We Don't Know What We Don't Know:

When and How the Use of Twitter's Public APIs Biases Scientific Inference

Rebekah Tromble
Corresponding author
Leiden University
r.k.tromble@fsw.leidenuniv.nl

Andreas Storz
Leiden University

Daniela Stockmann
Hertie School of Governance

WORKING PAPER
Comments welcome

November 29, 2017

Abstract: Though Twitter research has proliferated, no standards for data collection have crystallized. When using keyword queries, the most common data sources—the Search and Streaming APIs—rarely return the full population of tweets, and scholars do not know whether their data constitute a representative sample. This paper seeks to provide the most comprehensive look to-date at the potential biases that may result. Employing data derived from four identical keyword queries to the Firehose (which provides the full population of tweets but is cost-prohibitive), Streaming, and Search APIs, we use Kendall's-tau and logit regression analyses to understand the differences in the datasets, including what user and content characteristics make a tweet more or less likely to appear in sampled results. We find that there are indeed systematic differences that are likely to bias scholars' findings in almost all datasets we examine, and we recommend significant caution in future Twitter research.

Keywords: Twitter, data collection, APIs, bias

Twitter research has proliferated across academic disciplines (Bruns & Weller, 2014): from sociology to computer science, linguistics to political science. And yet no standards for the most basic function of Twitter research—data collection—have crystallized within most disciplines, let alone across the entire array of scholarly fields (Zimmer & Proferes, 2014). While some researchers write their own programs to query Twitter's two freely-accessible application programming interfaces (APIs), others employ third-party software that returns data from one of these APIs. Those with large budgets may purchase Twitter data from a vendor with access to either the entire population of Twitter data in real time—i.e., via the "Firehose"—or Twitter's historical archive.[1] Still others elect to scrape data via the web.

Each of these collection approaches has important implications for scholarly analysis. In the vast majority of cases, the researcher ends up with a sample of Twitter data but, due to proprietary limitations, has no way of assessing how representative that sample is. Indeed, when gathering data based on keyword searches (e.g., hashtags or @mentions), the only approach that (effectively) ensures one receives the full population of tweets is to pay a vendor to extract them in real time from the Firehose.[2] This is very costly and is beyond most researchers' means. Thus, in most cases, scholars cannot be confident that the conclusions they draw from the corpus of Twitter data at their disposal are sound, their inferences unbiased.

This paper seeks to provide the, to date, most comprehensive and in-depth look at the biases likely to be generated when collecting data via either of Twitter's cost-free application program interfaces: the Streaming and Search APIs. While several studies have compared data

[1] On November 14, 2017, Twitter announced it would be introducing new "premium APIs" with scalable access and pricing. At the time of writing, details about these new APIs were vague, with no specific information offered regarding costs or rate limits. See https://blog.twitter.com/developer/en_us/topics/tools/2017/introducing-twitter-premium-apis.html.

[2] Because tweets are removed from the historical archive whenever they are deleted, made private, or a user's account is suspended, even purchasing Twitter data from the archive will not return the full population of tweets.

collected via the Firehose to that generated by identical keyword queries of the Streaming API (Driscoll & Walker, 2014; Morstatter et al, 2013, 2014), to the best of our knowledge, this is the first research to compare the Firehose, Streaming, and Search APIs to one another.

How much of the total data are sampled from the free APIs? And do these data differ systematically from the full population of tweets?

We answer these questions by examining the results of three separate keyword queries across all three collection sources: #jointsession, #ahca, and #fomc. Each hashtag corresponds to a different political event in the United States during the first half of 2017. In order to provide additional insights into rate limits applied to the Streaming API, we also examine the results of a keyword query run during Donald Trump's January 20, 2017 inauguration. In this instance, we queried the PowerTrack and Streaming API for all tweets mentioning @realdonaldtrump. We use Kendall's-tau to assess how closely correlated the results are to one another across the different collection sources and then run logit regression analyses that allow us to explore what user and content characteristics make a tweet more or less likely to appear in either Streaming or Search results.

Our findings suggest that bias is extremely likely in the case of Search API results, as well as rate-limited Streaming API data. And we urge that scholars using data obtained via keyword queries of either source exercise significant caution when drawing inferences from such data.

The Twitter APIs

Twitter offers a number of options for members of the public to gather data from the platform. Researchers can purchase access to the Firehose, Twitter's real-time flow of all new tweets and

their related metadata, and those seeking historical data can purchase all relevant public, undeleted tweets from Twitter's archive. However, both of these options can be extremely cost prohibitive. Thus, academic researchers have largely turned to Twitter's free services, its public APIs.

The first of these, Twitter's Streaming API, provides tweets in real-time and can be queried using keyword, user ID, and geolocation parameters. When undertaking keyword queries of tweet content, the Streaming API will match keywords in the body of a tweet, the body of quoted tweets, URLs, hashtags, and @mentions. Twitter's documentation suggests that the Stream can return "up to" 100% of all tweets meeting one's query criteria, as long as the relevant tweets constitute less than 1% of the global volume of tweets at any given moment. When that 1% threshold is reached, the API begins imposing rate limits. Twitter's current global volume averages 6,000 tweets per second. However, this figure fluctuates significantly from day to day, hour to hour, and even minute to minute, driven in large part by unpredictable external events such as natural disasters, terrorist attacks, and other shocking or controversial news items. The API does provide rate limit messages, allowing a user to know when rate limits have been imposed, but these do not indicate what types of messages are missing—that is, if there is a systematic character to the tweets that are not captured. Moreover, Twitter's documentation does not suggest that 100% of tweets necessarily will be provided, even when no rate limits are imposed.

