

Research Tool Report

Twitter as research data *Tools, costs, skill sets, and lessons learned*

Kaiping Chen , University of Wisconsin–Madison

Zening Duan, University of Wisconsin–Madison

Sijia Yang, University of Wisconsin–Madison

ABSTRACT. Scholars increasingly use Twitter data to study the life sciences and politics. However, Twitter data collection tools often pose challenges for scholars who are unfamiliar with their operation. Equally important, although many tools indicate that they offer representative samples of the full Twitter archive, little is known about whether the samples are indeed representative of the targeted population of tweets. This article evaluates such tools in terms of costs, training, and data quality as a means to introduce Twitter data as a research tool. Further, using an analysis of COVID-19 and moral foundations theory as an example, we compared the distributions of moral discussions from two commonly used tools for accessing Twitter data (Twitter’s standard APIs and third-party access) to the ground truth, the Twitter full archive. Our results highlight the importance of assessing the comparability of data sources to improve confidence in findings based on Twitter data. We also review the major new features of Twitter’s API version 2.

Key words: Twitter, data collection tools, skill sets, cost, data quality evaluation, computational social science

Social media is a major platform that people use to consume, share, and discuss information about scientific innovations and controversies (Brossard & Scheufele, 2013). Around one in five U.S. adults use Twitter, and 42% of those use it to discuss politics (Hughes & Wojcik, 2019). Twitter has become a popular research site in recent years for studying issues at the intersection of the (life) sciences, politics, and policymaking. Many interesting problems have been investigated using Twitter data. For example, scholars have examined Twitter discourses on contentious science topics such as genetic modification technology (Singh et al., 2020; Wang & Guo, 2018) and the Zika virus (Wirz et al., 2018).

Beyond providing a lens into public discourse, Twitter data allow researchers to investigate how such

discourse is affected by political ideology, the rural-urban divide, and education levels across regions in the United States (Wirz et al., 2021). Twitter has also been used to study public health issues such as allergies, obesity, and insomnia; to track illness and risk behavior (Paul & Dredze, 2011); to monitor public perceptions of the H1NA pandemic (Chew & Eysenbach, 2010); and to communicate public health information (Vance et al., 2009). In addition, Twitter has emerged as a battleground for politicians to promote agendas on science controversies and blame other political entities for public health issues. Scholars have shown that Twitter is an active platform that nongovernmental organizations and politicians use to shape public perceptions about climate change (Fownes et al., 2018). During the COVID-19 pandemic, scholars found that the “China virus” stigma was perpetuated through Twitter (Budhwani & Sun, 2020). These important scholarly works demonstrate the promise of using Twitter data for biopolitical research.

doi: [10.1017/pls.2021.19](https://doi.org/10.1017/pls.2021.19)

Correspondence: Kaiping Chen, Life Sciences Communication, University of Wisconsin, Madison, WI. Email: kchen67@wisc.edu

For biopolitical research, Twitter data are attractive because they are particularly multilevel and multifaceted. The most common data type is tweet content. Plain tweet text provides researchers with rich information, such as the distribution of sentiments and topics (Cody et al., 2015; Haunschild et al., 2019; Walter et al., 2020), personality traits or other individual characteristics (Whittingham et al., 2020), and information-sharing behaviors (Singh et al., 2020). URLs (uniform resource locators) and mentions/retweets embedded in tweets are also essential to a variety of research questions, such as which external articles or websites people prefer to share, how people connect virtually (Ke et al., 2017), and how false news spreads on social media (Vosoughi et al., 2018). Analyzing Twitter images and videos might require different computational skills than analyzing plain tweets. Still, it could provide additional information for researchers, some critical for understanding how media biases are manifest in visual portrayals of politicians (Peng, 2018). A relatively small number of biopolitical studies have focused on visualization cues on Twitter, opening yet another promising avenue for future exploration (Peng, 2018).

This article explicates how researchers can get started using Twitter for research by discussing (1) the main ways to access and collect Twitter data, (2) the financial costs and skill sets needed, and (3) data quality issues. It also includes an empirical demonstration to assess data quality from different tools that researchers can use to collect Twitter data. Finally, this article alerts researchers to recent developments in Twitter's move to API version 2 (V2), which includes a dedicated Academic Research track. Overall, this article demonstrates the utility of Twitter data as a resource for research but also notes important considerations about Twitter as research data.

Tools, costs, and skill sets

This section first introduces two major ways for researchers to get started with Twitter data collection. These methods can satisfy most research needs, such as estimating the prevalence, temporal trajectory, or geographic distribution of content features (e.g., topics or sentiments). We then review the financial resources and skill sets needed when using these data collection methods. We end this section by introducing Twitter full-archive access, which is crucial to answering certain

research questions but can be inaccessible to some researchers.

Tools: Major ways to access Twitter data

Twitter's standard APIs. Twitter's standard application programming interfaces (APIs) are the most common entry points to Twitter data. Twitter provides a series of publicly available APIs that offer free but restricted access to data. The Streaming API and the Search API are two of them. Drawing from Twitter's official guide¹ and other research (Driscoll & Walker, 2014; Kim et al., 2020), we summarize the main features of and differences between these two APIs for research.

The *Streaming API* is designed to return tweets in a real-time stream. Two options are currently available: *filtered stream* and *sampled stream*. The filtered stream enables researchers to screen millions of new tweets in any given second and extract only those matching a specific set of filter rules (e.g., up to 400 key words, 5,000 user identities [IDs], and 25 locations). By default, the filtered stream returns 1% of tweets per hour. If the researcher wants to use the filtered Streaming API to collect all tweets with the hashtag #TwitterAPI, for example, and there are fewer matched tweets than the allowed cap, the researcher will obtain all tweets containing that hashtag; otherwise, the researcher will receive a sample of the #TwitterAPI tweets.

In comparison, the sampled stream is designed to return a likely random selection of all newly posted tweets (Pfeffer et al., 2018) in real time that are free from filtering constraints. Researchers who have specific topics to collect are better served using the filtered stream. When researchers do not have a specific research topic in mind and are interested in taking the temperature of all conversations occurring on Twitter for general monitoring purposes, the sampled stream is a better tool to use.

The *Search API* is another popular entry point for accessing Twitter data. It was designed to return historical tweets collected by matching rules defined by users. The standard Search API enables researchers to access at no cost a sample of tweets published in the past seven days. For researchers who only need the past seven days of Twitter data or who are willing to collect tweets every seven days to trace historical data, the standard Search API may be a good choice. If the research project needs

¹For more information, see <https://developer.twitter.com/en/docs/twitter-api>.

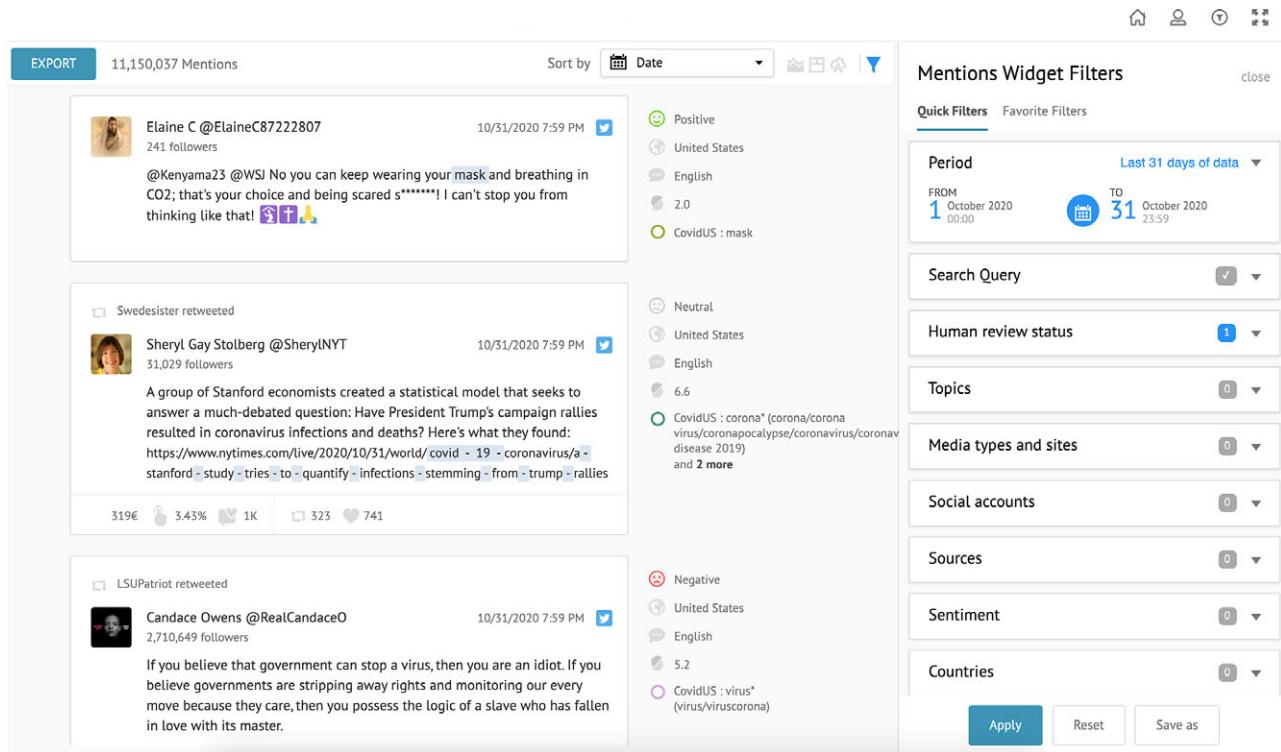


Figure 1. Example of a dashboard layout of a third-party platform.

historical tweets posted over a longer period and/or beyond the free-tier sampling rate, one would need to use paid premium plans of Twitter's Search API. As discussed later, two premium-level Search API products allow backtracking of tweets for a longer period (e.g., past 30 days) or tracing historical data back as far as 2006.

