Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age

Rebekah Tromble

Abstract
In the wake of the 2018 Facebook–Cambridge Analytica scandal, social media companies began restricting academic researchers’ access to the easiest, most reliable means of systematic data collection via their application programming interfaces (APIs). Although these restrictions have been decried widely by digital researchers, in this essay, I argue that relatively little has changed. The underlying relationship between researchers, the platforms, and digital data remains largely the same. The platforms and their APIs have always been proprietary black boxes, never intended for scholarly use. Even when researchers could mine data seemingly endlessly, we rarely knew what type or quality of data were at hand. Moreover, the largesse of the API era allowed many researchers to conduct their work with little regard for the rigor, ethics, or focus on societal value, we should expect from scholarly inquiry. In other words, our digital research processes and output have not always occupied the high ground. Rather than viewing 2018 and Cambridge Analytica as a profound disjuncture and loss, I suggest that digital researchers need to take a more critical look at how our community collected and analyzed data when it still seemed so plentiful, and use these reflections to inform our approaches going forward.

Keywords
digital data, APIs, social media research, research ethics

Introduction
When the Facebook–Cambridge Analytica (CA) scandal burst into the headlines in March 2018, it cast a spotlight on tech companies’ uses and abuses of personal data (Cadwalladr & Graham-Harrison, 2018). For years, platforms such as Facebook and Google had been gobbling up user data, tracking the minutiae of our daily lives, and selling or sharing that information with companies that desired our attention. Whether these companies sought to influence our consumer choices or, like CA, impact our political behavior, they found a treasure trove of information with such specificity that it could reveal a user’s race, religion, wealth, partisanship, and physical and mental health status, among many other sensitive personal characteristics.

Feeling the weight of public scrutiny in the wake of this scandal, many of the platforms quickly moved to restrict access to what were perhaps the most generous and least scrutinized sources of digital data: their Application Programming Interfaces (APIs). The APIs allowed anyone with a few programming skills to gather massive volumes of data about a given platform’s users and content. And this included academics. From anthropology to psychology, economics to health science, scholars from a wide variety of disciplines relied on these APIs to gather large amounts of data for research into the content and behaviors found in digital spaces, and the CA-inspired restrictions significantly undermined multiple lines of research (Freelon, 2018; Hemsley, 2019).

Not surprisingly, digital researchers have reacted to this new “post-API age” (Freelon, 2018), with a mixture of frustration and concern. Frustration because the platforms took an incredibly broad approach—altering, restricting, or shutting down APIs altogether—without considering the impacts on scholarly inquiry. Concern because of the damage this has done to scientific knowledge. Digital research provides...
important insights about the social, cultural, economic, and political phenomena that impact people's everyday lives. Digital research also helps hold the platforms to account, spotting and diagnosing problems perpetuated by the tech companies themselves. In short, vital work has been devastated by the platforms’ responses to the CA scandal.

Or at least, this has been the common refrain. In this essay, I argue that though certain means of data access have indeed changed since 2018, the basic relationship between researchers, the platforms, and digital data remains largely the same. The platforms and their APIs have always been proprietary black boxes, never intended for scholarly use. And even when researchers could mine these data spigots seemingly endlessly, we rarely knew what type or quality of data we were analyzing. Moreover, the largesse of the API era allowed many digital researchers to conduct their work with little regard for the rigor, ethics, or larger societal value we should expect from scholarly inquiry. In other words, our digital research processes and output have not always occupied the high ground. Rather than viewing 2018 and CA as a profound disjuncture and loss, I suggest that digital researchers need to take a more critical look at how our community collected and analyzed data when it still seemed so plentiful, and use these reflections to inform our approaches going forward.

