International Journal of Communication 8 (2014), 1745–1764 1932–8036/20140005

IJoC Big Data, Big Questions

Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data

KEVIN DRISCOLL[1]
University of Southern California, USA

SHAWN WALKER
University of Washington, USA

Twitter seems to provide a ready source of data for researchers interested in public opinion and popular communication. Indeed, tweets are routinely integrated into the visual presentation of news and scholarly publishing in the form of summary statistics, tables, and charts provided by commercial analytics software. Without a clear description of how the underlying data were collected, stored, cleaned, and analyzed, however, readers cannot assess their validity. To illustrate the critical importance of evaluating the production of Twitter data, we offer a systematic comparison of two common sources of tweets: the publicly accessible Streaming API and the "fire hose" provided by Gnip PowerTrack. This study represents an important step toward higher standards for the reporting of social media research.

Keywords: methodology, data, big data, social media

Introduction

> *The instruments register only those things they were designed to register. Space still contains infinite unknowns.*
>
> —Spock (Black, Roddenberry, & Daniels, 1966)

Twitter is increasingly integrated into the visual presentation of news and scholarly publishing in the form of hashtags, user comments, and dynamic charts displayed on-screen during conference presentations, in television programming, alongside political coverage in newspapers, and on websites. In

[1] This research is based in part upon work supported by the National Science Foundation under the INSPIRE Grant Number 1243170.

Kevin Driscoll: kedrisco@usc.edu
Shawn Walker: stw3@uw.edu
Date submitted: 2013–04–10

each case, the transformation of mass-scale Twitter data into statistics, tables, charts and graphs is presented with scant explanation of how the data are collected, stored, cleaned, and analyzed, leaving readers unable to assess the appropriateness of the given methodology to the social phenomena they purport to represent.

This article is about designing research projects that contend with unstable, ambiguous digital environments like Twitter or Facebook. As an increasing amount of everyday social interaction is mediated by these systems, their servers actively aggregate vast stores of information about user behavior. In combination with the falling costs of mass storage and parallel computing, such repositories offer new methodological opportunities that combine characteristics of the micro and the macro. The "big social data" paradigm, to borrow a term from Lev Manovich (2011), seems to combine the grand scale and generalizability of methods like national surveys with the granularity and detail of close textual analysis, ethnography, or participant observation.

And yet, researchers who pursue big social opportunities face a host of daunting challenges—theoretical and ethical as well as technical—that may not be obvious from the outset. Whereas the reliability and validity of established social scientific methods depend on their transparency, big social data are almost universally produced within closed, commercial organizations. In other words, the stewardship of this unprecedented record of public discourse depends on an infrastructure that is both privately owned and operationally opaque.

Among the many sites where big social data is collected, Twitter is particularly compelling because of its perceived accessibility. In comparison to Facebook, which is largely closed-off to the academic community, or a high-bandwidth site like YouTube, tweets are small in size, public by default, numerous, and topically diverse. With little more than a laptop, an Internet connection, and a few lines of scripting code, researchers can aggregate several million tweets in a short period of time using widely-available, low-cost tools.

This accessibility has contributed to an explosion in published research dealing with Twitter data. Unfortunately, as several scholars have previously noted, three recurring problems have limited the long-term utility of much of this research. First, the apparent ease with which tweets may be aggregated belies the difficulty of designing a reliable, reproducible data collection strategy. Second, Twitter itself is a dynamic system subject to near constant change, and the interface features, client software applications, and data formats provided today may be fundamentally different tomorrow. Third, researchers lack a common vocabulary for describing tweets and the metadata that accompany them. This terminology problem is confounded by the misleading use of familiar terms like "sample" and "reply" within the platform's internal ontology. Finally, for all its representation in mainstream media—and popularity among researchers—Twitter use is not evenly distributed among Internet users in general and, even among those that use it, a significant percentage rarely send a tweet of their own, preferring instead to "listen" to the tweets of others.

Fortunately, the three recurring problems are not intractable and can be addressed with good research design habits. This article offers a close look at two different aggregation methods to help researchers critically develop their own data collection strategies. This comparison builds on recent efforts to develop a common terminology for describing Twitter data. Although Twitter, Inc. is not likely to become more transparent, or to cease developing their platform any time soon, the combination of an independent descriptive lexicon and set of parameters for evaluating different collection strategies will enable researchers to mitigate the negative effects of the unstable commercial service on their scholarly work.

Producing a Common Language for Twitter Research

Axel Bruns and Jean Burgess have made the most concerted effort to develop a consistent terminology and set of metrics to enable broad comparison across different studies (Bruns & Burgess, 2011a, 2011b, 2012). In a series of papers and presentations, they systematically detail the various traces made publicly accessible by Twitter. These include the standard information one might expect, such as the text of the tweet, the handle of the sending user, and the time it was sent, as well as background metadata such as the sender's default language (according to the sender's user profile), any geographical information provided by the sending user's device, the application used to send the tweet, and a link to the sender's profile picture. A second order of metadata is then identified within the text of the tweet: hashtags, @-mentions, and URLs. Taken together, these primitive components provide a set of basic descriptive characteristics that might be reported about any collection of tweets.

The descriptive reports recommended by Bruns and Burgess—including the count and types of tweets with links, hashtags, and user mentions—are not themselves analytic, but rather provide a foundation for analysis. The meanings that users attach (or do not attach) to the traces we collect are always locally negotiated in response to a particular context. Taken-for-granted features of Twitter such as the hashtag and @-mention began as user-driven innovations. They were first implemented by third-party software developers and only later adopted by Twitter itself (Twitter, 2009). This history is helpful as it reminds us not to take the metadata provided by the platform at face value. Twitter reports more than 50 fields of metadata with each tweet—even more for retweets—but these must be interpreted by researchers to understand the polysemy of various hashtags, the positions and locations of users within a given community, and the correspondence of screen names to individuals, organizations, and nonhuman agents.

