using digital behavioral data to study the news-finds-me phenomenon. *Social Media + Society*, *7*(3). https://doi.org/10.1177/20563051211033820

Haim, M., & Nienierza, A. (2019). Computational observation. *Computational Communication Research*, *1*(1), 79–102. https://doi.org/10.5117/CCR2019.1.004.HAIM

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society*, *22*(11), 1567–1581. https://doi.org/10.1080/1369118X.2019.1627386

Jürgens, P., Stark, B., & Magin, M. (2020). Two half-truths make a whole? On bias in self-reports and tracking data. *Social Science Computer Review*, *38*(5), 600–615. https://doi.org/10.1177/0894439319831643

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 541–559. https://doi.org/10.1017/S1049096500057607

Kmetty, Z., & Németh, R. (2022). Which is your favorite music genre? A validity comparison of Facebook data and survey data. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, *154*(1), 82–104. https://doi.org/10.1177/07591063211061754

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543–556. https://doi.org/10.1037/a0039210

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Mancosu, M., & Vegetti, F. (2020). What You Can scrape and what Is right to scrape: A proposal for a tool to collect public Facebook data. *Social Media + Society*, *6*(3). https://doi.org/10.1177/2056305120940703

Marino, C., Finos, L., Vieno, A., Lenzi, M., & Spada, M. M. (2017). Objective Facebook behaviour: Differences between problematic and non-problematic users. *Computers in Human Behavior*, *73*, 541–546. https://doi.org/10.1016/j.chb.2017.04.015

Puschmann, C. (2019). An end to the wild west of social media research: A response to Axel Bruns. *Information, Communication & Society*, (11), 1582–1589. https://doi.org/10.1080/1369118x.2019.1646300

Quercia, D., Bodaghi, M., & Crowcroft, J. (2012, June 22–24). *Loosing "friends" on Facebook*. Proceedings of the 3rd annual ACM Web Science conference on - WebSci '12. https://doi.org/10.1145/2380718.2380751

RatSWD [German Data Forum]. (2020). Big data in social, behavioural, and economic sciences: Data access and research data management. *RatSWD Output*, *4*(6). Berlin, German Data Forum (RatSWD). https://doi.org/10.17620/02671.52

Sandvig v. Barr. (2020). Civil Action No. 16-1368 (JDB) (United States District Court for the District of Columbia March 27, 2020). https://www.aclu.org/sandvig-v-barr-memorandum-opinion.

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A total error framework for digital traces of human behavior on online platforms. *Public Opinion Quarterly*, 85(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Skeggs, B., & Yuill, S. (2016). The methodology of a multi-model project examining how Facebook infrastructures social relations. *Information, Communication & Society*, *19*(10), 1356–1372. https://doi.org/10.1080/1369118X.2015.1091026

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2020). Linking survey and Twitter data: Informed consent, disclosure, security, and archiving. *Journal of Empirical Research on Human Research Ethics*, *15*(1–2), 63–76. https://doi.org/10.1177/1556264619853447

Statista. (2021). *Number of monthly active Facebook users worldwide as of 3rd quarter* 2021. https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

Thomas, L. (2021, June). *Supreme Court ruling that limits hacking law supports U-M researcher*. https://news.umich.edu/supreme-court-ruling-that-limits-hacking-law-supports-u-m-researcher/

Thorson, K., Cotter, K., Medeiros, M., & Pak, C. (2021). Algorithmic inference, political interest, and exposure to news and politics on Facebook. *Information, Communication & Society*, *24*(2), 183–200. https://doi.org/10.1080/1369118X.2019.1642934

Vaidhya, M., Shrestha, B., Sainju, B., Khaniya, K., & Shakya, A. (2017, November 15–18). *Personality traits analysis from Facebook data*. The 21st International Computer Science and Engineering Conference (ICSEC), 1–5. https://doi.org/10.1109/ICSEC.2017.8443932

Van Atteveldt, W., Althaus, S., & Wessler, H. (2021). The trouble with sharing your privates: Pursuing ethical open Science and collaborative research across national jurisdictions using sensitive data. *Political Communication*, 192–198. https://doi.org/10.1080/10584609.2020.1744780

Vincent, J. (2021, August). Facebook bans academics who researched ad transparency and misinformation on Facebook. *The Verge*. https://www.theverge.com/2021/8/4/22609020/facebook-bans-academic-researchers-ad-transparency-misinformation-nyu-ad-observatory-plug-in

Wang, Y., Leon, P. G., Acquisti, A., Cranor, L. F., Forget, A., & Sadeh, N. (2014). A field trial of privacy nudges for Facebook. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2367–2376. https://doi.org/10.1145/2556288.2557413