The second common source of Twitter data, the Search API, is a component of the larger REST API and may be used to query historical data. This option is clearly advantageous for collecting data on issues or events that cannot be easily predicted in advance. However, Search API queries carry significant limitations. To begin, the Search API only reaches back ~7-10

days. Second, each call to the API can return a maximum of just 18,000 tweets, and Twitter limits users to 180 calls every 15 minutes. In addition, queries to the Search API provide matches only to the main text of a tweet. Finally, and perhaps most importantly, Twitter makes it clear that Search results are non-random. The company states that queries return "top" content, not all relevant tweets, but Twitter does not clarify what constitutes "top" content.

Previous Research

In practice, then, whether using the Streaming or Search API, researchers collecting Twitter data via keyword queries are left with a great deal of uncertainty regarding how much and what types of tweets they may be capturing at any given moment in time.

With this in mind, earlier work has attempted to gain a better understanding of the nature of these two APIs. González-Bailón et al (2014), for instance, compared social network metrics generated from datasets obtained via keyword queries of the Search and Streaming APIs and found that, compared to the Streaming results, the Search API tended to underestimate centrality scores. They did not, however, include data from the Firehose, making it difficult to assess whether the Streaming API metrics were truly more representative than those from the Search.

Driscoll and Walker (2014), in contrast, compared the volume of results generated by two keyword queries to the Firehose and Streaming APIs. In both instances, the Streaming API data were rate-limited. In the first case, which comprised tweets related to the final 2012 presidential debate between Barack Obama and Mitt Romney, Driscoll and Walker reported that 20% of the Firehose data were missing from the Streaming capture. In the second case—a 15-day capture of tweets concerning the 2011 "Occupy" protests—5.2% were not captured by the Streaming API.

Also interested in the Streaming API, Morstatter and colleagues (2013) collected data related to the conflict in Syria from the Firehose and Streaming APIs over a 28-day period from mid-December 2011 to early January 2012. They used a number of hashtag keyword parameters (e.g., #syria, #assad, #homs) but also collected tweets geotagged as originating from within Syria, as well as those from the user @SyrianRevo. Although they do not report whether any rate limit messages were received via the Streaming API, the researchers did not obtain full data on any of the 28 days in question. Morstatter et al compared the Firehose, Streaming API, and random samples drawn from the former in terms of the top hashtags produced by and topics extracted from each. In both instances they find that the Streaming API data provide poorer approximations of the full-population Firehose results than do randomly sampled data, suggesting biased inferences are likely to result from use of sampled data collected via the Streaming API.

These are important initial findings. However, they also leave several key gaps in our knowledge. First, these studies do not compare the Search API to the Firehose. Second, because Twitter's API algorithms can and do change on a regular basis and because the overall volume of tweets has increased dramatically in recent years, new research is needed to update our understanding of even the Streaming API's functioning. Third, though Morstatter et al (2013) find that sampled data captured via the Streaming API likely are biased, we do not yet know whether Twitter uses any criteria related to the characteristics of the tweets themselves when imposing rate limits on the Streaming API. Nor, of course, do we know what criteria Twitter might use to designate "top" tweets for the Search API. If we can uncover certain patterns that differentiate tweets sampled from either the Streaming or Search APIs from non-sampled tweets, researchers would have a better grasp of the biases they are likely to introduce into their work.

Methodology

Data Collection

In order to begin exploring these potential biases, we collected four sets of keyword query data: #jointsession, #ahca, #fomc, and @realdonaldtrump. The first, #jointsession, was captured during Donald Trump's address to a Joint Session of Congress on February 28, 2017. Twitter live-streamed the speech and officially promoted use of #jointsession. Our 3-hour PowerTrack query of the Firehose API captured 569,015 tweets containing the hashtag. The second query, #ahca, was conducted on March 24, 2017 and corresponded with the initial failed effort by Republican leadership to bring a healthcare reform bill to a vote in the US House of Representatives. American mass media outlets provided nearly non-stop coverage in the moments leading up to, during, and just after the bill was pulled from the floor of the House, and over a 7.5-hour period, the PowerTrack returned 122,869 tweets matching #ahca. The final hashtag query, #fomc, relates to the US Federal Open Market Committee, which reviews and executes monetary policy for the US Federal Reserve. We captured tweets containing this hashtag for a 1.5-hour period during the June 14, 2017 press conference at which the committee announced a rise in interest rates. The PowerTrack query returned 1,061 tweets containing #fomc. Finally, we captured tweets containing @realdonaldtrump for 12 hours surrounding Donald Trump's January 20, 2017 inauguration. The PowerTrack query captured 1,285,922 relevant tweets.

Due to a technical glitch, we have data from the Streaming API but not from the Search API for the inauguration. However, for all three other events, we collected data from the PowerTrack, Streaming, and Search APIs. Access to the PowerTrack was provided by the DiscoverText platform. We used the R library StreamR to access the Streaming API and the

Python library Tweepy to access the Search API. In all instances, we captured simultaneously from the PowerTrack and Streaming APIs and launched the historical query of the Search API immediately after ending the real-time captures. The connection to the Search API was automatically terminated after 500 consecutive calls returned no new results.

Data Analysis

Kendall's-tau

For those interested in examining trends in Twitter activity, obtaining sample data that preserve the appropriate rank order of content such as hashtags, @mentions, and users is crucial. However, the rank order of content is frequently important for researchers studying other aspects of Twitter as well. Given the size of most Twitter datasets, many scholars choose to more closely interrogate their data by examining smaller subsets, especially the most prominent observations, and for those providing visualizations of their data—network graphs, for example—top attribute information offers simple means for determining cut-points for inclusion in an illustration. We therefore began our analysis by comparing Kendall's-tau correlations across each of the APIs for the top hashtags, @mentions, and usernames (i.e., the account names of those who sent tweets) for our four events. Kendall's-tau gauges the ordinal association, or similarity of the rankings, between two lists. Thus, the lower the Kendall's-tau score, the more likely one is to misidentify the most prominent actors or hashtags when using the requisite Streaming or Search API data.