One weakness of the typology suggested by Bruns and Burgess is that it adheres closely to the labels provided by the Twitter platform. Terms such as "reply" and "friend" should be regarded with suspicion as they may or may not represent the meaning suggested by their labels (Howison, Wiggins, & Crowston, 2011). In addition, the meaning of behaviors such as retweeting, @-mentions, and hashtags is also murky. This is similar to the ambiguous meaning of a "Like" on Facebook (e.g., a friend's father dies and people click "Like" to acknowledge their support and attention, not because they literally *like* that the father died). To ascribe a single meaning to any of these behaviors masks the complexities of users' actual intentions and experiences. The ontology of native Twitter objects is subject to change without warning, and different data sources provide tweets in entirely different formats. It is important, then, for a common vocabulary to remain independent of any particular implementation of Twitter's features. The risk of taking the metadata provided by these services at face value becomes even more relevant in the development of data collection strategies.

Comparing Aggregation Approaches

A common descriptive terminology is one half of the methodological tool kit required to enable comparison among the many studies drawing on Twitter data. The second half is a shared set of expectations regarding the identification, aggregation, cleaning, and archiving of tweets. In spite of the considerable volume of published work drawing on Twitter data, misinformation abounds regarding the validity of different data collection strategies and their appropriateness for the analytic processes to which they are subjected. The negative consequences of this uncertainty have limited the development of this field: Some researchers abandon otherwise promising projects while others hesitate to get started.

The confusion springs from the opacity of Twitter's publicly accessible application programming interfaces, or APIs. These APIs provide external access to features of Twitter's software platform that would otherwise only be available to its employees. Public APIs are typically designed to encourage the development of third-party software—for example, a plugin for WordPress, or a client for the new BlackBerry phone. As with so many popular technologies, public APIs may also be adapted to the needs of researchers, and Twitter's APIs have proved particularly generative in this regard. With methods named "search" and "sample," Twitter's APIs initially seemed to provide researchers with a unique window into the inner workings of a mass-scale information system. Unfortunately, as many soon discovered, these API methods did not function as clearly or consistently as their research-friendly names suggested.

Twitter maintains three publicly accessible APIs: the Search API, REST API, and Streaming API. Each offers a different set of methods for interacting with the system, and each constrains the user in different ways. None of the publicly accessible APIs offers researchers the same degree of access that one of the company's own engineers might enjoy. Beginning in 2010, Twitter, Inc. partnered with a small number of firms to develop the growing flow of Twitter data as a commercial service (Gnip, 2010). The resulting products represent a new class of pay-as-you-go APIs, each with its own features and restrictions. In addition to the many different APIs currently in use, a number of earlier access options—such as the academic "whitelist" allowing specific researchers to be exempted from API limits—are no longer available.

Making sense of the many different Twitter APIs is challenging enough for an individual researcher, but the rapid pace of change confounds the slower, more deliberate tempo of academic publishing. Papers published in 2012 may well be referring to data collected in 2010 or earlier when the API offerings were very different. Perhaps attempting to avoid this problem, some papers opt for generic nomenclature such as "the Twitter API" or "the public API." Many of these papers also omit the procedure used to collect, process, analyze, and store the data. Unfortunately, this lack of specificity and detail limits the generalizability and rigor of these studies as they cannot be reliably replicated or compared with any other studies.

The need for experimental comparison among different APIs has been long noted in the informal discourse of conference Q&A sessions and online talk, but the published record remains quite small. In

2011, Pablo Rey Mazon organized a comparison of different data collection strategies and published the results on his website (Mazon, Morer, Um Amel, Lotan, 2012). Using each of the publicly accessible APIs along with two commercial platforms, Mazon and his collaborators attempted to aggregate tweets related to the Indignados social movement during a protest. Although Mazon's experiment was exploratory, the results clearly demonstrated significant differences among the various APIs that would fundamentally alter the outcome of any analytic inquiry.

With these differences in mind, a team from the Oxford Internet Institute conducted an experiment in 2012 to compare the tweets returned by similar queries of the publicly accessible Search and Streaming APIs (Gonzalez-Bailon, Wang, Rivero, Borge-Holthoefer, Moreno, 2012). According to Twitter's documentation of the two APIs, Search is "focused in relevance and not completeness" and "not all Tweets are indexed" which means that "some Tweets and users may be missing" from the results.[2] These clues indicate that the Search API will return fewer results than the Streaming API, but it does not indicate how the subset will be produced. Gonzalez-Bailon et al. aggregated tweets related to the Indignados from April 30, 2012 to May 30, 2012, and found that the Streaming API returned more than four times as many tweets as the Search API. Nearly all of the Search API results were found in the Streaming API results, with a few unexpected exceptions. In total, 2.5% of the tweets, 1% of the users, and 1.3% of the hashtags returned by the Search API were not found in the Streaming API results (Gonzalez-Bailon et al., 2012, p. 10). While these omissions could be due to interruptions in their connection to the API or other errors, the discrepancies provide further evidence that data from the public APIs are not complete and should not be considered the entirety of all public tweets matching the search criteria. After comparing the two data sets using social network analysis, the researchers determined that the Search API results skewed steeply toward central users and more clustered regions of the network. Conversely, peripheral users were less accurately represented and may have been absent altogether (Gonzalez-Bailon et al., pp. 14–15).

Crucially, Gonzalez-Bailon et al.'s study confirms that Search API results are not a random sample of overall Twitter activity. Rather, Twitter's internal software plays an editorial role in selecting and yielding tweets according to a set of heuristic algorithms that are not known to outside users. Delivered tweets are also subject to the limitations of the local cache on Twitter's servers: More popular tweets are kept in memory, and less popular tweets are archived. The logic underlying Twitter's Trending Topics system is similarly obscured, but because of its visibility, it has been a recurring point of contestation as users in precarious political situations accuse Twitter of "censoring" the Trending Topics system (Gillespie, 2011). Censorship may not be the most appropriate rubric for understanding the bias that exists in Search API results, but the protestations of users provides an important prompt for researchers who rely on commercial services for data. None of the available APIs provide an unfiltered, direct interface to Twitter's internal data store. To meet the needs of researchers, all such tools must be turned, to greater or less extent, away from the original purposes for which they were designed.