Third-party platforms. Researchers interested in other more user-friendly but costly access points can use third-party platforms. The market for social media data vendors has been growing rapidly, offering an alternative way for researchers to access social media data, including Twitter (see Figure 1). Examples such as Salesforce Social Studio, Crimson Hexagon, and Synthetis differ considerably in their pricing structure, data access type and cap, dashboard and interface, availability of historic search, and data output. A typical Twitter data output from these third-party platforms (see Figure 2) consists of several data fields, such as time stamp, tweet URLs, publisher names and handles, tweet content, location, language, and even a basic sentiment score calculated by the platform.

Compared with Twitter's APIs, these third-party platforms provide a user-friendly dashboard to organize

search term setup, data filtering and download, and summary reports, hence requiring a less steep learning curve, especially for those unfamiliar with the API language. Some of these platforms also provide access to other popular social media sites, including public accounts on Facebook, Instagram, and YouTube, as well as online news websites and discussion forums. Typically, researchers can filter on language, time period, location, and source, although the availability of such parameters varies from platform to platform. Importantly, it is rare for third-party platforms to offer unlimited access. Some may place a cap on the maximum return of mentions per query without limiting the total daily queries allowed to run (e.g., 50,000 mentions per search), others may impose a monthly cap (e.g., three million), and still others may restrict how far back in time researchers can search (e.g., one to three years). These varying restrictions impact data quality and should be scrutinized before subscribing.

Shared tweet IDs. Using tweet IDs shared by other researchers is another way to collect Twitter data besides using APIs or purchasing third-party services. Researchers sometimes share metadata in the spirit of open science (Dienlin et al., 2021) or for other purposes.

Twitter as research data: Tools, costs, skill sets, and lessons learned

Mention URL	Publisher Name	Publisher Username	Mention Title
er.com http://twitter.com/SvConsulting1/status/1278251748587683840#1078310699523325952	Sv Consulting	SvConsulting1	The latest Sv Consulting : Digital Leadership & Inform
er.com http://twitter.com/woofka/status/127825089057552993#16273834_1278156387936112642	woofka #BlackLivesMatter	woofka	RT @morethanmySLE: REMINDER: Dan Patrick is ALS
er.com http://twitter.com/etkwebster/status/1278241039332805124102543826866698752	Verminous Scot Liz	etkwebster	Scotland has effectively been dragged out of EU again
er.com http://twitter.com/Ronet98063285/status/1278242617281708032#1003670790972289024_1277966987432308737	Ronet	Ronet98063285	RT @aileenwthenews: FIU has had 43 students and 25
er.com http://twitter.com/mivinjivine/status/1278250894182645767#115108245081243648_1278235035715461125	Jave #JoinAnakbayan	junvel_tlm	GINO KO RT @ABCSCBNNews: Japan OKs 50-billion
er.com http://twitter.com/ApertureImage/status/1278242034122469383#14830798_1278230408688186496	Chris (#3.5%)	ApertureImage	RT @Matthew82069336: Alok Sharma, the human sree
er.com http://twitter.com/lameloy/status/1278243466556948480#20759233_127883107077693440	Laure Meloy	lameloy	RT @TheBrookSoton: @PetrocTrelawny @BorisJohnson
er.com http://twitter.com/omahashila/status/1278240571191169024#797298875291799552_1278049152514101248	OmahaShila	omahashila	RT @Drottliemom: @donnie_maga @RedPilledisBACK
er.com http://twitter.com/HarrisonChris38/status/127825009727607808620196357_127773070498996497160	Christine Harrison	HarrisonChris38	RT @suwillkinson13: So disappointed to read that son
er.com http://twitter.com/OfirMe/status/12782464834600907008#13179354	Ofir.Me	ofirme	Cartoon: Covid-19 Fast and Furious. See also: The pre
er.com Deleted or protected mention	Deleted or protected mention	Deleted or protected mention	Deleted or protected mention
er.com http://twitter.com/jaybafna95/status/1278240839634829313#2879292354_1278235840041320448	Jay Bafna	jaybafna95	RT @ANI: Maharashtra: Section-144 imposed in Mumbai
er.com http://twitter.com/RsaCovid/status/1278251601526960128#12506237265975538474	RSA COVID	RsaCovid	Oudthoom municipality loses head of legal services to
er.com http://twitter.com/rajaredde2345/status/1278249903701938178#4147359792_1278230112140394497	Rajasekhar Reddy	ursysraja	RT @purijagan: While the world is battling with corona
er.com http://twitter.com/VintiquesMark/status/1278242164796018688#30990878_1278236321602043904	Mark Stacey	VintiquesMark	RT @Suewilson10: Johnson fails on pledge to have 24-
er.com http://twitter.com/Kay_Lee23/status/1278238225144975362#140135009_1278178559039664128	Krylic	Kay_Lee23	RT @CNN: "You're thinking it's an over-hyped flu and I'
er.com http://twitter.com/wowlay/status/127823735161547776#719063869419728896_127808777689206784	wowlay	wowlay	RT @Bibaybravo: His concert is on my bucketlist and v
er.com Deleted or protected mention	Deleted or protected mention	Deleted or protected mention	Deleted or protected mention
er.com http://twitter.com/shubhamg18/status/1278250989749800961#177546392_1278248097081028609	Shubham Gupta	shubhamg18	RT @jeabe: #Gurugram #CovidRecovered Need #Bio
er.com http://twitter.com/JustinPB4140091/status/1278251776140021760#125512539699597314_1278249326368763904	Justin perkins	justinpB4140091	RT @I Love GlosUK: County health chief says spike of c
er.com http://twitter.com/AnuBomb/status/12782432734854796288#73416807#1278240080944140289	WONDER WOMAnu	AnuBomb	RT @AJEnglish: Thread. There is 500 times more mico
er.com http://twitter.com/DepixolDave/status/127824214833752928#1144937920526200832	David Peat	d6a6v6e	Ordered two nice soap bars from Net-A-Porter, to give
er.com http://twitter.com/renxmh/status/1278248447028652016#1026413680441716737_1277747133484011521	akaashi	renxmh	what if china's making viruses to easily take over the w
er.com http://twitter.com/SDSNMed/status/12782511254525501468883276673921113088_1278247626421436416	SDSN Mediterranean	SDSNMed	RT @primalitaly: 🌐SAVE THE DATE: July15,17,30 CET,
er.com http://twitter.com/AtulDizipro/status/1278244788847736834#802142784_127824321085940948	Atul Verma	AtulDizipro	What is this? Yesterday #AarogyasetuApp went down
er.com Deleted or protected mention	Deleted or protected mention	Deleted or protected mention	Deleted or protected mention
er.com http://twitter.com/alexsteacy/status/1278238100196519937#779529665	Alex Steacy	alexsteacy	Overthrow your government immediately. https://www .
er.com http://twitter.com/TDTNarsarai/status/19789470724#nR@11LRQ#97AfNvHvYV	Thomase Dharkikiran Trust	TDTNarsarai	

Figure 2. Example of data output from a third-party platform(not limited to the selected columns). Note: other third-party platforms might give a different output.

One of the data sets used in the empirical analysis here (Jiang et al., 2020) was collected using this method. Publicly released Twitter data sets that comply with Twitter’s terms of service can be found at specialized websites such as DocNow Catalog (n.d.). Using tweet IDs, one can relatively easily “rehydrate” back into full tweets using tools such as *rehydratR* (Coakley & Steinert-Threlkeld, n.d.) or packages such as *rtweet* (Kearney et al., n.d.).