In each instance, we calculated the Kendall's-tau score in steps for the top 10, 25, 50, 100, 250, 500, and 1000 hashtags, mentions, and usernames. After ranking the top observations in the PowerTrack (from highest to lowest), we determined the corresponding rank in the comparison case (Streaming API or Search API) for each observation. When a particular

observation did not appear at all in the comparison API, it received the lowest available rank plus one. The Kendall's-tau score was then calculated for the two resulting rank-ordered lists. In order to account for possible ties in rank, we used Kendall's-tau-b for all calculations.

Following Morstatter et al (2013), we then compare these results to Kendall's-tau estimates calculated from repeated random samples drawn from the PowerTrack datasets. For each Twitter event we generated 100 random samples from the PowerTrack of the size of the comparison (Streaming or Search) API. As Twitter APIs are intended to produce non-duplicate data, we performed resampling *without* replacement,. For each of the seven steps (top 10, 25, 50, etc.) we then calculated the corresponding Kendall's-tau-b scores for the three features (hashtags, @mentions, usernames) from each of the samples in comparison to the full PowerTrack dataset. At each step, this process yielded 100 estimates from which we calculated 95% confidence intervals to approximate a distribution of Kendall's-tau scores under conditions of random sampling. Our Kendall's-tau-b scores calculated from the datasets collected via the Streaming and Search APIs could then be compared to these confidence intervals to assess how closely they resemble random samples.

Logit regression

While our Kendall's-tau analysis provides an overall sense of whether bias is likely to appear in the Streaming and Search API data, it does not indicate how that bias might manifest. In order to better understand what parameters Twitter may use to sample data for the Streaming and Search APIs, we therefore employ a series of logit regression models. The dependent variable for each model is a binary variable indicating whether a tweet that appeared in the PowerTrack results

also appeared in the relevant Streaming or Search captures (1=captured, 0=not captured). Table 1 provides a summary of the predictor variables that were included the regression models.

Our variable choice was driven by our own reasoning regarding what would make some tweets "better" or more "valuable" than others from Twitter's business (i.e., monetary and reputational) standpoint. Following Twitter's own language regarding the Search API, each of these features seems to offer a plausible indicator of a tweet's potential for "top" status, but such features might also factor into sampling in the Streaming API.

Table 1: Logit Regression Predictor Variables

	Variable	**Description**
Tweet Characteristics	Retweet	Binary variable, retweet=1, not retweet=0
	Quote tweet†	Binary variable, quote tweet=1, not quote tweet=0
	Reply	Binary variable, reply=1, not reply=0
	Media	Binary variable, contains media (e.g., picture, video, hyperlink)=1, does not contain media=0
	Mention count	Number of @mentions included in the tweet
	Hashtag count	Number of #hashtags included in the tweet
User Characteristics	Verified	Binary variable, verified account=1, not verified account=0
	List count†	Number of Twitter lists on which the user appears
	Status count	Number of tweets the user has posted
	User like count	Number of tweets the user has liked
	Follower count	Number of followers the user has
	Following count	Number of others the user follows

† As a result of changes to PowerTrack features over time, these variables were not available in our earliest dataset, @realdonaldtrump.

The first three tweet characteristic variables speak to the "originality" of a tweet, with quote tweets and reply tweets also highlighting the conversational features of Twitter. Tweets containing external media point to the richness of a tweet's content, and mention and hashtag counts offer an indication of how engaged a tweet is with the larger Twitter community.

Given the scrutiny that Twitter has faced over bot activity, as well as the business incentives in place to promote tweets from popular users, we also investigated whether user characteristics factored into the API samples. Whether the user's account is verified provides a measure of two things: prestige and authenticity. The number of lists on which a user appears offers a sense of prestige and influence. This is also true of a user's follower count, while status count, user like count, and following count all point to how prolific and engaged the user is.

In each of our regression models, the PowerTrack served as the baseline. We only ran the statistical models in cases where there is a considerable discrepancy in coverage between the PowerTrack and the comparison API. For the Streaming API, this was the case for the two events of #jointsession and @realdonaldtrump. The Search API, on the other hand, produced a sub-sample in all cases, i.e. #fomc, #ahca, and #jointsession. Therefore, we calculated five models in total. In each case, we exponentiated the resulting coefficients in order to generate more interpretable odds ratios.

Findings

Volume Captured

Table 2 presents the volume of data captured by each query, trimmed by timestamp such that the start and end times are consistent across the three APIs. Table 3 in turn offers the percentage of unique tweet IDs in the PowerTrack results that appear in the Streaming and Search results, respectively. Note that this is not a simple comparison of the volume of data captured, but an ID match. Due to technical differences and glitches it is possible—and indeed occurred in a very small number of instances—that a tweet is captured in either the Streaming or Search datasets but does not appear in the PowerTrack

Table 2: Number of tweets captured

	@realdonaldtrump	#ahca	#fomc	#jointsession
PowerTrack	1,285,922	122,869	1,061	569,015
Streaming	872,458	122,935	1,067	368,146
Search	--	82,318	900	375,951

Table 3: Percent tweet ID matches

	@realdonaldtrump	#ahca	#fomc	#jointsession
Streaming	66.72	100	99.72	64.82
Search	--	66.99	84.54	66.06

When not rate-limited (i.e., for #ahca and #fomc), the Streaming API effectively captured all relevant tweets, but when rate limits were imposed (#jointsession and @realdonaldtrump), only about two-thirds of the data were captured. The figures are approximately the same for the two higher-volume Search API queries, with roughly two-thirds of the #jointsession and #ahca data returned. Even with the low volume event, however, the Search API did not obtain completeness. Instead, it retrieved 84.5% of the tweets containing #fomc.