[2] For more documentation and discussion of the limitations of the Search API, see https://dev.twitter.com/docs/using-search and https://support.twitter.com/groups/32-something-s-not-working/topics/118-search-problems/articles/66018-i-m-missing-from-search

That Mazon and Gonzalez-Bailon et al. both use data related to the Indignados is useful for thinking about the very real theoretical implications of different data collection strategies. As Gonzalez-Bailon et al. demonstrate, the Search API suppresses evidence of legitimate peripheral participation in the Indignados protest activities. The API may also suppress artifacts of new participants until they become more central in the network. As a result, the artifacts of an emerging protest or the activities of users at the periphery may not be represented in data collected from the Search API.

While Gonzales-Bailon et al. demonstrate the shortcomings of the Search API relative to the Streaming API, they cannot conclude that the Streaming API is not itself similarly biased. The present research takes up this lingering question by comparing the output of the publicly accessible Streaming API with that of Gnip PowerTrack, a commercial service from one of Twitter's corporate partners.[3] Consistent with the previous literature, the present comparison is based on the use of Twitter for everyday political talk by the Occupy movement and viewers of the third U.S. presidential debate in 2012. Central to this comparison is a discussion of the technical resources required to interface with these high-volume data streams. These characteristics are then linked to broader issues of designing research around big social data, and the emerging "digital divide" in access to data (boyd & Crawford, 2011; Manovich, 2011). Although the core examples in this article are drawn from recent projects regarding the use of Twitter and the specific technologies under observation will inevitably be replaced, the implications of these observations are relevant for any research conducted in similarly hybrid public/private online spaces via publicly available APIs.

Comparing the Streaming API to Gnip PowerTrack

The Streaming API is a publicly accessible interface for third-party software programs to collect data from the Twitter platform. Pertinent to the comparison with Gnip PowerTrack is the "filter" method that provides external clients with a real-time stream of tweets matching a set of keyword filters (Walker, Hemsley, Eckert, Mason, & Nahon, 2013). The volume of these results is constrained by an undocumented upper limit known as the "streaming cap," which is believed to be up to 1% of the entire Twitter stream at any point in time.[4] When this limit is reached, the API sends a single message indicating a running total of the number of tweets that were not sent, or rate limited, since the most recent connection to the Twitter API was initiated. Twitter is strategically ambiguous about this constraint, and its documentation suggests that users who find themselves frequently being "rate limited" should consider purchasing a subscription to a commercial data provider.

[3] Gnip PowerTrack and the Streaming API are also common sources of data for analytics software packages. For example, the text-analysis service DiscoverText uses Gnip PowerTrack, and the open-source archiving tool yourTwapperKeeper uses the Streaming API. Understanding the limitations of these underlying data sources is especially important for users of prebuilt software.

[4] See https://dev.twitter.com/docs/faq#6861 and https://dev.twitter.com/discussions/6349 for more information on rate-limiting of the Streaming API.

Gnip is one of a handful of partner firms authorized to resell Twitter data. Similar to the Streaming API, PowerTrack provides a real-time stream of tweets matching a set of keyword rules. Marketing materials for the PowerTrack service emphasize its completeness, assuring the prospective customer that it "delivers full coverage . . . extracted from the full Twitter fire hose." In contrast to the barebones Streaming API, Gnip offers a variety of technical features, services, and support, but the most valuable aspect of the service is the absence of a data cap. Whereas the Streaming API will eventually hit this rate limit on high volume keywords (think "#superbowl"), PowerTrack promises "full" results.

Both Gnip and Twitter transmit tweet data in the machine-readable JSON format, but each structures the data according to a different ontology. The Streaming API yields what we might consider the "native" Twitter format, while Gnip output conforms to the Activity Streams specification.[5] In other words, the same tweet delivered by the Streaming API and Gnip PowerTrack will require different software to read, take up different amounts of space on the disc, and include different supplementary metadata. The supplementary metadata added by Gnip includes such features as expanded versions of any shortened URLs. Users of the Streaming API must manually add this supplementary information to their data sets. The distinction between these two formats makes visible the intermediary role played by—in this case—Twitter and Gnip in the production of trace data. These data are not "raw" but rather shaped according to unspoken criteria regarding what is valuable to know and how it ought to be categorized.

Both services also provide tweets based on a set of predetermined keyword filters. The Streaming API accepts up to 400 individual key terms and returns any tweet that contains those terms in its text, hashtags, @-mentions, or URLs. It is not possible to limit the matching criteria of the Streaming API, so any tweet containing one of the keywords in any of the four fields will be returned. PowerTrack employs a more fine-grained system for keyword matching that allows the construction of rules using Boolean logic operators and special platform-specific filters.[6] Gnip also provides multiple methods for managing these rules, including a user-friendly Web interface and a special set of API methods. Beyond convenience, these affordances enable Gnip users to more narrowly define their filtering rules in order to avoid erroneous matches. As a result, the Streaming API and Gnip PowerTrack will return different data sets with the same keywords. Comparing the resulting data sets is possible, but the differences in query construction between the two APIs must be taken into consideration.

The description of the Streaming API and Gnip PowerTrack as "real-time" refers to their method of delivering tweets to the client. In the typical search engine or database scenario, queries and responses are carried out through a series of discrete, asynchronous messages. The user types in a few keywords

[5] For more information about the Activity Stream specification, see http://activitystrea.ms.

[6] Words and phrases with divergent meanings across languages will have expected effects on keyword-based data collection procedures. For example, "#oo", a hashtag adopted by Occupy Oakland, frequently matches an unrelated set of tweets in Tagalog (Walker et al., 2013). Whereas users of the Streaming API will find themselves occasionally rate limited by these false matches, Gnip implements human language detection filters to reduce such collisions.

and waits for the system to respond; the system presents some possible matches and waits for the user's next request. Real-time streams, on the other hand, operate less like a card catalog and more like an automated coin-sorting machine. The user constructs a set of rules—for example, "tweets mentioning occupy"—and establishes a persistent network connection to the service. As long as that network connection remains open, the system will pass along any tweet sent by any user that matches the given criteria—for example, "Just arrived at the #occupy camp in Atlanta."