Throughout the essay, I draw on my experiences working (and occasionally clashing) with the platforms to draw attention to scholarly concerns and to secure data for academic research. I was a member of the Social Science One (SS1) Commission’s European Advisory Group between June 2018 and October 2020. SS1 (n.d.) seeks to build academic–platform partnerships that allow platform data sharing, while “ensuring the highest standards of privacy and data security.” Social Science One’s first partnership with Facebook has experienced a number of delays, and the release of the first data set came about almost a year-and-a-half behind schedule. Yet, the initiative is also providing a number of crucial insights into the barriers, both new and old, to conducting better, more responsible digital research, and it offers some promising avenues moving forward. I am also the lead investigator on an academic research project selected by Twitter to independently assess the “health of conversations” on the platform. It took 18 months for our team to receive data, and what we did receive was far less than originally promised. Our team has been struggling across the academic–industry divide, and we have learned a great deal about why this is so difficult, as well as what might help researchers and the platforms bridge their differences. Ultimately, in writing this essay, I hope that the lessons from each of these endeavors, coupled with research I have conducted into digital data quality, will prove instructive to a wide array of researchers.

I develop my argument in several steps. First, I tackle the claim that API restrictions have unfairly swept up academic work—the sense that we are being punished for something we did not do. I unpack the relationship between the platforms and academic research before examining the ethics of digital research in both the “Data Golden Age” and the post-API era. Next, I turn a critical eye to the argument that digital research is of great value and significance. I suggest that digital research has not always been adequately rigorous, has failed to examine and acknowledge the limitations of digital data, and has too often been motivated by expedience, rather than societal value. The essay concludes with a number of suggestions for moving digital research forward in the post-API age.

**How Could the Platforms Do this to Us?**

Much of the frustration expressed by digital researchers following the CA scandal has contained a healthy dose of indignation, with some suggesting that the platforms have gained an advantage from the scandal—that they have found a convenient excuse for keeping data out of the hands of those who could otherwise hold them accountable (Bruns, 2019). Others see less overtly hostile intentions from the platforms, but still plenty of shortsightedness. In their haste, the tech companies did not consider the larger consequences of shutting down the APIs, including their impact on important academic work (Hemsley, 2019).

There is clearly some truth to both sets of suspicions. These are large tech companies, facing competing interests and incentives across different internal units. My personal experiences dealing with Facebook and Twitter have shown that while certain executives wish to share data with academics—to truly shed light on the good, the bad, and the ugly—others are much more cynical. The latter tend to regard data sharing as a “damned if we do; damned if we don’t” prospect. Perhaps most crucially, however, it is risk-averse corporate lawyers who tend to win these debates, placing limitations on data sharing not because they see academic research as threatening, but because they fear liability under regulatory schemes such as the European Union’s General Data Protection Regulation (GDPR), the California Consumer Privacy Act, or the United States Federal Trade Commission Act.

Yet, the very fact that academics are having these conversations with the platforms and their lawyers is a positive signal. Indeed, in many ways, CA—as well as myriad controversies concerning misinformation, bots, abuse and harassment, hate speech, and so on—have actually opened some actors’ eyes to the value of external academic research. Before these scandals, platforms rarely engaged with or supported outside, independent scholarship. The platforms’ top priorities were to scale and make profit, and their work was focused on engineering, design, and performance, not on understanding the larger social implications of the systems they were building. Some scholars conducted research for tech companies—as employees or contractors—but the results were rarely vetted by peer review or shared publicly. Occasionally, the platforms provided special data sets to external researchers (e.g., Vosoughi et al., 2018), and in a few instances, they collaborated directly with outside
scholars (e.g., Kramer et al., 2014). But the modus operandi was still more or less passive apathy toward independent research, especially in the social sciences and humanities (Tromble & McGregor, 2019).

And the platforms’ API design and implementation naturally fit this pattern. Academics may have been mining APIs at will, but we were not the APIs’ intended users. The APIs were designed for developers whose games and other apps would bring more users to the platforms. They were designed for corporations to monitor their customer bases, brand identity, and advertising. In short, they were designed for the platforms’ profit. Academics were free to go about their research using the APIs, but the platforms were not deeply concerned about the results. When the platforms began restricting their APIs, they were not considering the implications for academic research, primarily because they did not think of the APIs as tools for academic research. The new rules that they have laid out for gaining access to the APIs in the wake of CA make this particularly clear. Facebook, for instance, requires verification as a “business entity” for approval of apps seeking access to the Pages API, which was long popular with academic researchers (Hemsley, 2019).