Kendall's-tau

Tables 4-6 offer the results of the Kendall's-tau correlations for the ranked lists of top hashtags, mentions, and usernames, respectively. Following Wang and colleagues (2012), we presume that any bias introduced by very low levels of error, 0.0500 or less, is likely to be trivial but above this level, bias is likely to have a more substantial impact on one's findings. The non-rate-limited Streaming API results are generally below or very near the 0.0500 threshold (i.e., the correlation statistics are near or above 0.9500). However, for the rate-limited #jointsession and @realdonaldtrump data, the correlations are above this mark in all but two instances: the top 10 #jointsession hashtags and mentions. And the levels of error are particularly high for usernames. For @realdonaldtrump, the top 10 usernames produces a correlation of -0.0449. The username correlations are also extremely low for the #ahca and #jointsession Search API results, but even

#fomc correlations are poor. The error in the correlation of top hashtags for #fomc is consistently low. However, in all other instances, error is well above the 0.0500 threshold for the Search API.

Table 4: Kendall's-tau correlations of top hashtags

Top	Stream realdonaldtrump	Stream ahca	Search ahca	Stream fomc	Search fomc	Stream jointsession	Search jointsession
10	0.8222	1.0000	0.7333	1.0000	1.0000	0.9556	0.8667
25	0.9267	0.9467	0.7800	0.9770	0.9447	0.9000	0.8267
50	0.9404	0.9792	0.8402	0.9808	0.9386	0.8608	0.7002
100	0.8923	0.9782	0.8221	0.9897	NA	0.8125	0.6846
250	0.8586	0.9606	0.7794	NA	NA	0.8337	0.6830
500	0.8569	0.9500	0.7606	NA	NA	0.8336	0.6566
1000	0.8593	0.9536	0.7365	NA	NA	0.8435	0.6535

Table 5: Kendall's-tau correlations of top mentions

Top	Stream realdonaldtrump	Stream ahca	Search ahca	Stream fomc	Search fomc	Stream jointsession	Search jointsession
10	0.8667	1.0000	0.8667	1.0000	0.8098	0.9556	0.6889
25	0.8267	1.0000	0.6467	0.9876	0.7580	0.9267	0.8200
50	0.8624	1.0000	0.7040	0.9968	0.7711	0.7971	0.7192
100	0.9021	0.9965	0.6637	NA	NA	0.8310	0.5509
250	0.8836	0.9892	0.5513	NA	NA	0.8182	0.6169
500	0.8821	0.9887	0.5328	NA	NA	0.8511	0.5990
1000	0.8839	0.9922	0.5801	NA	NA	0.8504	0.5642

Table 6: Kendall's-tau correlations of top usernames

Top	Stream realdonaldtrump	Stream ahca	Search ahca	Stream fomc	Search fomc	Stream jointsession	Search jointsession
10	-0.0449	1.0000	0.2608	1.0000	0.7342	0.6591	-0.1648
25	0.2609	1.0000	0.1681	0.9802	0.7571	0.5859	0.2471
50	0.3520	1.0000	0.1402	0.9874	0.6760	0.5580	0.2827
100	0.4616	1.0000	0.0218	0.9795	0.6760	0.5093	0.2707
250	0.4797	0.9994	0.1451	NA	NA	0.5668	0.3750
500	0.5677	0.9984	0.2711	NA	NA	0.5414	0.4095
1000	0.5981	0.9993	0.2616	NA	NA	0.5899	0.4356

When comparing the Kendall's-tau scores obtained from the Search and rate-limited Streaming APIs with the results generated by 100 repeated random samples drawn from each PowerTrack dataset, concerns about bias only grow. These findings, including the 95% confidence intervals associated with the repeated random samples, are displayed in Figures 1-5. With one exception—the #fomc Search results—each of the API samples consistently under-performs the random samples. Aside from the #fomc Search data, just two Kendall's-tau scores—the top 10 #jointsession Streaming hashtags and mentions—fall within the 95% confidence intervals generated by the random samples.

Figure 1. Kendall's-tau correlations, @realdonaltrump Streaming results compared to 95% confidence intervals generated by repeated random samples