If the network connection is interrupted because the user's local machine crashes, is turned off, or the API disconnects, there will be a temporal gap in the final data set. For this reason, the demands on local infrastructure are much higher for real-time streams than for asynchronous systems. In an ideal scenario, one or more computers will be dedicated to the data collection process and remain connected to the Streaming or PowerTrack API 24 hours a day. Merely catching tweets as they come across the network connection is just one aspect of the local data apparatus, of course. Additional human and technology resources are required to monitor and manage the incoming stream, clean and categorize new tweets, and maintain an archive of past activity.

Methods

The Social Media (SoMe) Lab at the University of Washington aggregates tweets using the Streaming API with two virtual servers hosted by Amazon Web Services. At the end of each day, an automated process backs up the previous 24 hours of data and begins a new file to collect the incoming data. The backed-up tweets are then subjected to a battery of postprocessing tasks that produce a set of additional metadata for each tweet, including:

- The expanded version of any shortened URLs
- A lowercase list of hashtags in the tweet
- A lowercase list of @-mentions in the tweet
- A count of the number of hashtags, URLs, and mentions in the tweet
- A list of data collection keywords that matched within the tweet and the location of the match (hashtag, @-mention, text, or URL) in order to explain why each tweet made it into the archive

The original tweet, augmented by these metadata, is then inserted into a MongoDB database accessible to researchers in the lab. A Web interface enables less technical researchers to explore and export data in multiple formats that can imported into analysis software such as R and Gephi (Walker et al., 2013).

The Annenberg Innovation Lab at the University of Southern California maintains a similar apparatus for collecting tweets via Gnip PowerTrack. Four servers are housed in a building on campus and the hardware is maintained by university IT services. On one of these machines, a shell script uses the cURL command-line tool to maintain a connection with Gnip PowerTrack. Like the SoMe Lab, a second process automatically backs up the most recent 24 hours of data to the other three servers. Multiple research projects within the lab share the same PowerTrack connection so the 24-hour chunks are filtered a second time using a set of custom Python scripts and stored in project-specific databases.

The financial burden of Gnip PowerTrack places it beyond the reach of most researchers within the academy. A monthly subscription begins in the low thousands of dollars and scales up linearly with the volume of data being collected. To properly assess the cost of implementing either service, however, it is necessary to consider the local infrastructure required to support it. In both scenarios, a data management system was assembled and maintained by members of each lab. While the Streaming API is free of charge and Gnip PowerTrack may cost tens of thousands of dollars, neither service can feasibly be run at mass scale without significant computing resources and the cooperation and expertise of an interdisciplinary team.

Case 1: Rate Limiting During the October 22, 2012, U.S. Presidential Debate

To better understand rate limiting of the Twitter's public Streaming API, we collected tweets during an extremely popular political event—the third and final 2012 presidential debate between President Barack Obama and Governor Mitt Romney. Tweets with the hashtags #debate or #debates were simultaneously collected using both the Gnip PowerTrack API and the Twitter Streaming API from October 22, 2012, 19:00 EST to October 23, 2012, 0:30 EST (two hours before and after the debates). The filter criteria used for the Gnip PowerTrack API specified only tweets with the hashtags *#debate* or *#debates*. The Streaming API does not include a similar hashtag-matching feature, so the keywords "debate" and "debates" were used instead. To manage these more inclusive criteria, tweets returned by the Streaming API were parsed locally, and messages that did not include either of the two hashtags were discarded.

To ensure that there were no errors in the data collection, a random sample of 1,500 tweets was compared between the two data sets by matching the unique tweet ID, author's screen name, and timestamp of each tweet. After the successful verification of the two data sets, tweets in the Gnip data set were coded as to whether there was a matching tweet ID in the Streaming API data set. The results of this data collection, in 15-minute intervals, can be seen in Figure 1.