However, real conversations are now occurring between the platforms and independent academics, and many more executives seem to genuinely support efforts to bring more rigorous social scientific and humanistic approaches into the heart of their companies’ work, as well as to find ways to support outside research. Twitter has been enhancing its academic outreach and support efforts and has hired staff dedicated to better understanding how academics use its data (Twitter, n.d.). Facebook has offered a number of funding opportunities for academics proposing research into WhatsApp, Instagram, and Facebook itself, and though those grants do not offer data, they are unrestricted gifts—meaning Facebook has no say over how researchers use the money and cannot control results or publications. In August 2020, Facebook also announced that it had launched an initiative in which independent external scholars were collaborating with internal Facebook researchers to study the platform’s impacts on the 2020 US presidential elections. The independent scholars are leading the research design, and all results from the project will be made public (Facebook, 2020).

This is not to suggest that the platforms have shared some perfect epiphany. Profit remains king. Internal incentives are still conflicting. And lawyers continue to have the final say. Our Twitter “healthy conversations” project was significantly delayed in large part because creating understanding across the tech-academic-legal divide is immensely difficult.

Simply understanding one another’s needs and priorities—even just one another’s jargon—is laborious. And yet the fact that we are having these conversations, we are trying to understand one another, represents progress. Facebook and Social Science One’s partnership is similarly behind schedule, generating a number of justifiable frustrations. Yet the initiative has produced a new state-of-the-art system that uses differential privacy to mitigate data abuses and is likely to serve as a model for future platform–academic data sharing endeavors (DeGregorio et al., 2019). Facebook deserves credit for investing significant resources into developing this system, and those within the company who continue to push back against the internal naysayers also deserve acknowledgment. There are true allies of academic research within these companies, and our work depends on identifying and supporting their efforts. Of course, this should not mean backing away from critique of the platforms and their policies. Critique should remain vociferous. But it does mean being realistic about the challenges ahead. It also means giving credit where credit is due.

**Looking in the Mirror**

At the same time, the digital research community also needs to take a closer look in the proverbial mirror. Indeed, we frequently seem to forget that the CA scandal was also an academic research scandal. Aleksandr Kogan, then a scholar at Cambridge University, developed and then shared the GSRApp that CA used to harvest Facebook users’—and friends of those users’—data. Although Kogan shared the GSRApp with CA under the guise of his private company, Global Science Research, he developed and implemented the app as an academic researcher and created his company specifically to work with CA. In July 2019, the Federal Trade Commission announced that it had reached a consent agreement with Kogan. To prevent charges for employing “deceptive tactics to harvest personal information from tens of millions of Facebook users for voter profiling and targeting” (Federal Trade Commission, 2019b), Kogan agreed to a lengthy list of reporting requirements and restrictions on his business, research, and data collection activities (Federal Trade Commission, 2019a).

It is tempting to think of Kogan as a solitary academic “bad apple.” But he is not. While we cannot know how many among us have used deceptive tactics to gather data, it is no secret that many in our community have long been focused, first and foremost, on amassing as much digital data as possible. This is particularly true in my own computational research sub-community, where far too frequently data acquisition, rather than theory—or even a research question—leads the research endeavor. This tendency to foreground and, in some cases, fetishize the data themselves (Mosco, 2016, pp. 205–206) has led researchers to adopt a number of questionable practices.

To be sure, even in the midst of the “Data Golden Age,” scholars were exploiting bugs in platforms’ code and breaking terms of service to gather far more data than was technically permissible. Many engaged—and continue to engage—in these practices with a shrug. Who cares if we violate the terms laid out by these mega corporations? We are not obliged to protect their business model. Our responsibility is to the research, to knowledge, and to science. Yet all too often, foregrounding the data has led us to neglect the users behind the data. Numerous studies have collected and...
analyzed platform data without users’ informed consent (e.g., Catanese et al., 2011; Gjoka et al., 2010; Kramer et al., 2014; Lewis et al., 2008; Traud et al., 2012). And a variety of popular academic Facebook apps, such as NameGenWeb (Hogan, n.d.) and Netvizz (Rieder, 2013), allowed researchers to gather data from users’ friends without the friends’ knowledge. This friends-of-friends functionality was built into the Facebook API at the time, meaning that scholars were not violating Facebook’s terms of service. But such compliance does not absolve researchers of the ethical responsibilities to those whose data they extract, analyze, and in many cases, share with others.