Financial costs

Twitter’s standard APIs. Standard APIs under version 1.1 (V1.1) allow researchers to collect a small sample of tweets and associated metadata for free. They are good choices for pilot-testing initial research ideas. Researchers can search tweets for the past seven days with desired filtering and sampling and interact with the Twitter platform at no cost. If a researcher is primarily interested in gathering a random sample of topic-matched or general Twitter discourses, the standard APIs are probably sufficient as long as the researcher has the skills to set up the API connection to cover the required time period. However, researchers who need longer historical coverage or higher sampling density beyond the free standard APIs will need to switch to a premium subscription, which allows researchers to collect full-

archive tweet data with a current monthly cap of 1.25 million. The current pricing for premium Search Twitter APIs ranges from \$149 to \$2,499 per month, depending on the types of features and the level of access needed (see Table 1).

Third-party access. Based on our own research on several popular third-party platforms on the market, the subscription costs of third-party platforms vary considerably, ranging from \$7,000 to \$50,000 for an annual contract based on quotes received in 2019. It is important to compare multiple platforms and closely examine data availability and cost structure, including, but not limited to, how the maximum cap is placed, whether the number of topics is restricted, historical data access, coverage of social media sources, user access authorization, and data output. For example, some platforms might not be able to provide a tweet’s original ID, making it hard for researchers to further extract profile information and other metadata about a matched tweet. Or a platform might claim that overall data access is unlimited but still impose a daily cap on the total number of mentions allowed to output. Some platforms charge by the number of topics to be examined, offering access to three years of historical data and allowing separate dashboards for different users on the same research team. Other platforms charge by a monthly maximum, going back to when Twitter was founded and requiring

Table 1. Financial costs of different data collection circumstances.

	Standard	Premium	Enterprise
Twitter			
Search API: 7 days	Free	Not applicable	Not applicable
Search API: 30 days	Not applicable	\$149–\$2,499	By negotiation
Search API: Full archive	Not applicable	\$99–\$1,899	By negotiation
Filtered Streaming API	Free	Not applicable	By negotiation
Sampled Streaming API	Free	Not applicable	By negotiation
Third-party platforms	Prices vary across vendors		

Table 2. Skill sets required for using various Twitter tools.

	Data collection	Data storage	Data analysis
Twitter's standard APIs	1. Basic knowledge about how to use a programming language such as R or Python 2. Basic knowledge about how Boolean operation works in order to build the search strings	1. Knowledge about saving common formats of structural data such as CSV or XLSX on a local machine	1. Knowledge about using basic functions in Excel or programming software such as R or Python 2. When the data size is large, knowledge about how to use parallel computing packages in R and Python is necessary
Third-party platforms	1. Knowledge about building Boolean operations and filters 2. Knowledge about how to design automated scripts but not mandatory	1. Knowledge about saving common formats of structural data such as CSV or XLSX on local machine	1. Knowledge about working with common formats of structural data such as CSV or XLSX
Full-archive access	1. Knowledge about connecting Twitter full-archive server with local storage	1. Knowledge about managing multiple workstations/servers and working on distributed processing frameworks such as Hadoop or Spark 2. Knowledge about making local filtering and sending queries (e.g., SQL)	1. Knowledge about processing a large batch of JSON files

team members to share the same dashboard. Most platforms cover data from the big four social media websites (Facebook, Twitter, Instagram, and YouTube) plus some online news websites. Some also cover foreign social media sites such as China's Weibo. It is useful to request a demonstration from a sales representative, download a sample data output, and examine whether all the required information is available.

Skill sets

Twitter's standard APIs. As indicated in Table 2, several technical skills are necessary for using the standard APIs for data collection, storage, and analysis. To collect tweets through standard APIs, researchers need some basic knowledge about a data science programming language such as R or Python. That fundamental knowledge will allow researchers to employ several useful open-source packages available in R or Python to access Twitter's APIs, such as *twitteR* (Gentry, n.d.), *python-twitter* (Python-Twitter Developers, n.d.), and

tweepy (Roesslein, n.d.). There are many useful online tutorials that teach researchers how to use these tools, though it can take a few weeks of intensive study to grasp these off-the-shelf packages without diving into the technical details of these programming languages. Additionally, researchers need to understand Boolean operation, which is necessary to build search queries. Once the search query is developed and activated to start data collection, researchers also must be able to debug their code to address error messages. For instance, errors such as a broken connection or reaching an API quota might occur. Researchers need to understand these messages or patiently access the resources available to understand them in order to write code to fix the errors. Pilot tests are often an effective way to identify where and when the codes might break.

Regarding data storage and data analysis, researchers can often process up to several million tweets on a single updated laptop as long as the analytical strategy allows processing tweets sequentially in batches. Researchers

can store their raw tweets spread across several files based on day or hour. While the data can be slowly analyzed in this basic form, to analyze these files more quickly, researchers can use “embarrassingly parallel”² programming in R or Python to examine each file separately if the job of each file is independent of the others (i.e., the result of one file does not need to be finished in order to serve as the input to the next file).

Third-party access. Third-party platforms with well-organized query structures and interactive interfaces can relieve researchers from massive programming work but still require some basic technical skills. Resources for researchers regarding platform functionality are generally available through reviewing official instruction manuals, watching video tutorials, and consulting technical support teams.

Researchers should be aware of platform policies and limits while designing a project. *Policies* refers to the platform-imposed rules that specify ways to extract matched tweets, available Boolean operations to support query construction, and format and size to standardize data export. It is common for platforms to implement their own policies to keep services consistent across customers. Common caps include data rates, data volume per request, request cap for a period, and request length. All of these can affect data quality as well as researchers’ time and computational resources requirements. It is worth noting that automation scripts (e.g., Selenium) can reduce time and resource burdens.

Specialized access: Full-archive access and researcher access

While many research questions can be addressed using conventional access points through Twitter APIs, third-party platforms, and shared tweet IDs, certain research questions require full-archive access. For example, projects aiming to analyze message diffusion structure and dynamics typically need to reconstruct the complete diffusion chains for tweets of interest such as fake news (Vosoughi et al., 2018) and health messages (Meng et al., 2018). In these situations, only the complete collection of tweets matched with filtering criteria can

²For details, here is a useful presentation on “embarrassingly parallel computation”: <http://cs.boisestate.edu/~amit/teaching/530/handouts/ep.pdf>. Embarrassingly parallel processing can be achieved on a single laptop because a modern laptop is often equipped with multiple cores which can complete batches of tasks independent of each other. The snow/parallel package in R (Tierney et al., 2018) and the Joblib package in Python (Joblib, n.d.) are useful for multicore processing.

offer the granularity, resolution, and time stamps necessary to accurately infer who retweeted whom and when along the diffusion chain. Relatedly, research projects involving networks, such as the follower-followee network or the network formed by quoting, replying, or retweeting, are better served by obtaining full-archive access, because random samples of tweets remain prone to systematic biases in recovering key network-level attributes (González-Bailón et al., 2014; Lee et al., 2006). For these reasons, we briefly introduce methods for access, financial cost, and skill sets relevant to the full-archive access to Twitter data.

Twitter’s enterprise-level APIs enable users to access its full archive dating back to the first tweet in March 2006. Compared with the standard version, enterprise APIs remove a lot of usage restrictions, but they are financially costly for individual researchers and graduate students. In addition, users must request enterprise-level access from and negotiate related terms with Twitter. One of the enterprise packages is the full-archive Search API. According to Twitter, this API provides complete and instant access to the full corpus of Twitter data. It is query-based access to the tweet archive with minute granularity. Using this approach, users define the filter rules, and tweets matching those queries become available from the Search API about 30 seconds after being published.

Often, these data are accessed via Twitter’s PowerTrack APIs, commonly known as the Firehose endpoints. PowerTrack API, which also provides access to the full Twitter archive, is the enterprise-level package based on the standard filtered Streaming API. Full-archive access, previously operated by data providers such as Gnip and now available directly from Twitter, offers 100% data coverage and allows complex search filters to match tweets. By applying the PowerTrack filtering language (e.g., geo-location and key words), users can filter the real-time stream of tweets. Decahose is another enterprise-level API based on the sampled Streaming API. It delivers a 10% random sample of the real-time Twitter Firehose. Table 3 summarizes two major types of Twitter APIs, Search APIs and Streaming APIs, and their three subscription levels: standard, premium, and enterprise.

Around mid-2020, Twitter made available to invited institutions and research groups the streaming endpoint as part of its PowerTrack API, to provide free Firehose access to COVID-19-related tweets. In October 2020, Twitter tested the new Academic Research product track in a private beta program and publicly launched it in

Table 3. Twitter's standard, premium, and enterprise levels of Search and Streaming APIs.

	Standard	Premium	Enterprise
Search API: 7 days	✓	Not applicable	Not applicable
Search API: 30 days	Not applicable	✓	✓
Search API: Full archive	Not applicable	✓	✓
Filtered Streaming API	✓	Not applicable	✓
Sampled Streaming API	✓	Not applicable	[Firehose via PowerTrack] ✓ [Decahose]

January 2021. In particular, Twitter released a new version of its APIs (V2) that is built on new functionalities, making it easier to collect and analyze the public conversation and simpler to scale up or down without changing APIs, as well as making it friendlier to academic researchers.³ Previously, full-archive data were inaccessible to most researchers.