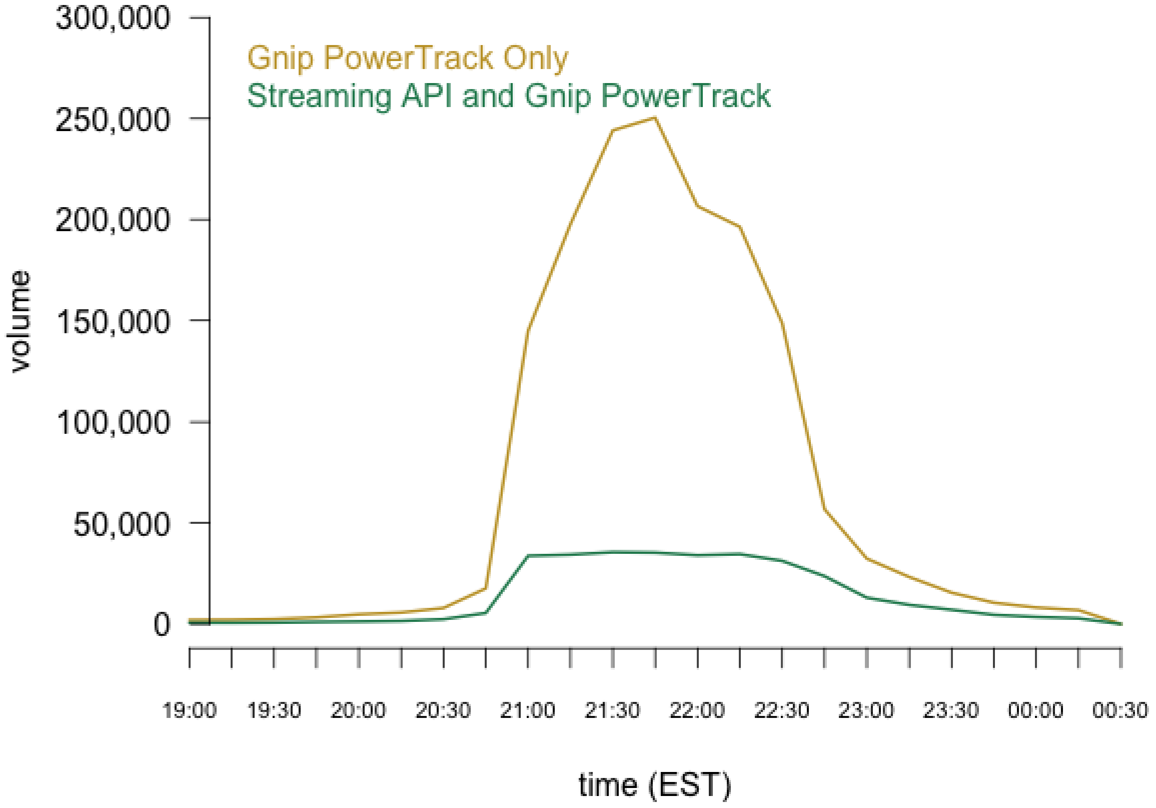

Figure 1. Comparison of tweets with the hashtags #debate and #debates collected via Gnip PowerTrack and Twitter's Streaming API.

Of the 1,588,392 tweets collected from the Gnip PowerTrack API, 1,271,730 (20%) were not found in the Streaming API data set. As seen in Figure 1 and Table 1, the Streaming API drops an increasing number of tweets during the two hours before the debates begin. Tweet activity spikes 45 minutes into the debates, with over 25,000 tweets a minute, and the rate-limited Streaming API is unable to keep up with this high volume and loses more than 22,400 tweets a minute. The loss of data continues well after the debates have finished.

One crucial feature of this comparison is the continued data loss that occurs after 23:00 due to the Streaming API’s generous keyword-matching function. While the Gnip PowerTrack API returns only those tweets with one of the two debate-related hashtags, the Streaming API returns any tweet containing the bare string “debate” or “debates.” Although the use of debate-related hashtags declines sharply at the end of the televised program, the number of tweets matching “debate” or “debates” falls more slowly. In spite of being immediately discarded, the tweets without hashtags nonetheless contribute to a sufficiently high volume of activity to trigger the rate-limiting function of the Streaming API (see Figure 2.) This subtle inefficiency illustrates the interdependence of the filtering functions and the rate-limiting system of the Streaming API.

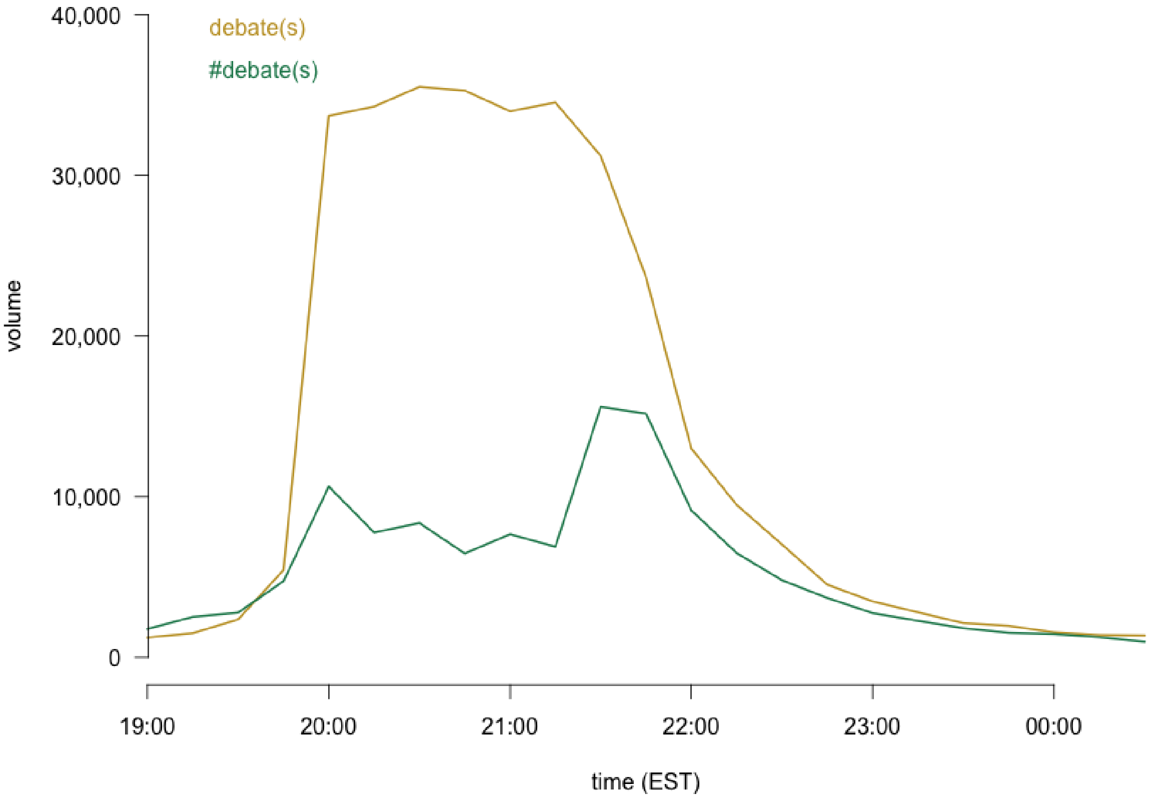

Figure 2. Comparison of tweets with the keywords “debate(s)” and “#debate(s)” collected via Twitter’s Streaming API.

The data loss observed in this analysis indicates that the Streaming API is not an appropriate tool for studies that require comprehensive collections of tweets concerning high-volume events or topics.

Table 1. Comparison of Tweets with the Hashtags #debate and #debates Collected via Gnip PowerTrack and Twitter's Streaming API.