This latter point is particularly important. Kogan’s clearest and most egregious ethical violation occurred when he decided to share the GSRApp and data with CA. Direct, for-profit data sharing with private companies may not be common among academics, but with open data practices on the rise, and particularly widespread among computational researchers, many of our data sets are being used by non-academics. Yet we lack guidelines for considering the risks inherent in releasing our data publicly (Zimmer, 2010). Of course, we also share data with other academics—for replication, to further collaborations, and to help address resource inequalities. These are all admirable pursuits. Yet we also lack standards and concrete mechanisms for evaluating researcher integrity. How do we know whom to trust? How do we keep track of whom is using data and how they are doing so? How do we more quickly and effectively identify abuses? How do we stop those abuses?

If we are going to continue demanding access to platform data, we must have clearer answers to such questions. Current accountability mechanisms such as peer review and institutional review board (IRB) approval are inadequate. IRBs are notoriously ill-equipped to assess the implications of digital research practices (Bloss et al., 2016; McKee & Porter, 2008; Torous & Roberts, 2018), and volumes of questionable studies have made it through peer review. In some cases, we will need to establish practices for research monitoring and audits—for example, using keystroke logging and analytical “privacy budgets” (McSherry, 2009). We will also need to develop standard protocols for encryption and secure storage of digital data, as well as clear guidelines about when and how to share data.

And these guidelines should include already public data. As Zimmer (2010) argues, data opened to the platforms’ “public square” are not analogous to the data we might gather while observing people in physical public spaces. People often post sensitive information such as their location, political views, and religious affiliation online without understanding the potential implications (Crawford & Finn, 2015; Fiesler & Proferes, 2018). Even when not offered publicly, network analysis and machine learning techniques may allow researchers to generate much more precise inferences about sensitive traits than would be possible based on mere observations in a physical space. What is more, public information that we typically do not consider sensitive can still be used for harm. For example, birthdate, family, work, and education information all create vulnerabilities to hacking and doxing.

The popularity of social media services such as Snapchat and Instagram Stories, built on the notion of ephemerality, also calls for more careful consideration. We are already beginning to see research that collects, stores, and analyzes data intended to disappear (e.g., Juhász & Hochmair, 2018), and digital researchers have barely grappled with the implications of the “right to be forgotten” for our data management, retention, and replication practices (Tromble & Stockmann, 2017). This also points to the need for clearer standards regarding data reproduction in presentations and publications. Images and text that seem relatively innocuous today could have serious consequences for people down the road. Platform users might see no harm in sharing them publicly now. Certain social media influencers or political activists might even revel in knowing that scholars are sharing their posts more widely. Yet these posts are likely to carry more permanence in scholarly publications than they do on the platforms themselves. If an influencer later regrets party pictures or the activist begins facing harassment for provocative political statements, they can delete the original posts. They cannot delete our books and journal articles. We therefore need to consider guidelines for determining when reproduction is appropriate, as well as when and how to remove content from our presentations (e.g., publicly available video or slides) and publications.

These are all difficult issues, especially as core concepts such as “privacy” remain contested (Bloustein, 1964; Fuchs, 2011; Nissenbaum, 2010). And guidelines on many of these matters have been suggested before, including the Association of Internet Researchers’ extremely useful list of ethics questions (Markham & Buchanan, 2012) and its “Ethical Guidelines 3.0” (franzke et al., 2020). However, unless and until ethical considerations are taken more seriously by the entire academic digital research community, with more robust guidelines and mechanisms for accountability put into place, our clarion calls for evermore data will rightfully be met with skepticism.

Our Work Is so Important!

Our cries for more data also require that we demonstrate the value of our work. Scholars of the digital are particularly well-positioned to analyze and unpack some of the most pressing social, cultural, economic, and political questions of our time. However, our work must actually live up to this potential. And crucially, we must be able to convincingly communicate this potential to tech companies, policymakers, and the public alike.