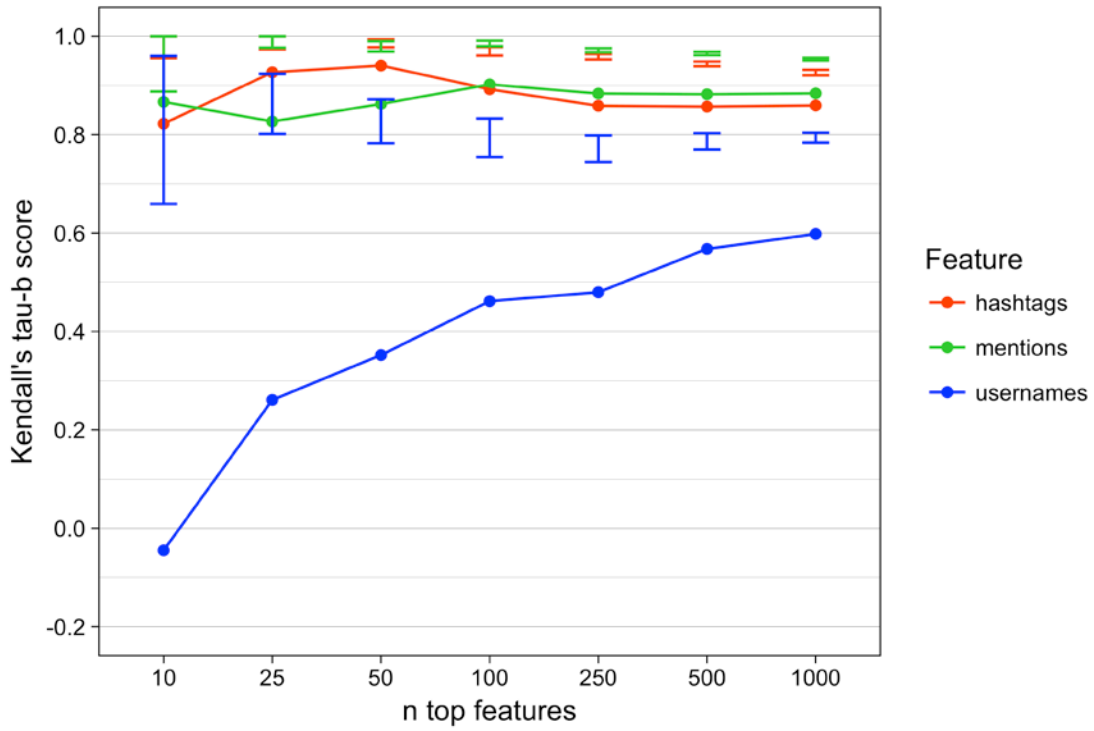


Figure 2. Kendall's-tau correlations, #jointsession Streaming results compared to 95% confidence intervals generated by repeated random samples

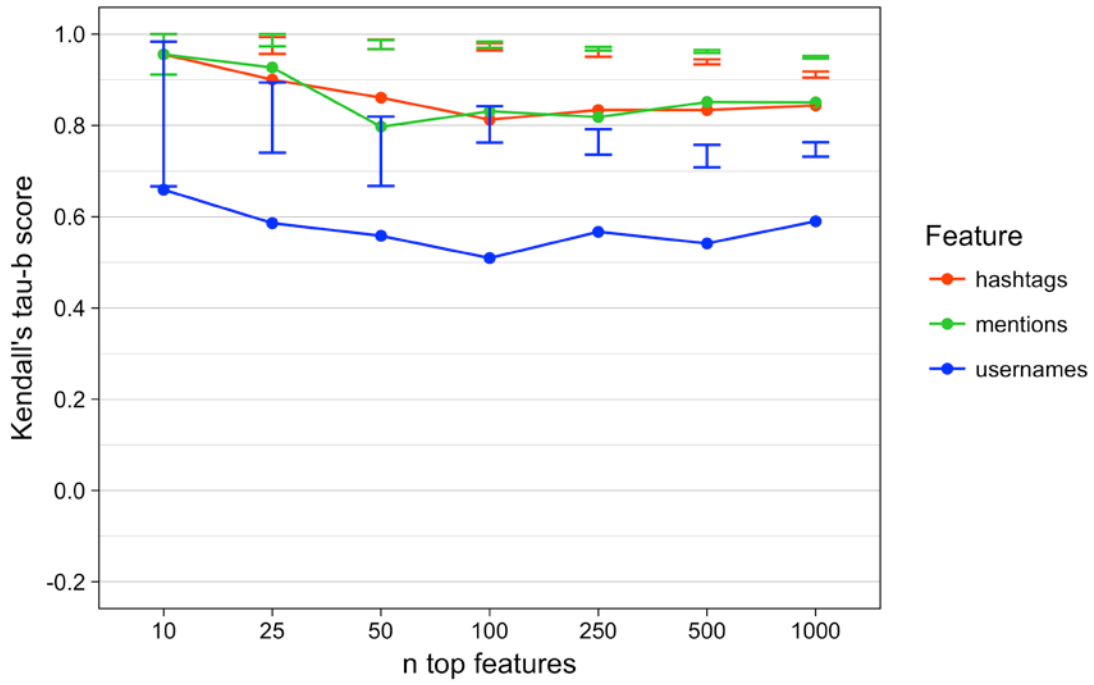


Figure 3. Kendall's-tau correlations, #ahca Search results compared to 95% confidence intervals generated by repeated random samples