Enterprise APIs, such as Firehose via the Historical PowerTrack API, offer the highest level of access and reliability. Twitter currently does not have a fixed pricing structure. Accessing this type of APIs needs to be negotiated directly with Twitter, and users are charged on a case-by-case basis by assessing the nature and volume of requested data.

Given the large volume of data output from Firehose access, multiple workstations/servers can considerably enhance efficiency in handling data storage and processing. For example, in the analysis below of the COVID-19 Firehose corpus, a mere week's volume in June easily reached more than 30 million tweets, and we had to find dedicated servers to store the entire corpus covering the second half of 2020, with the size measured in terabytes. This volume of data can overwhelm the most powerful laptops on the market. Raw tweets usually come in JSON files, which are better reorganized using distributed processing frameworks such as Hadoop or Spark to enable more efficient storage, filtering, and retrieval (e.g., via SQL).

Analytically, while analytical strategies such as dictionary-based scaling can be carried out in batches, almost independently of each other, other approaches such as topic modeling might require more advanced distributed processing to update parameter estimates globally across the entire corpus. Also, one needs to pay attention to whether the preferred analytical strategy is scalable and friendly to distributed processing. For

example, although structural topic modeling is a preferred method to examine the linkage between metadata and topic structures (Roberts et al., 2019), it is typically less scalable and slower to run than the latent Dirichlet allocation algorithms adapted to distributed processing on large-scale data sets.

To bypass the technical complexities of setting up the distributed data storage and analytical infrastructure (e.g., Hadoop, distributed processing), one feasible option is to seek collaboration with computer and data scientists. We should caution that not all such interdisciplinary collaborations bear out. It is critically important to understand and resolve disciplinary differences such as incentives, research goals (e.g., explanatory theory building versus prediction improvement), publication processes (e.g., slower journal papers versus faster conference proceedings), and authorship norms. Our experiences suggest that grand challenges of common interest to multiple disciplines, such as the COVID-19 crisis, and opportunities for grant applications might create strong enough incentives to initiate and sustain such collaborations.

Ethical and legal issues are high-priority concerns

For researchers who use or plan to use Twitter data, several ethical and legal issues are worth considering. The first ethical challenge is whether and how to obtain informed consent from Twitter users whose information may be collected. As argued by Lomborg and Bechmann (2014), collecting informed consent is unrealistic for large-scale quantitative research since there is no direct contact between researchers and research participants; however, consent may be viable in small-sample qualitative studies using APIs (Lomborg & Bechmann, 2014). Another ethical and legal challenge is how to analyze tweets related to sensitive topics such as people's health status, sexual orientation, and religion. In terms of legal compliance, researchers should keep in mind that

³For more information, see <https://developer.twitter.com/en/solutions/academic-research/products-for-researchers#early-access-v2-api>.

sharing the raw data set is prohibited under Twitter's current developer policy.⁴ Permittable workarounds include sharing tweet IDs or simply reporting the key words and retrieval time frame to allow other researchers to replicate data collection. Fiesler and Proferes (2018), Norval and Henderson (2020), and Webb et al. (2017) provide detailed discussions of ethical concerns with using Twitter data.

Data quality

In this section, we highlight three major issues that researchers need to keep in mind when evaluating the quality of Twitter data. We then present empirical evidence of the representativeness of standard API and third-party platform samples of Twitter full-archive data, the ground truth.

The pros and cons of Twitter as organic data

Twitter data differ in nature from the data that researchers collect from traditional quantitative methods such as surveys or experiments. Survey data are researcher controlled and designed, whereas social media data can be viewed as more organic (Groves, 2011). We highlight several nuances (i.e., advantages coexist with challenges) about data quality that researchers should consider due to this organic nature. The first nuance is data quality versus data newness. Data quality is more likely ensured in surveys and experiments, as researchers have more control of which participants to recruit and what questions to ask. Nevertheless, the emergent nature of social media discussions may offer researchers opportunities to identify new, previously unidentified perspectives and frames (Klašnja et al., 2017, p. 17). "Newness" is a strength of social media data and is especially useful for studying emerging life sciences issues where the right questions to ask are elusive. However, data newness comes with a data quality challenge that requires researchers to develop methods to indirectly evaluate user characteristics such as user identity and motivations (Chen & Tomblin, *in press*).

The second nuance is control and cost. Researchers have more control of the data generation process in survey and experiments. Yet, it is costly to collect survey

⁴See <https://developer.twitter.com/en/developer-terms/policy>. This policy, which may be changed without notice, provides rules and guidelines for developers who interact with Twitter's ecosystem of applications, services, website, web pages, and content.

data, especially if a large sample of panel data is needed. For using Twitter to collect social media data, researchers cannot control sample composition, but they can assess sample representativeness of targeted populations through the limited profile information provided by the users. If researchers have domain knowledge about social media platform users, then social media is potentially much cheaper for collecting large and time-series data sets. Researchers also need to be aware that the data generation from social media can suffer problems such as algorithmic bias, polarization, and segregation (Baeza-Yates, 2018; Colleoni et al., 2014; Lawrence et al., 2010). In fact, social media companies regularly conduct A/B testing (i.e., a simple controlled experiment with two variants) to understand their users. Yet, researchers know little about what A/B testing these companies are doing and how it may confound research results (Freiling et al., *in press*). For instance, suppose the platform is experimenting with misinformation correction messages and the treatment is effective at changing users' opinions. If the researcher collects data on users' opinions during the experimental time frame, he or she might reach biased conclusions because the researcher does not know about this misinformation correction intervention.

Finally, there is also a theoretical consideration of whether the 1% random sample from the Streaming API is indeed random. As Twitter points out,⁵ its standard API (V1.1) focuses on relevance and not completeness. Research also suggests that the universality of the 1% ceiling with millisecond filtering is unclear, and therefore there is no assurance of a random sample (Kim et al., 2013; Pfeffer et al., 2018). We provide more details on how we can assess the randomness later.

Bot intervention

Bot accounts are prevalent and evolving (Cresci, 2020). Bots account for 37.2% (Imperva, 2020) of all internet traffic. Existing literature shows that at least 39 countries, including the United States, the United Kingdom, Canada, Russia, and China, have reported cases of political manipulation involving social bots during political events (e.g., elections) or in normal times (Bastos & Mercea, 2019; Bolsover & Howard, 2019; Cresci, 2020; Luceri et al., 2019; Stukal et al., 2017).

It is necessary to assess whether bots warrant concern about the reliability and representativeness of social

⁵See <https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview>.

media data, as bots and humans have become increasingly difficult to distinguish from each other. The answer depends upon the nature of the research question. For example, when the interest is the discursive strategies of legitimate users such as political elites, bot detection might not be necessary, since politicians are seldom willing to hand over their social media accounts to algorithms. However, when the goal is to depict online discourses or topical trends in public opinion, naively treating social media data as exclusively human traces without considering the existence of bots may lead to misjudgment (Duan et al., n.d.).

Methods have been developed for detecting social bots. One off-the-shelf bot detection tool is Botometer (Observatory on Social Media, n.d.), a machine learning-based system which evaluates the activity of a Twitter account and generates a bot-like score (Davis et al., 2016). Besides the machine learning approach, other approaches are graph-based (Hurtado et al., 2019), crowdsourcing-based (Wang et al., 2012), and anomaly-based (Echeverria & Zhou, 2017; Orabi et al., 2020).

Sample representativeness

For big data research, the “bigness” of data size can never guarantee representativeness of the target population, a critical challenge facing scholars who use social media data. Despite its popularity as a source of accessible digital trace data, Twitter attracted a mere 22% of Americans as self-reported ever users in 2019 (Perrin & Anderson, 2019). Even for Facebook, the most popular social networking site to date, still around 30% of Americans had never used its service as of 2019 (Perrin & Anderson, 2019). The sample representativeness of the general population of Twitter data is still debated, but it appears to be highly contingent on the research question. For certain topics, research has shown that Twitter is a credible source to reproduce real-world outcomes such as protest and violence (Muchlinski et al., 2021; Sobolev et al., 2020; Steinert-Threlkeld, 2018; Zhang & Pan, 2019). However, for other research questions, such as studying public opinion across the United States, we align with other scholars who express caution about extrapolating Twitter users’ opinions to other Americans, such as about people’s happiness level (Jensen, 2017). Twitter users are younger and are more likely to be Democrats compared with the general public, and 10% of users created 80% tweets (Wojcik & Hughes, 2019). As Barberá & Rivero (2015) suggested,

Twitter provides researchers with the opportunity to study public opinion, yet the validity of generalizations needs to be assessed along with the biases within political discussions on Twitter.