We did not have to make this case in the era of data abundance—or at least, we did not have to do so as often, nor as compellingly. Data largesse allowed digital researchers to produce a plethora of small-scale, narrowly focused studies.
It complemented and fed into insular academic “publish or perish” incentive structures. With data readily available, researchers could undertake narrowly tailored studies, adding another small case or slightly shifting research questions with each additional publication. In computational fields, researchers could focus on making moderate improvements to algorithmic performance with each new data set. We needed to justify our work to one another via peer review, but because we could so readily scrape digital data from a variety of websites and platforms, we rarely had to convince others of the profound value of our research. The data were there, and we used them as our academic needs demanded. This is not to say that all of the work undertaken in the time of API largesse was shallow, nor even that more narrowly focused studies are unimportant. However, for many—and particularly for early-career researchers who face the greatest pressure to publish or perish—convenience and speed frequently win out over societal significance.

As we cry foul about new data restrictions and issue pleas for better access, we need to be much more honest with ourselves about this. In the wake of CA, each new request for data will rightfully be met with a long list of privacy and data use concerns, and we must be able to offer credible justifications for opening the public—and even the platforms—to these risks.

The “Golden Age” that Never Was

We also need to be much more forthright about the analytical limitations found in much of the work we produced in the “Data Golden Age.” Today’s calls for data access come tinged with a sense of nostalgia, but this nostalgia masks significant problems with the data, both then and now.

In an ongoing research project examining digital data quality, my colleagues and I unpack the limitations of Twitter data in particular (Tromble et al., 2017; Tromble & Stockmann, 2017). Twitter is the most (over-)studied social media platform precisely because it offers relatively open data access. Its public Search API allows researchers to gather tweets posted up to 7 days earlier, while the public Streaming API permits capture of tweets in real time. Following the CA scandal, Twitter now requires scholars to undergo review for API access, and the company only allows each researcher use of one app to query the APIs. Otherwise, however, the APIs return much the same data they did before, and because they are free to use, academics can gather large amounts of Twitter data no matter their financial resources.

Yet, the non-randomness of data captured via these APIs means that, even in the best of times, many Twitter studies have drawn conclusions based on substantially biased inferences. Neither of the public APIs guarantees one will capture all tweets matching a query’s parameters. In fact, Twitter’s developer documentation makes it clear that the Search API will not return all tweets,2 and the Streaming API throttles captures when one’s query parameters match more than 1% of the total volume of tweets produced globally at any given moment in time.3 By comparing data collected using identical keyword queries to the free Search and Streaming APIs with the full population of tweets purchased in real time (at substantial cost) from Twitter’s PowerTrack API, our research shows that conclusions based on data from the public APIs are likely to be biased. This is particularly true for analysis of tweet content or interactions between Twitter users, as tweets with hashtags are over-represented in both Streaming and Search APIs, while user mentions are over-represented in Search API results (Tromble et al., 2017).

Researchers with substantial resources could simply purchase all tweets in real time, but this is too cost-prohibitive for most researchers. It is also possible, and typically less expensive, to purchase data from Twitter’s historical archive, which covers all tweets generated since the company was founded. However, the archive does not include tweets that have been set to private, nor any deleted from the platform. This means that the longer one waits to capture tweets, the fewer will be available (though note that because of temporary privacy settings or short-term account suspensions, some do reappear over time). And because tweets do not disappear randomly, we are likely to draw biased conclusions from these historical data (Tromble & Stockmann, 2017).

To illustrate just how serious this problem can be, let me offer a brief example. Scholars interested in information flows, virality, and discursive patterns on social media frequently use social network analysis techniques to examine clusters of interaction and identify central actors or concepts within a broader discourse or event (e.g., Isa & Himelboim, 2018; R. Wang & Chu, 2019; Xiong et al., 2019). One might, for instance, investigate the network generated when Twitter users @-mention one another, using this to identify clusters of interaction and central, influential actors. Beginning with a data set captured in real time during Donald Trump’s first address to a Joint Session of Congress in February 2017 (and captured using the PowerTrack API), I analyzed whether and how waiting three, six, nine, and twelve months to capture such data4 would impact two network centrality scores—betweenness and eigenvector centrality—that are often used to identify influential actors. Figure 1 compares lists of the top 10, 25, 50, and 100 users based on these centrality scores. The initial real-time data set from February 2017 serves as baseline, and I use Kendall’s tau-b to assess how closely these results correlate with the lists generated as tweets disappear (and reappear) over time. Following D. J. Wang et al. (2012), I presume that any bias introduced by levels of error greater than 0.05 (i.e., a correlation of 0.95 or less) is likely to be non-trivial and provide the dashed red lines in Figure 1 to represent this benchmark. The lower the correlations fall below these lines, the greater the concern about bias. While the betweenness centrality scores for the top 10 hashtags show relatively low error throughout, all other observations fall well below the target. By the 6-month mark, correlations for betweenness centrality among the top 25 to 100 user mentions range between 0.5167 and 0.6052. And the results
for eigenvector centrality are remarkably poor throughout, with the list of top 10 users generating negative correlations. Whether one were using these metrics in quantitative research or to dig into information flows and other dynamics using primarily qualitative approaches, the timing of data capture could dramatically impact one’s results.