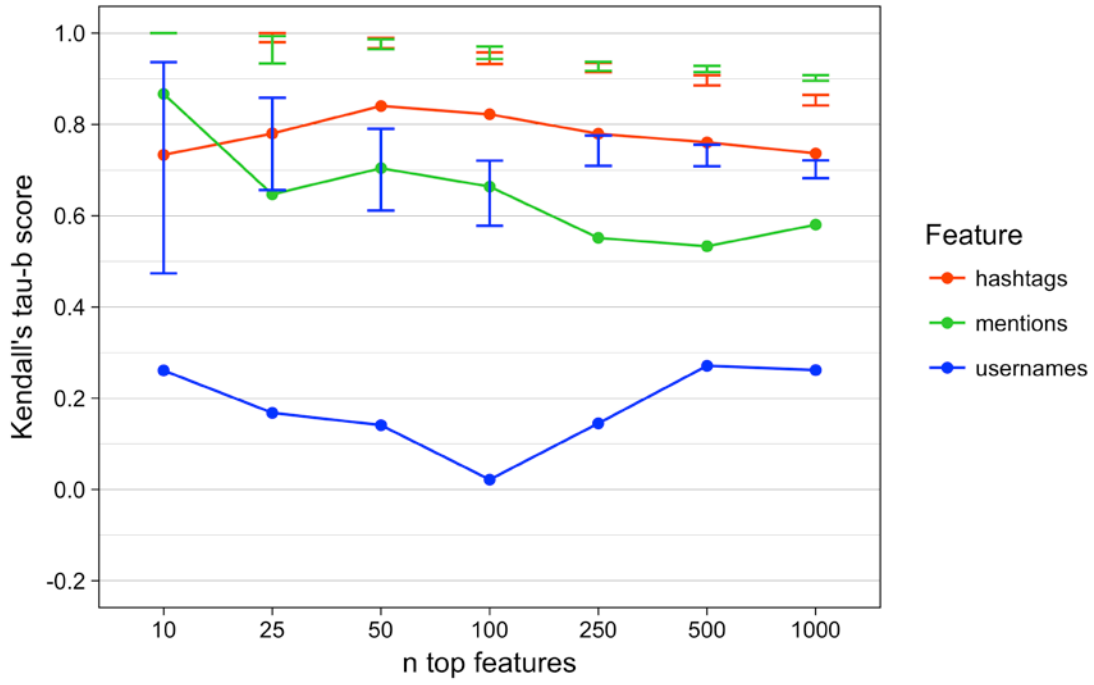


Figure 4. Kendall's-tau correlations, #jointsession Search results compared to 95% confidence intervals generated by repeated random samples

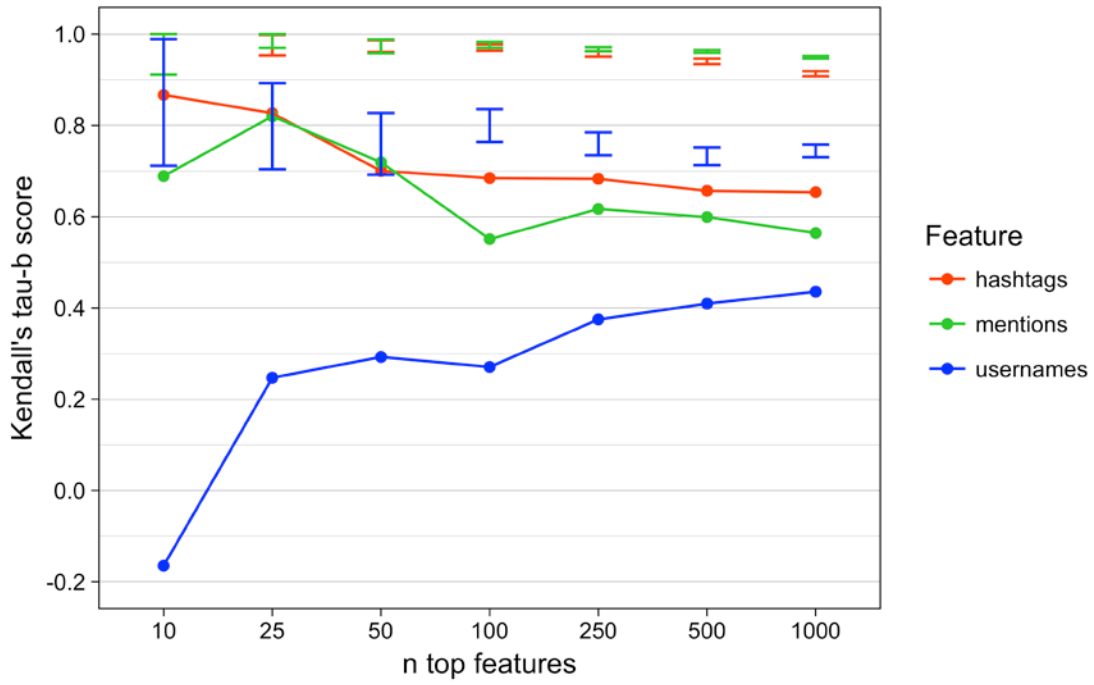
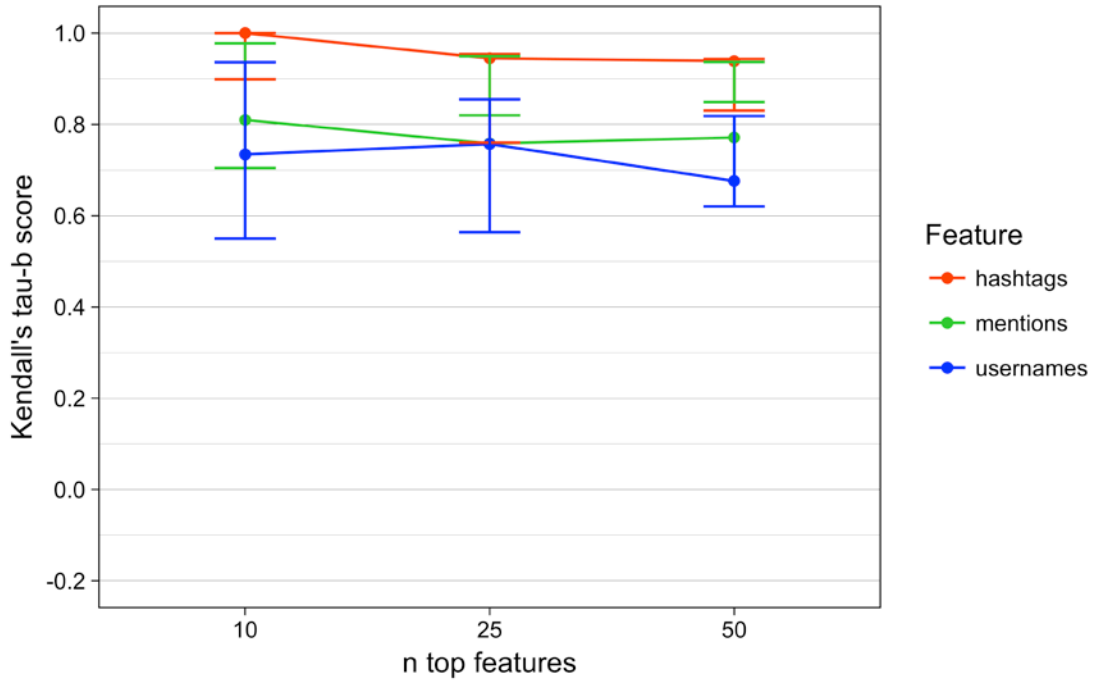