Other sources of data bias

Two additional sources of potential biases may further threaten the quality of obtainable big social media data: (1) the quality of search terms for data retrieval and (2) the black box of APIs and third-party platforms especially their sampling logic. There have been extensive discussions about search term development and evaluation (Kim et al., 2016; King et al., 2017), but little is known about the sampling quality of APIs and third-party platforms, which is our focus in the empirical assessment.

Researchers relying upon APIs and third-party platforms rarely have the opportunity to verify how sampling is handled on the back end. For example, according to Twitter documentation, compared with the latest API (V2), its free-access standard API (V1.1) is less likely to return true randomly sampled data, because it is designed to improve relevance for consumer-use cases. Although this potential deviation from true random sampling has been corrected in API (V2), such biases might have permeated numerous third-party platforms that rely upon Twitter’s APIs.

To evaluate the extent to which such biases in sampling of full-access data might exist in Twitter’s standard API (1% data coverage) and third-party platforms, we next use the same list of search terms related to COVID-19 to compare the distributions of one specific type of content attribute—moral appeals—in three collections of COVID-19 tweets for a two-week period of data (July 1–July 14, 2020, Coordinated Universal Time): Twitter’s standard API (1% coverage), a third-party platform that claims to provide 10% random sampling, and Twitter’s full-archive access (the ground truth).

Third-party platform and Twitter API representativeness of full-archive data

Over the past decade, interest has grown in applying moral foundations theory (MFT) (Graham et al., 2013; Haidt, 2012) to investigate the roles of moral appeals and moral values in political ideology, public opinion, and individuals’ processing of information related to controversial medical and scientific issues (Clifford & Jerit, 2013; Skitka & Morgan, 2014; Wolsko et al., 2016; Yang et al., 2018). MFT has become a useful

Twitter as research data: Tools, costs, skill sets, and lessons learned

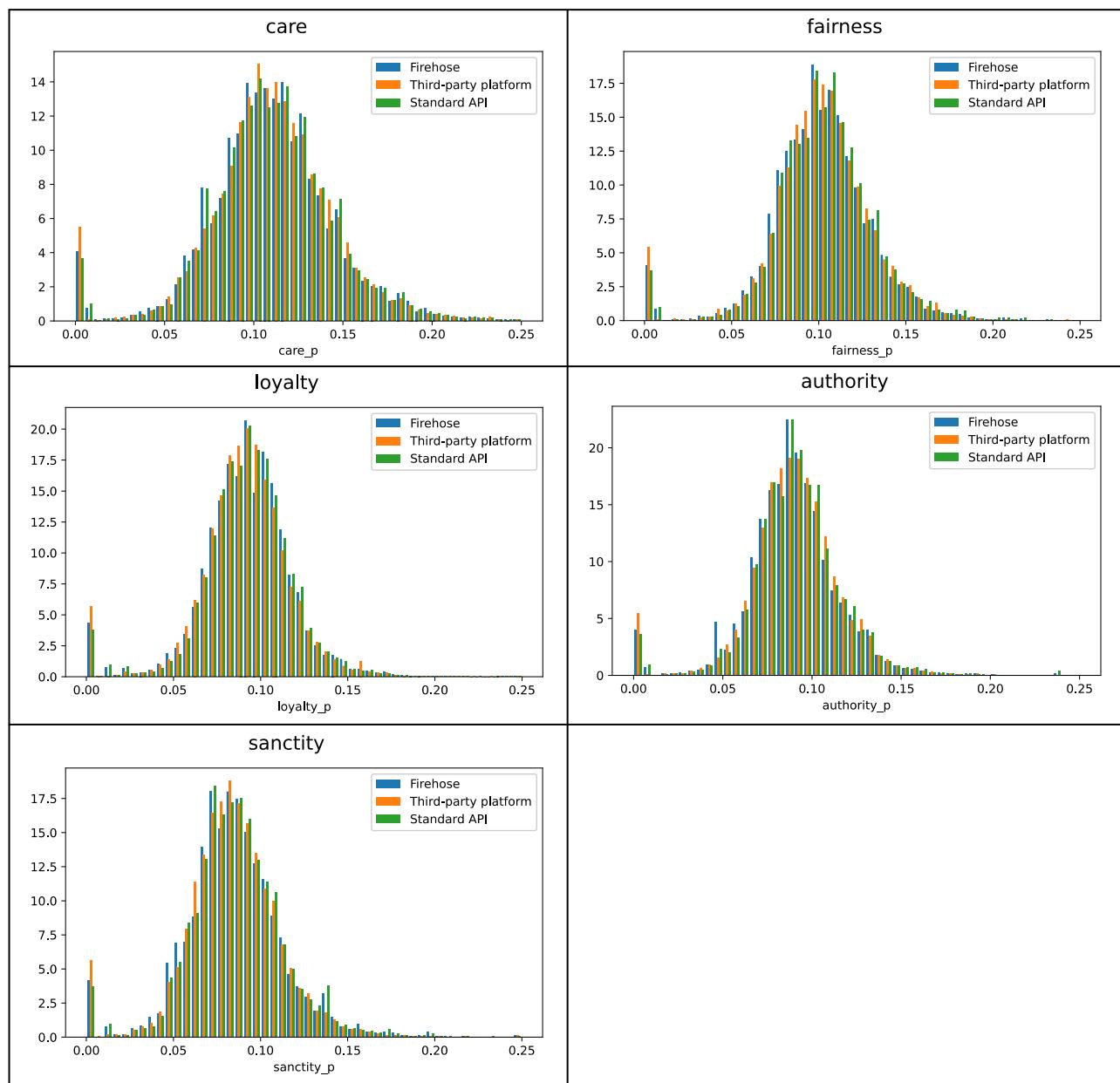


Figure 3. Comparisons of score distributions of five moral appeals.

theoretical framework for scholars studying politics and life science communication. In this analysis, we use the most recent extended Moral Foundations Dictionary (Hopp et al., 2020) to measure each tweet's mentioning of the five types of moral appeals (care, fairness, loyalty, authority, and sanctity). We focus on relevance scoring and do not distinguish moral virtues (e.g., a tweet praising mask-wearing as caring for others) from vices (e.g., a tweet condemning violation of social distancing that harms others). Since many social science questions

concern extracting theoretical constructs similar to moral appeals from social media data, this empirical assessment is intended to provide insights into the prevalence (or the lack of) of potential sampling biases likely to generalize to other research contexts.

If Twitter's standard API and third-party platforms indeed return random samples of matched tweets under identical screening criteria (e.g., language, time period, and key words), we should expect to see similar distributions of *any* content feature—including moral appeals

Table 4. Comparison of five basic statistics across three data sources.

	Care	Fairness	Loyalty	Authority	Sanctity
<i>Full archive</i>					
Mean	0.110	0.102	0.092	0.089	0.086
Median	0.109	0.101	0.093	0.089	0.084
Max	0.750	0.600	0.563	0.750	0.667
Min	0.000	0.000	0.000	0.000	0.000
SD (σ)	0.037	0.033	0.027	0.028	0.031
<i>Third-party platform</i>					
Mean	0.110	0.102	0.091	0.090	0.084
Median	0.110	0.102	0.091	0.090	0.084
Max	0.750	0.556	0.562	0.750	0.520
Min	0.000	0.000	0.000	0.000	0.000
SD (σ)	0.037	0.031	0.027	0.028	0.029
<i>Standard API</i>					
Mean	0.110	0.103	0.092	0.091	0.086
Median	0.110	0.103	0.093	0.090	0.085
Max	0.609	0.458	0.500	0.750	0.520
Min	0.000	0.000	0.000	0.000	0.000
SD (σ)	0.037	0.031	0.026	0.028	0.030

—across the three data sets. Figure 3 presents the density plots comparing the three data sources, stratified by moral appeal type. Similar to previous studies (Hopp et al., 2020), the empirical distributions largely resemble a normal distribution. The overall shape, central tendency, and spread of these distributions appear comparable from visual inspection, which is consistent with similar estimated means and standard deviations across the three data sets, per moral appeal category (see Table 4 for details on basic descriptive statistics).

Further, we perform a series of Kolmogorov-Smirnov (K-S) tests to further investigate the statistical comparability across the three data sources. In the two-sample case, the K-S test statistic is an indicator of the maximum distance between the empirical cumulative distribution functions of two samples (Massey, 1951). In our study, two steps were taken: first, we repeatedly sampled 10,000 tweets without replacement from the full-archive Firehose 1,000 times, and then we calculated the K-S statistics between each sample against the entire Firehose corpus to derive an empirical distribution of the K-S test statistics under the null hypothesis, respectively for each moral appeal category. Next, for each moral appeal type, we estimated the K-S statistics comparing the entire third-party corpus and also the entire standard API corpus with the Firehose, respectively (see the two dashed vertical lines in Figure 4).