The timing of data capture mattered in the “Data Golden Age,” and it continues to matter in the post-API era. Yet we rarely talk about these issues. We frequently treat digital data as if, once generated, they are permanent and invariant, and we fail to acknowledge the potential consequences of data loss. Sometimes we even fail to reveal that data were captured (well) after the event in question. We similarly lack open, frank discussions about the ways in which the APIs themselves dictate data quality. We have instead often treated them as “neutral tools” from which to gather similarly “neutral” data (Bucher, 2013). Part of the problem, of course, is that the APIs are, and always have been, proprietary black boxes. The platforms are under no obligation to reveal whether, why, and how some data are made available via an API while other data are not. However, as we express frustration about new API restrictions—as we lament that we cannot get the data we once could—we would do well to think more critically about the quality of the data we actually gathered.

The Post-API Era—Moving Forward

In the post-API era, simply calling for what we once had is both unrealistic and unwise. We do not want the data the APIs once provided. We want better. Rather than focusing on getting more data from the platforms, we must focus on getting high-quality data. We must also focus on doing better, more ethical work with those data. Instead of foregrounding and fetishizing the data themselves, socially significant questions should serve as our starting point. And we must recognize the responsibilities we carry when working with digital data, particularly to the people represented in those data.

Figure 1. Kendall’s tau-b correlations of top hashtags in co-occurrence networks, as tweets disappear over time.
With these priorities in mind, I suggest two broad strategies for pursuing further data access.

First, rather than approaching platforms one researcher (or research team) at a time, we need a more coordinated strategy for pressing this issue forward, both with the platforms themselves and with policymakers. We should therefore work within our professional associations to develop tactics and approaches for outreach. Social Science One involves dozens of researchers from around the world, but most are political communication scholars who employ computational methods (myself included). This has led to a rather narrow focus on obtaining very large data sets, as well as access to specific APIs (including the Facebook Pages API). But are these necessarily the best types of data? Would it be better to focus on securing regular access to carefully curated, smaller data sets? This is of course a matter of debate. But it is a debate sorely needed within our community. And, as part of that debate, we need to think creatively about what types of data would simultaneously serve broad constituencies of researchers and the public at large. Only then can we approach the platforms and policymakers with our most compelling, most socially responsible requests.

Second, we should work, again within our associations, to improve guidelines for ethical research, as well as to identify stronger mechanisms for bolstering their adoption. The Association of Internet Researchers’ guidelines (franzke et al., 2020) offer a particularly good starting point. Unfortunately, however, these guidelines are not especially well-known, let alone followed, across the wide variety of disciplines and institutions undertaking digital research. We also need to revisit these guidelines, and other relevant frameworks, with the data access question at the forefront of our minds. If we hope to convince platforms, policymakers, and the public that further academic data access is warranted, we will have to offer concrete proposals for specific protocols and safeguards. (It will certainly be better if our community identifies and proposes such safeguards, rather than having them dictated to us by the platforms.) We also need to think more carefully about how we manage data risks that vary over time. Although the likelihood of harm coming to the subjects of our research may seem very low when data are first collected, analyzed, or shared, we need much better mechanisms in place for reacting when those risks change.

API restrictions have had a major impact on scholarly inquiry; there can be no doubt. However, rather than viewing this as an overwhelming and unmitigated loss, our research community can take advantage of this moment for critical reflection and improvement. To take full advantage of this opportunity, however, we must be open, honest, and acknowledge our mistakes. Too much of our work has involved questionable ethics; been driven by data and experience, rather than larger societal value; and overlooked critical limitations in the data we so eagerly amassed. We must do better moving forward.