Figure 5. Kendall's-tau correlations, #fomc Search results compared to 95% confidence intervals generated by repeated random samples



Logit Analyses

Table 7 displays the results of the logit models for the two rate-limited Streaming API queries. The results prove consistent in terms of statistical significance and coefficient direction across the two datasets for seven of the ten shared variables. Among the user characteristic variables, the number of tweets a user has liked, the number of accounts they follow, and the number of tweets they have posted are all positively correlated with the likelihood of appearing in the Streaming API data. Tweets originating from verified accounts, on the other hand, are less likely to be captured by the Streaming API. Considering tweet characteristic variables, the presence of media, higher hashtag counts, as well as higher mention counts are all associated with a greater likelihood of Streaming API capture.

However, there are also some intriguing differences between the two datasets. In both cases, the coefficient for retweets is statistically significant, but it is positive for @realdonaldtrump and negative for #jointsession. In the #jointsession data, the coefficient for reply tweets is positive but not statistically significant, while in the @realdonaldtrump data, the coefficient is significant at the 0.000 level and is negative. If we turn to the odds ratios (Table 8), we see that the odds of replies appearing in the @realdonaldtrump Streaming capture are almost 50% lower than non-reply tweets. For retweets, the odds of appearing in the #jointsession Streaming data increase by 42.2% but decrease in the @realdonaldtrump capture by a full 73.8%.

The odds ratios point to a number of other interesting results as well. The effect sizes for media and mention count are particularly high in the @realdonaldtrump results, with an additional mention increasing the odds of Streaming API capture by nearly 600%. And while the odds of tweets generated by users with verified accounts are 9% less likely to appear in the #jointsession Streaming data, they are 35.2% less likely in the @realdonaldtrump dataset. This is

especially remarkable given the increased focus since the 2016 presidential elections on the impact of Twitter bots on political news and discourse.

Table 9 displays the results of the logistic regression models for the three Search API datasets. There is limited consistency across all three, with only list count, hashtag count, and quote tweet demonstrating statistical significance and similar coefficient direction. However, because the #fomc model has a small number of observations, with many independent variables, including quite a few blunt dummy variables, we concentrate attention on the two larger datasets. The small dataset is important for data completeness, but we would expect more stable results in the models with a higher number of observations. Comparing these two, we find nine of the twelve variables prove consistent. List count, verified, hashtag count, and reply are statistically significant and positively correlated with the likelihood of Search API capture. The following count, status count, mentions count, retweet, and quote variables are also statistically significant, but all are negatively associated with the likelihood of appearing the Search API results.

The most striking difference between the two models relates to media. In the #ahca data, the odds of tweets containing some form of media being captured are 99.1% higher, but for #jointsession, they are 2.4% lower (see Table 10). Otherwise, the core differences relate to the magnitude of effect sizes. The odds of replies appearing in #jointsession are 18.2% higher than non-replies but 277.1% higher for #ahca. In contrast, the odds of retweets being capture in #ahca are 5.4% lower than standard tweets but 56% lower for #jointsession. In both datasets, the effects of the number of hashtags and user verification are particularly high. Each additional hashtag increases the odds of appearing in the #ahca and #jointsession captures by 95.4% and 525.8%, respectively. Verified accounts have 213.7% and 74.2% higher odds of being captured in the #ahca and #jointsession Search datasets, respectively.

Table 7: Logistic regression results - Streaming API

	Dependent variable:	
	Tweet in Streaming API	
	#jointsession	@realdonaldtrump
List count	-0.00001*	
	(0.00001)	
User like count	0.00000***	0.00000***
	(0.00000)	(0.00000)
Follower count	0.00000	-0.00000***
	(0.00000)	(0.00000)
Following count	0.00000***	0.00000***
	(0.00000)	(0.00000)
Status count	0.00000*	0.00000***
	(0.00000)	(0.00000)
Verified	-0.095***	-0.434***
	(0.019)	(0.021)
Media	0.053***	0.428***
	(0.010)	(0.007)
Hashtag count	0.067***	0.130***
	(0.003)	(0.003)
Mention count	0.084***	1.791***
	(0.006)	(0.004)
Reply	0.015	-0.642***
	(0.029)	(0.010)
Retweet	0.352***	-1.338***
	(0.007)	(0.007)
Quote tweet	0.247***	
	(0.012)	
Constant	0.269***	0.073***
	(0.007)	(0.007)
Observations	555,966	1,238,662
Log Likelihood	-353,409.400	-544,910.200
AIC	706,844.800	1,089,842.000
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table 8: Odds ratios - Streaming API

	Dependent variable:	
	Tweet in Streaming API	
	#jointsession	@realdonaldtrump
List count	1.000	
User like count	1.000	1.000
Follower count	1.000	1.000
Following count	1.000	1.000
Status count	1.000	1.000
Verified	0.910	0.648
Media	1.054	1.535
Hashtag count	1.069	1.139
Mention count	1.087	5.994
Reply	1.015	0.526
Retweet	1.422	0.262
Quote tweet	1.280	
Constant	1.309	1.076
Observations	555,966	1,238,662
Log Likelihood	-353,409.400	-544,910.200
AIC	706,844.800	1,089,842.000
Note:	*p<0.05; **p<0.01; ***p<0.001	