	Gnip Only	*Streaming API & Gnip*	*Difference*
19:00	2,007	638	1,369
19:15	2,092	624	1,468
19:30	2,432	736	1,696
19:45	3,296	1,018	2,278
20:00	4,802	1,240	3,562
20:15	5,694	1,512	4,182
20:30	7,930	2,376	5,554
20:45	17,635	5,444	12,191
21:00	144,958	33,766	111,192
21:15	197,717	34,351	163,366
21:30	244,115	35,567	208,548
21:45	250,440	35,341	215,099
22:00	206,635	34,054	172,581
22:15	196,378	34,597	161,781
22:30	148,657	31,258	117,399
22:45	56,825	23,695	33,130
23:00	32,357	13,021	19,336
23:15	23,296	9,486	13,810
23:30	15,525	7,052	8,473
23:45	10,551	4,550	6,001
0:00	8,178	3,512	4,666
0:15	6,865	2,820	4,045

Case 2: Rate Limiting and Keyword Matching During the Occupy Wall Street Protests

To better understand the effect of rate limiting on a medium-volume, longer-term observation, we collected tweets during the Occupy Wall Street protests. Using the same method as the previous case study, tweets with the hashtags *#occupy* or *#ows* were simultaneously collected using both the Gnip PowerTrack API and the Twitter Streaming API over a 15-day period from November 8−22, 2011. As with the first case, the tweet selection procedure was different for the two APIs. The filter criteria used for the Gnip PowerTrack API specified only tweets with the hashtags *#occupy* or *#occupy*. A curated list of 109 popular hashtags, keywords, and Occupy city accounts related to the Occupy movement was used to collect data from the Streaming API. The keywords "occupy" and "ows" were included in that list. Tweets with the hashtags "occupy" and "ows" were extracted from the collected data. The two collections were successfully verified and compared using the same methods as the previous case study.

Of the 1,589,210 tweets collected from the Gnip PowerTrack API, 82,960 (5.2%) were not found in the Streaming API data set. As seen in Figure 3 and Table 2, the missing tweets from the Streaming API are concentrated around November 15, 2011. This date coincides with the eviction of the Occupy Wall Street protesters from Zuccotti Park in New York City, an anomalous high-volume event during an otherwise medium-volume observation. With the exception of November 15, the Streaming API continuously collected all of the tweets with either *#occupy* or *#ows* without being rate limited.

The initial outcome of this observation is that Gnip PowerTrack and the Streaming API return comparable collections of tweets for medium-volume events and topics that persist over longer periods of time.

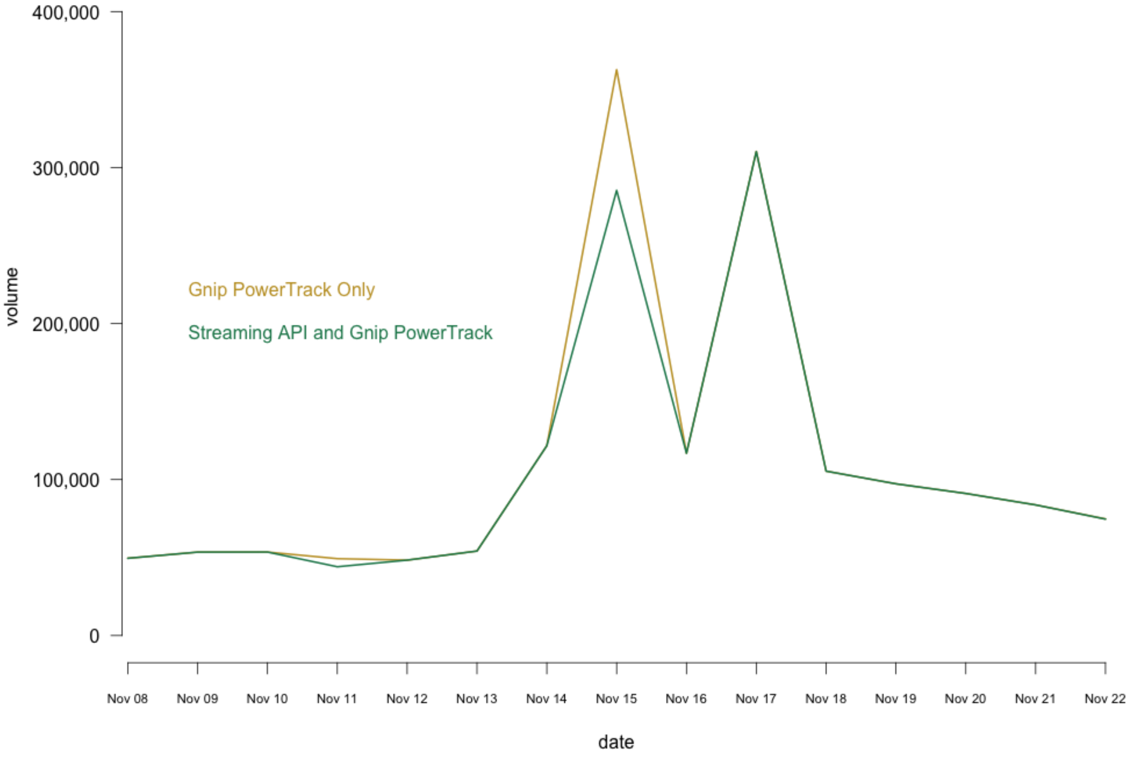

Figure 3. Comparison of tweets with the hashtags #ows and #occupy collected via Gnip PowerTrack and tweets with the keywords "ows" and "occupy" collected via Twitter's Streaming API.

Table 2. Comparison of tweets with the Hashtags #ows and #occupy Collected via Gnip PowerTrack and Twitter's Streaming API.