Declaration of Conflicting Interests
The author(s) declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: The author has received a grant funded by Twitter and a gift provided by Facebook, both to support independent research. Neither company has had any input in this piece, and funding did not support any of the research found herein.

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

ORCID iD
Rebekah Tromble https://orcid.org/0000-0003-3875-2729

Notes
1. The Pages API is just one of several APIs for which Facebook requires “business verification.” See https://developers.facebook.com/docs/apps/review/business-verification for this list and further explanation of its policy.
2. See https://developer.twitter.com/en/docs/tweets/search/overview#standard.html.
3. See https://developer.twitter.com/en/docs/tweets/filter-real-time/guides/connecting.html.
4. I used Twitter’s REST API for this task, which permits the retrieval of specific tweets based on their unique IDs.

References
Bloss, C., Nebeker, C., Bietz, M., Bae, D., Bigby, B., Devereaux, M., Fowler, J., Waldo, A., Weinbel, N., Patrick, K., Klemmer, S., & Melichar, L. (2016). Reimagining human research protections for 21st century science. Journal of Medical Internet Research, 18(12), Article e329.
Bloustein, E. (1964). Privacy as an aspect of human dignity: An answer to Dean Prosser. New York University Law Review, 39, 962–1007.
Burns, A. (2019). After the “APIcalypse”: Social media platforms and their fight against critical scholarly research. Information, Communication & Society, 22(11), 1544–1566.
Bucher, T. (2013). Objects of intense feeling: The case of the Twitter APIs. Computational Culture. http://computationalculture.net/objects-of-intense-feeling-the-case-of-the-twitter-api/#:~:text= Whilst%20contributing%20one%20such%20glimpse, forms%20of%20contestation%20and%20identification%20%20C
Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. The Guardian. https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election
Catanesi, S. A., De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2011). Crawling Facebook for social network analysis purposes. In Proceedings of the international conference on web intelligence, mining and semantics (p. 52). Association for Computing Machinery. http://www.cogprints.org/7663/1/71-catanesi.pdf
Crawford, K., & Finn, M. (2015). The limits of crisis data: Analytical and ethical challenges of using social and mobile data to understand disasters. GeoJournal, 80(4), 491–502.
DeGregorio, C., Hillenbrand, B., Li, D., Messing, S., & Nayak, C. (2019). Preserving privacy while fostering meaningful research on elections and democracy. Facebook Research. https://research.fb.com/blog/2019/06/preserving-privacy-while-fostering-meaningful-research-on-elections-and-democracy/

Facebook. (2020). New Facebook and Instagram research initiative to look at US 2020 presidential election. https://about.fb.com/news/2020/08/researchimpact-of-facebook-and-instagram-on-us-election/

Federal Trade Commission. (2019a). In the matter of Cambridge Analytica, LLC: Case summary. https://www.ftc.gov/enforcement/cases-proceedings/182-3107/cambridge-analytica-llc-matter

Federal Trade Commission. (2019b). In the matter of Aleksandr Kogan, an individual, and Alexander Nix, individually and as Chief Executive Officer of Cambridge Analytica, LLC: Agreement containing consent order as to respondent Aleksandr Kogan. https://www.ftc.gov/system/files/documents/cases/182_3106_acco_kogan.pdf

Fiesler, C., & Proferes, N. (2018). “Participant” perceptions of Twitter research ethics. Social Media+ Society, 4(1), Article 763366.

franzke, a.s., Bechmann, A., Zimmer, M., Ess, C. & The Association of Internet Researchers. (2020). Internet research: Ethical guidelines 3.0. https://aoir.org/reports/ethics3.pdf

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

Fuchs, C. (2011). An alternative view of privacy on Facebook. Information, 2(1), 140–165.

Gjoka, M., Kurant, M., Butts, C. T., & Markopoulou, A. (2010, March). Walking in Facebook: A case study of unbiased sampling of OSNs [Conference session]. 2010 Proceedings IEEE INFOCOM, San Diego, CA, United States.