As Figure 4 indicates, for most moral appeal categories, the two observed K-S test statistics lay beyond the 97.5% quantile cut-off points (two-tailed tests) in the empirical K-S distributions under the null, suggesting

statistically significant discrepancies in the shape of most moral appeal distributions between the Firehose and the other two data sets. This pattern casts doubt on the assumption that the other two corpora could be treated as random samples from the Firehose. With that said, we should note that K-S tests are sensitive to sample size and can easily pick up noise in local discrepancies. Given our large sample sizes, we would like to emphasize the lack of substantial discrepancies between the three data sets in terms of the descriptive statistics reported in Table 4 and the highly comparable density plots in Figure 3.

New development: Twitter Academic Research API

On August 12, 2020, Twitter released Twitter API V2, including a dedicated Academic Research track. In this version, Twitter rebuilt the foundation of its API services, redesigned the access levels and developer portal, and introduced new product tracks for different use scenarios. In particular, the free Academic Research track, available to researchers upon application and Twitter's approval, provides access to the full Twitter archive, though this access is currently subject to some limitations. Although anyone can apply for the new API V2, as of spring 2021, it seemed that applicants with academic institutional affiliations having a clearly defined research project and pledging to adhere to Twitter's Developer Policy had a greater chance for approval for the Academic Research track. Next, we outline major updates to API V2, especially the Academic Research track, in terms of access restrictions, scope of the data set available, and changes in the data organization unit.

Regarding data access, under API V2, researchers could access the full Twitter archive for free, subject to a monthly cap of 10 million tweets, as of spring 2021. This access is similar to the enterprise level in V1.1 mentioned in Table 1; yet in V1.1, most requests at this level would incur substantial financial cost. The Academic Research track supports full-archive search for any topic that the researcher is interested in, as long as the topic is not out of the scope delineated in the initial application. It is not clear how closely Twitter monitors the scope of queries; however, we encourage interested researchers to submit revised requests to Twitter should they decide to substantially modify the scope.

The third noteworthy feature is an added organizing logic for retrieved data, namely, the logic of conversation thread. API V2 assigns a shared conversation ID to all

Twitter as research data: Tools, costs, skill sets, and lessons learned

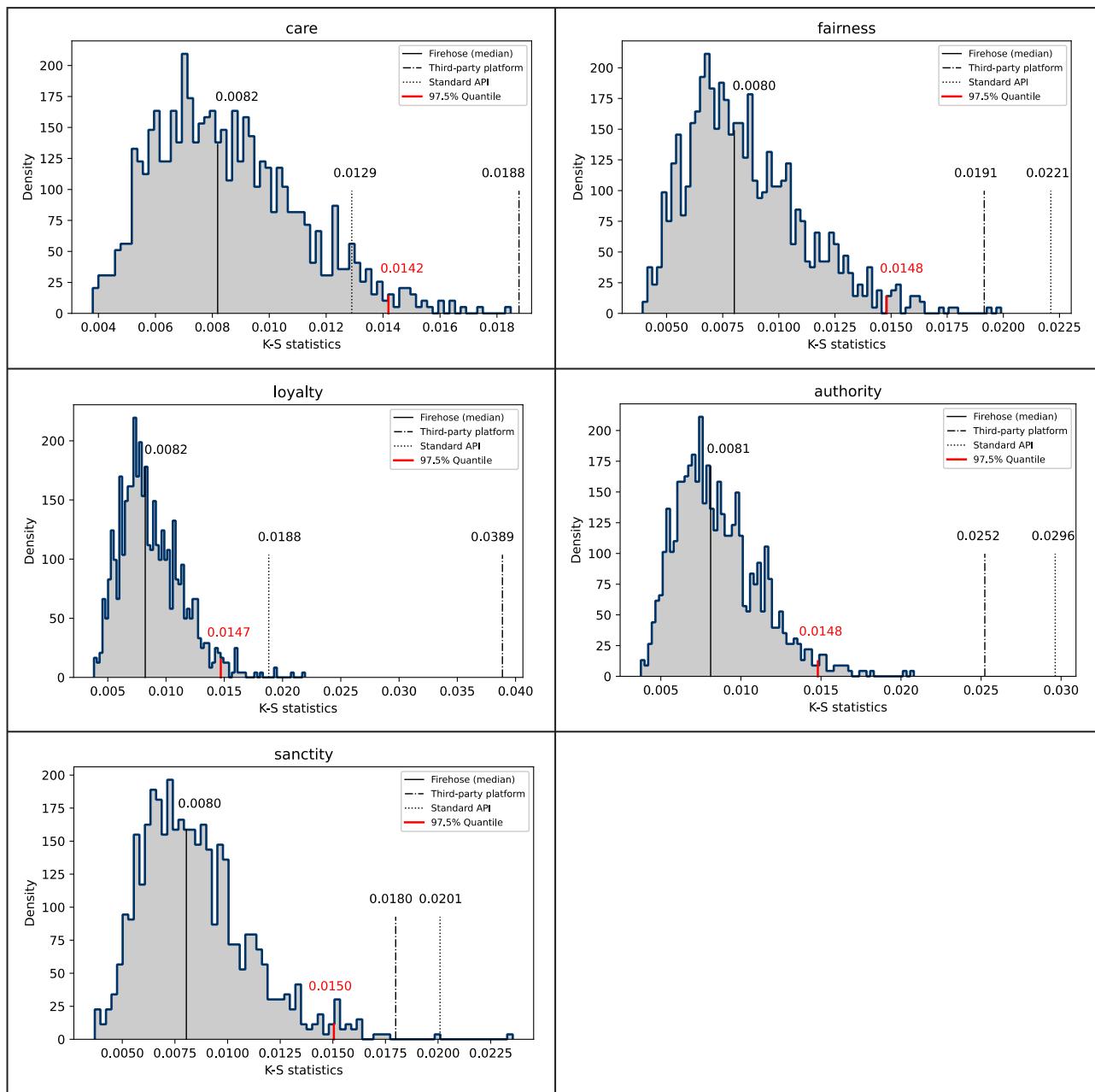


Figure 4. K-S statistics distribution comparison.

replies to an original tweet, as well as replies to a reply. Therefore, researchers can retrieve and reconstruct the entire thread of the evolving conversation linked by the conversation ID and indexed by time stamps. Previously in API V1.1, only segments of the conversation could be retrieved, lacking key information to accurately track how a conversation grows in length and complexity. Now, this critical information is readily accessible under API V2.

Under the Academic Research track, researchers are now able to access a larger Twitter archive, both historical and streaming, at nearly no cost, which can benefit researchers who have access to limited funding. In addition, researchers have the opportunity to explore research questions related to public discourses and information diffusion and mutation more easily than before. For instance, researchers can track the evolution of Twitter conversations over time by taking advantage of

conversation IDs. Also, it is much easier to calculate public engagement with content made feasible by using publicly available metrics (e.g., number of likes or retweets) and internal metrics (e.g., impression count or URL link clicks), a capacity available now yet previously hidden.

Conclusion

Twitter data are well suited to a number of research questions and have been used in published research on several important topics (e.g., Budhwani & Sun, 2020; Chew & Eysenbach, 2010; Fownes et al., 2018; Paul & Dredze, 2011; Singh et al., 2020; Vance et al., 2009; Wang & Guo, 2018; Wirz et al., 2018). But collecting Twitter data can be costly and time-consuming, and it often requires scholars to learn new skills. Building an interdisciplinary team is one way to efficiently address these challenges as well as to study meaningful problems at the intersection of life sciences and politics, which by nature are already multidisciplinary. Our goal in this article has been to introduce Twitter data as a viable research resource and to highlight key issues surrounding access, costs, training, and data quality that researchers need to consider when deciding whether and how to use these data. We summarize the main points of our paper as follows:

Tools, costs, and training

- Scholars can use Twitter's standard APIs, third-party platforms, and the full archive to access Twitter data. Before thinking about the costs and training issues, researchers need to base their choice of data tools on what research questions they are interested in studying. When the focus is on the *content* of Twitter discourse about life sciences and politics, accessing a sample of tweets collected by the standard APIs or third-party platforms will likely suffice. Studying the *network* among users, however, requires researchers to access the Twitter full archive in order to construct a complete network.
- Budget is almost always a constraint for researchers. Twitter offers free but limited access to data in some cases, but the standard APIs and third-party platforms require researchers to expend money if they need a large sample of tweets rather than 1% or limited historical tracing. However, budget constraints can be worked around if researchers have a flexible schedule about data collection. For instance, Twitter's

Search API is free going back seven days of data. Yet, researchers can write scripts to collect data every seven days to build a time-series data set for free. The new Academic Research track for Twitter API V2 helps reduce the financial costs for Twitter-approved researchers.

- Different learning curves are required for various tools. Third-party platforms require the least programming skills. Researchers, though, are restricted by certain types of Twitter data that third-party platforms give to them. While acquiring new skills is essential, building an interdisciplinary team is also beneficial to accelerate the process.