	Gnip Only	*Streaming API and Gnip*	*Difference*
11/8/2011	49,551	49,510	41
11/9/2011	53,472	53,469	3
11/10/2011	53,506	53,504	2
11/11/2011	49,276	44,081	5,195
11/12/2011	48,335	48,335	0
11/13/2011	54,158	54,157	1
11/14/2011	121,683	121,660	23
11/15/2011	362,753	285,390	77,363
11/16/2011	116,850	116,849	1
11/17/2011	310,370	310,194	176
11/18/2011	105,385	105,379	6
11/19/2011	97,261	97,256	5
11/20/2011	91,080	91,075	5
11/21/2011	83,738	83,730	8
11/22/2011	74,752	74,621	131

Discussion

The superficial conclusion of this study is simple and unsurprising: Gnip PowerTrack provides access to a greater volume of Twitter activity at considerable monetary cost. But the similarity of both services, particularly in the case of the Occupy movement, invites more critical evaluation of Twitter data collection strategies. Both the Streaming API and Gnip PowerTrack are real-time data streams that yield tweets based on a set of keyword filters. Each service requires significant local expertise and a sophisticated infrastructure for the aggregation, organization, and analysis of mass-scale data. Neither method offers retroactive historical data.[7]

The comparison of Gnip PowerTrack and the Streaming API has some unexpected implications for future research design. Principally, it reveals that costly commercial data providers are hardly a panacea for social media researchers. Although PowerTrack provides a comprehensive account of very high volume events such as the U.S. presidential debate, it does little to resolve the persistent lack of transparency in the production of social media data. Researchers must still approach the collected data with a critical eye and read prepackaged metadata against the grain.

The differences between the publicly accessible Streaming API and a private service like Gnip PowerTrack are not merely matters of cost and raw volume. Rather, the affordances of the two services respond differently to various rhythms of activity over time and incur different material burdens. To evaluate the appropriateness of either service to a given project, researchers should first be fully immersed in the flows of information they plan to analyze. Preliminary fieldwork should exceed casual use and may include such simple techniques as creating a new account dedicated to the project, acquiring different mobile devices, significantly reducing or increasing the number of accounts one follows, or checking in on the site at different times of day. Instrumentally, this work will enable researchers to determine a comprehensive set of key terms or phrases, estimate the duration of the planned observation, and anticipate the volume of tweets they expects to encounter.[8] Due to the ephemeral nature of Twitter, however, this preliminary fieldwork will inevitably be tempered by the urgency of setting the data collection process in motion. While too little fieldwork will lead to a noisy, unfocused data set, waiting too long to begin may result in irrecoverable data loss.

Observation Period

Identifying the duration of the observation period is important for designing an appropriate data collection apparatus. For an open-ended, transnational protest discourse like Occupy, it made sense to aggregate tweets on a similarly open-ended basis. In spite of the cost, round-the-clock data collection may capture emergent events that could not have been anticipated at the outset. Other phenomena may be adequately served with shorter, less costly periods of observation. For example, a project concerned with the use of Twitter as ambient media throughout the workday may not need to collect tweets in the middle of the night. Conversely, long-term observation remains one of the weakest areas in the field (boyd & Crawford, 2011, p. 4). Although the gap may be explained to some extent by the additional

[7] Third-party data providers Gnip and DataSift offer access to historical tweets on a limited basis, but the cost of these programs will exceed the resources of the majority of researchers.

[8] Preliminary fieldwork should also be used to learn about the privacy norms and expectations of different user populations. (See Zimmer, 2010 for a more thorough discussion of the ethical dimensions of privacy.) Accepting users' privacy settings at face value may not be sufficient, especially within default-public spaces such as Twitter.

technological challenges of collecting and interpreting longitudinal data, the lack of attention to historical inquiry reflects an emphasis on short-term return characteristic of commercial approaches to social media analytics. As the Occupy case demonstrates, long-term observations of lower-volume phenomena using the Streaming API are not only possible but may be less costly than short-term, high-visibility projects.

Keywords/Terms

All the data aggregation technologies discussed in this article depend on a list of keyword-based rules or filters. This architecture is reflected in the predominant use of hashtags in the design of Twitter studies. Bruns and Burgess note that while "hashtag-based approaches" (2012, p. 8) offer a clear point of entry into an unfolding live event, they yield a rather skewed picture of the overall discursive space. Tributaries of conversation often flow away from the dominant hashtag as users address one another directly using @-mentions. To access a more complete sense of the discursive scene, it is necessary to assemble a strategic set of terms and phrases informed by preliminary fieldwork.

An alternative to the single hashtag procedure is to approach keyword matching and filtering as a two-step process. In the first step, a promiscuous set of keywords is assembled that will yield a large volume of tweets. Next, researchers experimentally filter this big, noisy collection using a subset of the original keywords. By iterating on these two steps, the research team can determine a set of keywords that balances the cost of false matches with the benefit of capturing peripheral messages.

The list of keyword rules driving data collection should be as dynamic as the discourse it is designed to match. Whereas certain key terms will remain stable for the duration of a study—for example, "Obama," "Romney," or "#debate"—others will be impossible to anticipate, such as "binders" or "big bird." Different research projects will approach this dynamism differently, but it is important to plan ahead and track changes that are made to the keyword rules as they evolve. It is not possible to collect "all" of the tweets related to a given phenomenon, but keen participant observation paired with flexible rule management can mitigate egregious blind spots.

Although keyword-filtering is an unavoidable aspect of using the available APIs for data collection, it is not necessary to organize research around a set of key terms or hashtags. The keyword systems of both the Streaming API and Gnip PowerTrack can be used to match user names or substrings from within URLs. This means that a data collection scheme might be devised to follow a known network of users or to track the circulation of one or more unique links. In each case, preliminary fieldwork and playful engagement with the keyword filtering system may reveal new data collection strategies beyond simply following a hashtag.

One final note on the development of keyword lists concerns their interrelationship with the data aggregation apparatus. The cost of adding new rules is different for the Streaming API and Gnip PowerTrack, and the careless addition of an untested rule can be quite costly. For Gnip subscribers, false positives add to the monthly charges, and an errant keyword rule may unexpectedly match millions of tweets. The Streaming API, on the other hand, restricts the total number of tweets that may be aggregated in a given period of time. As demonstrated in the second case, false positives will crowd out desired matches during periods of high-volume activity.