Table 9: Logistic regression results - Search API

	Dependent variable:		
	Tweet in Search API		
	#fomc	#ahca	#jointsession
List count	0.001**	0.0001***	0.0001***
	(0.0004)	(0.00002)	(0.00001)
User like count	0.00000	-0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)
Follower count	-0.00000**	-0.00000**	-0.00000
	(0.00000)	(0.00000)	(0.00000)
Following count	-0.00002	-0.00000**	-0.00001***
	(0.00003)	(0.00000)	(0.00000)
Status count	-0.00000	-0.00000***	-0.00000***
	(0.00000)	(0.00000)	(0.00000)
Verified	0.504	0.759***	0.555***
	(0.594)	(0.054)	(0.028)
Media	-1.026***	0.689***	-0.025**
	(0.298)	(0.026)	(0.011)
Hashtag count	0.453***	0.670***	1.660***
	(0.094)	(0.007)	(0.006)
Mention count	-0.332	-0.159***	-0.074***
	(0.283)	(0.009)	(0.007)
Reply	1.317	1.019***	0.167***
	(1.339)	(0.061)	(0.048)
Retweet	0.533**	-0.056***	-0.821***
	(0.237)	(0.017)	(0.009)
Quote tweet	-2.501***	-2.479***	-2.095***
	(0.278)	(0.028)	(0.015)
Constant	1.942***	0.377***	0.069***
	(0.302)	(0.020)	(0.010)
Observations	1,061	122,869	569,015
Log Likelihood	-311.707	-58,586.360	-254,965.400
AIC	649.415	117,198.700	509,956.900

Note: *p<0.05; **p<0.01; ***p<0.001

Table 10: Odds ratios - Search API

	Dependent variable:		
	Tweet in Search API		
	#fomc	#ahca	#jointsession
List count	1.001	1.000	1.000
User like count	1.000	1.000	1.000
Follower count	1.000	1.000	1.000
Following count	1.000	1.000	1.000
Status count	1.000	1.000	1.000
Verified	1.655	2.137	1.742
Media	0.359	1.991	0.976
Hashtag count	1.572	1.954	5.258
Mention count	0.718	0.853	0.929
Reply	3.733	2.771	1.182
Retweet	1.704	0.946	0.440
Quote tweet	0.082	0.084	0.123
Constant	6.972	1.458	1.071
Observations	1,061	122,869	569,015
Log Likelihood	-311.707	-58,586.360	254,965.400
AIC	649.415	117,198.700	509,956.900

Discussion

Taken together, these results should give serious pause to researchers using the Search and Streaming APIs for capturing keyword-based datasets. The good news is that, when not rate-limited, the Streaming API does appear to return nearly all relevant tweets. However, when rate limits are imposed, the likelihood of introducing bias into one's scientific inferences is substantial. In the two rate-limited datasets, Kendall's-tau correlations for top hashtags, mentions, and usernames fell outside of the values expected based on random sampling of the full population of tweets in 40 out of 42 instances. What is more, we found systematic differences between the PowerTrack and Streaming data for the vast majority of variables, and in most cases, effect sizes were large enough to suggest the impact of these differences would be high. If, for example, one were examining retweet networks during Donald Trump's inauguration, findings would be based on data in which retweets are substantially underrepresented, but if examining the same question during Trump's first address to a Congressional joint session, findings would be rooted in a sample in which retweets are actually considerably more likely to appear than standard tweets. Yet without purchasing the full population data, one simply cannot know which—if either—type of dataset one has captured.

The risks are even higher for the Search API, which consistently returns less data than the Streaming API and does not achieve completeness, even when the total volume of relevant tweets is low. Based on Twitter's own (vague) statements about the Search API, we know that there are systematic differences between data captured via the API and the full population of tweets. That Twitter prioritizes "top" tweets tells us as much. But until now, we had little inkling of what parameters Twitter is using to designate a tweet "top". Though just a first cut using a small number of datasets, our results suggest that original tweets (as opposed to retweets) receive

priority, as do tweets with more hashtags, and those generated by verified accounts. Still, the fact that our results are not entirely consistent across models highlights the need for more research into this question.

It also points to a fundamental problem affecting all Twitter research: *We simply do not know what we do not know.* It might be that with a large number of comparisons we can detect consistent patterns in the results returned by each of the public APIs. Yet it might also be the case that any inconsistencies found in our results are a product of changes to the APIs themselves, rather than the inadequacies of our data and models. There is simply no way for us to know. Though we have presented data captured over a relatively short timeframe—with the first, @realdonaldtrump, dataset captured in late January 2017 and the last, #fomc, captured less than six months later—it is likely that the APIs changed, at least in small ways, during that period. But were there larger changes? And even if just small, were these enough to impact our comparisons? Again, we simply do not know.

What we do know is that based on the evidence presented in previous research and expanded upon here, we cannot and should not take for granted that data drawn from the Search or Streaming APIs are representative of the underlying population of relevant tweets. Indeed, it is much more appropriate to assume the opposite—that these data are systematically different and likely to introduce bias into any findings upon which they are based.

References

Bruns, A., & Weller, K. (2014). Twitter data analytics – or: the pleasures and perils of studying Twitter. *Aslib Journal of Information Management*, *66*(3).

Driscoll, K., & Walker, S. (2014). Big Data, Big Questions| Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication*, *8*(0), 20.

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*, 16–27. https://doi.org/10.1016/j.socnet.2014.01.004

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *arXiv:1306.5204 [Physics]*. Retrieved from http://arxiv.org/abs/1306.5204

Zimmer, M., & Proferes, N. (2014). A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250–261.