Data quality

- Twitter data are organic, which has advantages and disadvantages for researchers. It is a recommended practice to assess and describe who the users are from collected tweets and whether the presence of bots is a concern for key research questions.
- Critically assessing the promise of the 1% *random* sample from standard APIs or a higher percentage of a random sample from third-party platforms is a key task facing future research in order to assess the representativeness of data returned by various data tools.
- We compared moral appeal distributions to assess sample representativeness relative to full-archive access. If researchers have access to multiple Twitter data sets, it is useful to carry out similar evaluations using content features tailored to their own research projects. Dictionary-based content scaling could be used as an easy-to-implement and efficient means for such comparative analyses.
- In our empirical examination comparing the full-archive Firehose data to the 1% sample from Twitter's standard API or to the third-party platform, discrepancies in the distributions of moral content were statistically significant, based on the K-S tests, but not substantial. These findings support using the less costly, sampled Twitter data sets for social science research, such as studying moral messaging in the case of COVID-19. With that said, before we fully understand the representativeness of these sampled tweets from standard APIs or third-party platforms, we emphasize the value of assessing comparability between different data sources to improve confidence in findings from using Twitter data.

Acknowledgments

Kaiping Chen would like to thank the National Science Foundation for supporting this work under Grant No. 2027375. The authors thank Yachao Qian for research assistance, particularly for help with compiling data and conducting analyses. We are also thankful for support from the Mass Communication Research Center and the Center for Communication and Civic Renewal at the University of Wisconsin-Madison as well as grant support from The John S. and James L. Knight Foundation, the Hewlett Foundation, and the Journal Foundation. We thank Twitter, Inc. for providing the COVID-19 stream. Additional support for this research was provided by the University of Wisconsin-Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation.

All authors contributed equally to this article.

References

- Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM*, 61(6), 54–61. <https://doi.org/10.1145/3209581>
- Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, 33(6), 712–729. <https://doi.org/10.1177/089443931455836>
- Bastos, M. T., & Mercea, D. (2019). The Brexit Botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1), 38–54. <https://doi.org/10.1177/0894439317734157>
- Bolsover, G., & Howard, P. (2019). Chinese computational propaganda: Automation, algorithms and the manipulation of information about Chinese politics on Twitter and Weibo. *Information Communication and Society*, 22(14), 2063–2080. <https://doi.org/10.1080/1369118X.2018.1476576>
- Brossard, D., & Scheufele, D. A. (2013). Science, new media, and the public. *Science*, 339(6115), 40–41. <https://doi.org/10.1126/science.1232329>
- Budhwani, H., & Sun, R. (2020). Creating COVID-19 stigma by referencing the novel coronavirus as the “Chinese virus” on Twitter: Quantitative analysis of social media data. *Journal of Medical Internet Research*, 22(5), e19301. <https://doi.org/10.2196/19301>
- Chen, K., & Tomblin, D. (in press). Using data from Reddit, public deliberation, and surveys to measure public opinion about Autonomous Vehicles. *Public Opinion Quarterly*.
- Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1 outbreak. *PLOS ONE*, 5(11), e14118. <https://doi.org/10.1371/journal.pone.0014118>
- Clifford, S., & Jerit, J. (2013). How words do the work of politics: Moral foundations theory and the debate over stem cell research. *Journal of Politics*, 75(3), 659–671. <https://doi.org/10.1017/S0022381613000492>
- Coakley, K., & Steinert-Threlkeld, Z. (n.d.). *rehydratorR, version 0.1.0*. Retrieved June 3, 2021, from <https://cran.r-project.org/web/packages/rehydratorR/readme/README.html>
- Cody, E. M., Reagan, A. J., Mitchell, L., Dodds, P. S., & Danforth, C. M. (2015). Climate change sentiment on Twitter: An unsolicited public opinion poll. *PLOS ONE*, 10(8), 1–18. <https://doi.org/10.1371/journal.pone.0136092>
- Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, 64(2), 317–332.
- Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM*, 63(10), 72–83. <https://doi.org/10.1145/3409116>
- Davis, C., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot: A system to evaluate social bots. *Proceedings of the 25th international conference companion on world wide web* 59(7), 273–274. <http://doi.org/10.1145/2872518.2889302>
- Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B. K., Huskey, R., Schneider, F. M., Breuer, J., Parry, D. A., Vermeulen, I., Fisher, J. T., Banks, J., Weber, R., Ellis, D. A., ... de Vreese, C. (2021). An agenda for open science in communication. *Journal of Communication*, 71(1), 1–26. <https://doi.org/10.1093/joc/jqz052>
- DocNow Catalog. (n.d.). *Documenting the now: DocNow Catalog*. Retrieved June 5, 2021, from <https://catalog.docnow.io/>
- Driscoll, K., & Walker, S. (2014). Working within a black box: Transparency in the collection and production of big twitter data. *International Journal of Communication*, 8(1), 1745–1764.
- Duan, Z., Li, J., Josephine, L., Yang, K.C., Chen, F., Shah, D., & Yang, S. (n.d.). *Bot as strategic communicator in the digital public space: Evidence for algorithmic agenda-setting during the COVID-19 pandemic*. Working paper.
- Echeverria, J., & Zhou, S. (2017). Discovery, retrieval, and analysis of the “Star wars” botnet in twitter. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in*

- Social Networks Analysis and Mining, ASONAM 2017* (pp. 1–8). ACM. <https://doi.org/10.1145/3110025.3110074>
- Fiesler, C., & Proferes, N. (2018). “Participant” perceptions of Twitter research ethics. *Social Media + Society*, 4(1), 2056305118763366. <https://doi.org/10.1177/2056305118763366>
- Fownes, J. R., Yu, C., & Margolin, D. B. (2018). Twitter and climate change. *Sociology Compass*, 12(6), e12587. <https://doi.org/10.1111/soc4.12587>
- Freiling, I., Krause, N., Scheufele, D. A., & Chen, K. (in press). The science of open (communication) science: Toward an evidence-driven understanding of quality criteria in communication research. *Journal of Communication*.
- Gentry, J. (n.d.). *twitteR, version 1.1.9*. Retrieved June 3, 2021, from <https://www.rdocumentation.org/packages/twitteR/versions/1.1.9>
- González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, 38(1), 16–27. <https://doi.org/10.1016/j.socnet.2014.01.004>
- Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. *Advances in Experimental Social Psychology*, 47, 55–130. <https://doi.org/10.1016/B978-0-12-407236-7.00002-4>
- Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, 75(5), 861–871. <https://doi.org/10.1093/poq/nfr057>
- Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., & Marx, W. (2019). Does the public discuss other topics on climate change than researchers? A comparison of explorative networks based on author keywords and hashtags. *Journal of Informetrics*, 13(2), 695–707. <https://doi.org/10.1016/j.joi.2019.03.008>
- Haidt, J. (2012). Moral psychology and the law: How intuitions drive reasoning, judgment, and the search for evidence. *Ala. L. Rev.*, 64, 867.
- Hopp, F. R., Fisher, J. T., Cornell, D., Huskey, R., & Weber, R. (2020). The extended Moral Foundations Dictionary (eMFD): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. *Behavior Research Methods*, 53, 232–246. <https://doi.org/10.3758/s13428-020-01433-0>
- Hughes, A., & Wojcik, S. (2019, August 2). *10 facts about Americans and Twitter*. Pew Research Center. <https://www.pewresearch.org/fact-tank/2019/08/02/10-facts-about-americans-and-twitter/>
- Hurtado, S., Ray, P., & Marculescu, R. (2019). Bot detection in reddit political discussion. *SocialSense 2019: Proceedings of the 2019 4th International Workshop on Social Sensing* (pp. 30–35). ACM. <https://doi.org/10.1145/3313294.3313386>
- Jensen, E. A. (2017). Putting the methodological brakes on claims to measure national happiness through Twitter: Methodological limitations in social media analytics. *PLOS ONE*, 12(9), e0180080. <https://doi.org/10.1371/journal.pone.0180080>
- Jiang, J., Chen, E., Yan, S., Lerman, K., & Ferrara, E. (2020). Political polarization drives online conversations about COVID-19 in the United States. *Human Behavior and Emerging Technologies*, 2(3), 200–211. <https://doi.org/10.1002/hbe2.202>
- Joblib (n.d.). *Joblib: Embarrassingly parallel for loops*. Retrieved June 6, 2021, from <https://joblib.readthedocs.io/en/latest/parallel.html>
- Ke, Q., Ahn, Y., & Sugimoto, C. R. (2017). A systematic identification of scientists on Twitter. *PLOS ONE*, 12(4), e0175368. <https://doi.org/10.1371/journal.pone.0175368>
- Kearney, M. W., Heiss, A., & Briatte, F. (n.d.). *rtweet: Collecting Twitter data, version 0.7.0*. Retrieved June 3, 2021, from <https://cran.r-project.org/web/packages/rtweet/index.html>
- Kim, A. E., Hansen, H. M., Murphy, J., Richards, A. K., Duke, J., & Allen, J. A. (2013). Methodological considerations in analyzing Twitter data. *Journal of the National Cancer Institute Monographs*, 2013(47), 140–146.
- Kim, Y., Huang, J., & Emery, S. (2016). Garbage in, Garbage out: Data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *Journal of Medical Internet Research*, 18(2), e41. <https://doi.org/10.2196/jmir.4738>
- Kim, Y., Nordgren, R., & Emery, S. (2020). The story of goldilocks and three twitter’s APIs: A pilot study on twitter data sources and disclosure. *International Journal of Environmental Research and Public Health*, 17(3), 1–15. <https://doi.org/10.3390/ijerph17030864>
- King, G., Lam, P., & Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, 61(4), 971–988. <https://doi.org/10.1111/ajps.12291>
- Klašnja, M., Barberá, P., Beauchamp, N., Nagler, J., & Tucker, J. (2017). Measuring public opinion with social media data. In L. R. Atkeson & R. M. Alvarez (Eds.), *The Oxford Handbook of Polling and Survey Methods* (pp. 555–582). Oxford University Press.
- Lawrence, E., Sides, J., & Farrell, H. (2010). Self-segregation or deliberation? Blog readership, participation, and polarization