On the "New" Digital Divide and Privacy

Mass-scale aggregation of digital traces requires material resources, institutional capital, and the coordination of researchers with a diversity of skills, expertise, and interest. Employees of data-rich organizations like Visa, AT&T, Facebook, or Google will have access to information and infrastructure that is simply beyond the reach of all but the most highly capitalized academic and state institutions. These conditions have made certain types of inquiry inaccessible to the majority of the research community, a situation that critics have described as a new "digital divide" (boyd & Crawford, 2011, p. 12; Manovich, 2011, pp. 2–5). Some see these problems as intractable, but the extent to which private industry excels at making sense of big social data is overstated (boyd & Crawford, 2011, p. 13, footnote 4). Deep engagement with local communication cultures, long-term historical analysis, and the production of platform-independent theory are just a few of the areas neglected by industrial data science.

It would seem that the problem of unequal access to data might be solved by sharing data openly among peer researchers at different institutions. From the start, this was a recurring topic of discussion among participants in the Occupy Research network. The terms of service for most publicly available APIs, specifically Twitter, forbid researchers from sharing data outside of their research teams. But amassing data is just one of the challenges of studying mass-scale information systems. Other researchers would not have been able to make use of the data without a local infrastructure and appropriate technical knowledge to manage and access it. Furthermore, radically open sharing of social media data may violate the privacy of individuals who find themselves unwillingly included. This risk is amplified in the context of social movement studies where the data may be used against the interests of activists and movement supporters.[9] Researchers with privileged access to mass-scale data should take extra precaution to guard against accidental exposure of user data, a standard which may exceed the expectations of their institutional review boards (Walker et al., 2013; Zimmer, 2010).

Conclusion

Researchers in the humanities and social sciences are not accustomed to thinking of their work in terms of laboratory science—we are not technicians in white coats, pipetting solutions and peering into microscopes, after all! But mass-scale observation is possible only with the assistance of specialized technologies. As such, the sociology of science has come home in surprising ways. The infrastructures we construct to aggregate and store tweets—servers, scripts, and databases—mirror the internal architectures of commercial systems such as Twitter. The analyses we perform, then, are carried out against this mirror, not "Twitter" itself. The conclusions we draw from this work are, in part, mediated by the contours of our local data management systems. There are many different interfaces to Twitter—Web

[9] The risks here are not hypothetical, as demonstrated by the subpoena of Occupy activists' Twitter history by police. See Williams (2012).

interfaces, cell phone applications, and so on—but the mirror worlds we create are unique. The picture of Twitter visible in the data we collect will differ from the day-to-day experience of any human user.

At first blush, Gnip PowerTrack and the Streaming API seem to enable the construction of very different types of mirrors. As we have shown, the Streaming API excels at longitudinal data collection, but is a poor choice for massive, short-term events. PowerTrack, on the other hand, offers very large collections of tweets sent within short periods of time, but is extremely costly in the long term. However, the two access methods are more alike than they are different and having access to the "fire hose" does not necessarily enable more meaningful research design. Both APIs require researchers to organize their data collection strategies in terms of keywords, an architecture that is reflected in the dominance of hashtag-based data collection in the field. Curiously, while many studies focus on short-term events, few researchers have explored the opportunity for longitudinal data collection that the publicly accessible Streaming API makes possible.

References

Black, J. D. F. (Writer), Roddenberry, G. (Writer), & Daniels, M. (Director) (1966). The naked time [Television series episode]. In Roddenberry, G. (Producer), *Star trek*. Culver City, CA: Desilu Studios.

boyd, d., & Crawford, K. (2011, September). Six provocations for big data. Paper presented at the Oxford Internet Institute's A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society, University of Oxford, Oxford, UK.

Bruns, A., & Burgess, J. E. (2011a, September). New methodologies for researching news discussion on Twitter. Paper presented at the Future of Journalism conference, Cardiff University, Cardiff, UK.

Bruns, A., & Burgess, J. E. (2011b, August). The use of Twitter hashtags in the formation of ad hoc publics. Paper presented at the 6th European Consortium for Political Research General Conference, University of Iceland, Reykjavik.

Bruns, A., & Burgess, J. (2012, August). Notes towards the Scientific Study of Public Communication on Twitter. Keynote presented at the Conference on Science and the Internet, Düsseldorf, Germany.

Gillespie, T. (2011, October 19). Can an algorithm be wrong? Twitter Trends, the specter of censorship, and our faith in the algorithms around us. *Culture Digitally*. Retrieved from http://culturedigitally.org/2011/10/can-an-algorithm-be-wrong

Gnip. (2010, November 17). Commercial Twitter data now available through Gnip. Retrieved from http://gnip.com/pr_announcing_commercial_twitter_data

Gonzalez-Bailon, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2012, December 4). Assessing the bias in communication networks sampled from Twitter. Retrieved from http://dx.doi.org/10.2139/ssrn.2185134

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 767–797.

Manovich, L. (2011). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 460–475). Minneapolis: University of Minnesota Press.

Mazon, P. R., Morer, I., Um Amel, V. J., & Lotan, G. (2012, May). #12m15m Twitter archive experiment. Retrieved from http://numeroteca.org/12m15m

Twitter (2009, August 13). Project Retweet: Phase one [Web log post]. Retrieved from http://blog.twitter.com/2009/08/project-retweet-phase-one.html

Walker, S., Hemsley, J., Eckert, J., Mason, R. M., & Nahon, K. (2013). Tools for social media research. *iConference 2013 Proceedings* (p. 971). doi:10.9776/13496

Williams, M. (2012, September 14). Twitter complies with prosecutors to surrender Occupy activist's tweets*. The Guardian*. Retrieved from http://www.guardian.co.uk/technology/2012/sep/14/twitter-complies-occupy-activist-tweets

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics of Information Technology*, *12*, 313–325. doi:10.1007/s10676-010-9227-5