Hemsley, J. (2019). Social media giants are restricting research vital to journalism. Columbia Journalism Review. https://www.cjr.org/tow_center/facebook-twitter-api-restrictions.php?bclid=IvAR1uULkcGqOrQYRsagYkJtISaKcIKGK5t2x_Q5SnoOd38CGsO2NdovULdpns

Hogan, B. (n.d.). Software. https://blogs.oii.ox.ac.uk/hogan/software/

Iba, D., & Himelboim, I. (2018). A social networks approach to online social movement: Social mediators and mediated content in #freejastaff Twitter network. Social Media+ Society, 4(1), Article 8768087.

Juhász, L., & Hochmair, H. (2018, June 12). Analyzing the spatial and temporal dynamics of Snapchat [Conference session]. Analysis, Integration, Vision, Engagement (VGI-ALIVE) Workshop, Lund, Sweden.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. Proceedings of the National Academy of Sciences of the United States of America, 111(24), 8788–8790.

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.

Markham, A., & Buchanan, E. (2012). Ethical decision-making and internet research: Recommendations from the AoIR Ethics Working Committee (Version 2.0). Association of Internet Researchers. http://aoir.org/reports/ethics2.pdf

McKee, H., & Porter, J. E. (2008). The ethics of digital writing research: A rhetorical approach. College Composition and Communication, 59(4), 711–749.

McSherry, F. D. (2009, June). Privacy integrated queries: An extensible platform for privacy-preserving data analysis. In Proceedings of the 2009 ACM SIGMOD international conference on management of data (pp. 19–30). Association for Computing Machinery.

Mosco, V. (2016). To the cloud: Big data in a turbulent world. Routledge.

Nissenbaum, H. (2010). Privacy in context: Technology, policy, and the integrity of social life. Stanford University Press.

Rieder, B. (2013). Studying Facebook via data extraction: The Netvizz application. In Proceedings of the 5th annual ACM web science conference (pp. 346–355). Association for Computing Machinery.

Social Science One. (n.d.). Social Science One: Building industry-academic partnerships. https://socialscienceone.org/

Torous, J., & Roberts, L. W. (2018). Assessment of risk associated with digital and smartphone health research: A new challenge for institutional review boards. Journal of Technology in Behavioral Science, 3(3), 165–169.

Traud, A. L., Mucha, P. J., & Porter, M. A. (2012). Social structure of Facebook networks. Physica A: Statistical Mechanics and Its Applications, 391(16), 4165–4180.

Tromble, R., & McGregor, S. (2019). You break it, you buy it: The naiveté of social engineering in tech – and how to fix it. Political Communication, 36(2), 324–332.

Tromble, R., & Stockmann, D. (2017). Lost umbrellas: The “right to be forgotten” in Internet research. In M. Zimmer & K. Kinder-Kurlanda (Eds.), Internet research for the social age: New cases and challenges (pp. 75–91). Peter Lang.

Tromble, R., Storz, A., & Stockmann, D. (2017). We don’t know what we don’t know: When and how the use of Twitter’s public APIs biases scientific inference. Social Science Research Network. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3079927

Torous, (n.d.). Academic research: Data on everything and anything, at your fingertips. https://developer.twitter.com/en/solutions/academic-research

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. Science, 359(6380), 1146–1151.

Wang, D. J., Xiaolin, S., McFarland, D. A., & Leskovec, J. (2012). Measurement error in network data: A re-classification. Social Networks, 34(4), 396–409.

Wang, R., & Chu, K. H. (2019). Networked publics and the organizing of collective action on Twitter: Examining the #Freebassel campaign. Convergence, 25(3), 393–408.

Xiong, Y., Cho, M., & Boatwright, B. (2019). Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of Twitter during the# MeToo movement. Public Relations Review, 45(1), 10–23.

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

Author Biography

Rebekah Tromble is director of the Institute for Data, Democracy & Politics and Associate Professor in the School of Media & Public Affairs at The George Washington University. Her research focuses on political communication, digital research methodology, and research ethics, with particular interests in political discourse on social media, as well as the individual and social impacts of online disinformation and abusive content. She is currently leading a team of international researchers from both the social and computer sciences investigating the “health” of political conversations on Twitter.