Twitter as research data: Tools, costs, skill sets, and lessons learned

- in American politics. *Perspectives on Politics*, 8(1), 141–157. <https://doi.org/10.1017/S1537592709992714>
- Lee, S. H., Kim, P. J., & Jeong, H. (2006). Statistical properties of sampled networks. *Physical Review E*, 73(1), 016102.
- Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *Information Society*, 30(4), 256–265. <https://doi.org/10.1080/01972243.2014.915276>
- Luceri, L., Badawy, A., Deb, A., & Ferrara, E. (2019). Red bots do it better: Comparative analysis of social bot partisan behavior. In *WWW '19: Companion Proceedings of the 2019 World Wide Web Conference* (pp. 1007–1012). ACM. <https://doi.org/10.1145/3308560.3316735>
- Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253), 68–78.
- Meng, J., Peng, W., Tan, P. N., Liu, W., Cheng, Y., & Bae, A. (2018). Diffusion size and structural virality: The effects of message and network features on spreading health information on twitter. *Computers in Human Behavior*, 89, 111–120. <https://doi.org/10.1016/j.chb.2018.07.039>
- Muchlinski, D., Yang, X., Birch, S., Macdonald, C., & Ounis, I. (2021). We need to go deeper: Measuring electoral violence using convolutional neural networks and social media. *Political Science Research and Methods*, 9(1), 122–139. <https://doi.org/10.1017/psrm.2020.32>
- Norval, C., & Henderson, T. (2020). Automating dynamic consent decisions for the processing of social media data in health research. *Journal of Empirical Research on Human Research Ethics*, 15(3), 187–201. <https://doi.org/10.1177/1556264619883715>
- Observatory on Social Media. (n.d.). *Botometer, version 4*. Retrieved June 7, 2021, from <https://botometer.osome.iu.edu/>
- Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: A systematic review. *Information Processing and Management*, 57(4), 102250. <https://doi.org/10.1016/j.ipm.2020.102250>
- Paul, M. J., & Dredze, M. (2011). *You are what you tweet: Analyzing twitter for public health* [Paper presentation]. Fifth International AAAI Conference on Weblogs and Social Media, Barcelona.
- Peng, Y. (2018). Same candidates, different faces: Uncovering media bias in visual portrayals of presidential candidates with computer vision. *Journal of Communication*, 68(5), 1–22. <https://doi.org/10.1093/joc/jqy041>
- Perrin, A., & Anderson, M. (2019, April 10). *Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018*. Pew Research Center. <https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/>
- Pfeffer, J., Mayer, K., & Morstatter, F. (2018). Tampering with Twitter's sample API. *EPJ Data Science*, 7(1), article 50. <https://doi.org/10.1140/epjds/s13688-018-0178-0>
- Python-Twitter Developers. (n.d.). *python-twitter*. Retrieved June 3, 2021, from <https://python-twitter.readthedocs.io/en/latest/index.html>
- Roberts, M., Stewart, B., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2). <https://doi.org/10.18637/jss.v091.i02>
- Roberts (2020). Bad bot report 2020: Bad bots strike back. Published in Imperva: <https://www.imperva.com/blog/bad-bot-report-2020-bad-bots-strike-back/>
- Roesslein, J. (n.d.). *Tweepy*. Retrieved June 3, 2021, from <https://docs.tweepy.org/en/stable/>
- Singh, L., Bansal, S., Bode, L., Budak, C., Chi, G., Kawintiranon, K., Padden, C., Vanarsdall, R., Vraga, E., & Wang, Y. (2020). *A first look at COVID-19 information and misinformation sharing on Twitter*. ArXiv. <https://arxiv.org/abs/2003.13907v1>
- Skitka, L. J., & Morgan, G. S. (2014). The social and political implications of moral conviction. *Political Psychology*, 35 (Suppl. 1), 95–110. <https://doi.org/10.1111/pops.12166>
- Sobolev, A., Chen, M. K., Joo, J., & Steinert-Threlkeld, Z. C. (2020). News and geolocated social media accurately measure protest size variation. *American Political Science Review*, 114 (4), 1343–1351. <https://doi.org/10.1017/S0003055420000295>
- Steinert-Threlkeld, Z. C. (2018). *Twitter as data*. Cambridge University Press.
- Stukal, D., Sanovich, S., Bonneau, R., & Tucker, J. A. (2017). Detecting bots on Russian political Twitter. *Big Data*, 5(4), 310–324. <https://doi.org/10.1089/big.2017.0038>
- Tierney, L., Rossini, A. J., Li, N., & Sevcikova, H. (2018). *Package "snow": Simple network of workstations, version 0.4-3*. Retrieved June 6, 2021, from <https://cran.r-project.org/web/packages/snow/snow.pdf>
- Vance, K., Howe, W., & Dellavalle, R. P. (2009). Social internet sites as a source of public health information. *Dermatologic Clinics*, 27(2), 133–136. <https://doi.org/10.1016/j.det.2008.11.010>
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146–1151. <https://doi.org/10.1126/science.aap9559>
- Walter, D., Ophir, Y., & Jamieson, K. H. (2020). Russian twitter accounts and the partisan polarization of vaccine

- discourse, 2015–2017. *American Journal of Public Health*, 110(5), 715–724. <https://doi.org/10.2105/AJPH.2019.305564>
- Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M., Zheng, H., & Zhao, B. Y. (2012). *Social turing tests: Crowdsourcing sybil detection*. ArXiv. <http://arxiv.org/abs/1205.3856>
- Wang, W., & Guo, L. (2018). Framing genetically modified mosquitoes in the online news and Twitter: Intermedia frame setting in the issue-attention cycle. *Public Understanding of Science*, 27(8), 937–951. <https://doi.org/10.1177/096362518799564>
- Webb, H., Jiroka, M., Stahl, B. C., Housley, W., Edwards, A., Williams, M., Procter, R., Rana, O., & Burnap, P. (2017). The ethical challenges of publishing Twitter data for research dissemination. In *Proceedings of the 2017 ACM on Web Science Conference* (pp. 339–348). ACM.
- Whittingham, N., Boecker, A., & Grygorczyk, A. (2020). Personality traits, basic individual values and GMO risk perception of twitter users. *Journal of Risk Research*, 23(4), 522–540. <https://doi.org/10.1080/13669877.2019.1591491>
- Wirz, C. D., Howell, E. L., Brossard, D., Xenos, M. A., & Scheufele, D. A. (2021). The state of GMOs on social media: An analysis of state-level variables and discourse on Twitter in the United States. *Politics and the Life Sciences*, 40(1), 40–55. <https://doi.org/10.1017/pls.2020.15>
- Wirz, C. D., Xenos, M. A., Brossard, D., Scheufele, D., Chung, J. H., & Massarani, L. (2018). Rethinking social amplification of risk: Social media and Zika in three languages. *Risk Analysis*, 38(12), 2599–2624.
- Wojcik, S., & Hughes, A. (2019, April 24). *Sizing up Twitter users*. Pew Research Center. <https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/>
- Wolsko, C., Ariceaga, H., & Seiden, J. (2016). Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. *Journal of Experimental Social Psychology*, 65, 7–19. <https://doi.org/10.1016/j.jesp.2016.02.005>
- Yang, S., Maloney, E. K., Tan, A. S. L., & Cappella, J. N. (2018). When visual cues activate moral foundations: Unintended effects of visual portrayals of vaping within electronic cigarette video advertisements. *Human Communication Research*, 44(3), 223–246. <https://doi.org/10.1093/hcr/hqy004>
- Zhang, H., & Pan, J. (2019). Casm: A deep-learning approach for identifying collective action events with text and image data from social media. *Sociological Methodology*, 49(1), 1–57. <https://doi.org/10.1177/0081175019860244>