

Orthogonal Sampling: An Approach to Reducing Selection Bias in the Collection of Social Media Data

Jeffrey Sachs

Department of Sociology
Yale University
ORCID: 0000-0002-6451-2502
email: jeffrey.sachs@yale.edu

Daniel Karell

Department of Sociology
and
Institute for Social and Policy Studies
Yale University
ORCID: 0000-0001-6709-6535

Acknowledgements

We thank the participants of the session on Improving the Transparency and Reproducibility of Social Research at the 2022 Annual Meeting of the American Sociological Association, organized by Jeremy Freese on behalf of the Methodology Section, for their helpful comments.

Abstract

Social scientists studying contemporary public debates, discourses, and sentiments often turn to social media platforms for data. They typically collect these data by querying platforms' application program interfaces (APIs) using a term of interest, as well as terms related to the target term, or "topic-based sampling". Unfortunately, because of how APIs work, this risks introducing bias into the samples, likely resulting in inaccurate estimates of how prevalent a discourse is on the platform. We For Peer Review introduce and demonstrate an approach to topic-based sampling that reduces bias and allows for more reliable inferences about the population. Building on theory from animal ecology and previous efforts to address social media sampling problems, and using tools from natural language processing, we first develop and detail the procedures for our "orthogonal sampling approach". The method consists of incorporating layers of independence into data collection. It is based on the core idea that a discursive space, such as a social media platform, can be treated geometrically and thus leveraged to set initial locations for data collection to topics that are orthogonal to the topic of interest. After explaining our approach, we provide evidence that our approach produces samples of a discourse that are representative of the true abundance of the discourse. Our results additionally indicate that as the samples' sizes increase, so should their representativeness. We also find that gathering social media by querying a target term and related terms – a common practice in the scholarship – can have strong negative effects on properties that are characteristics of a representative sample, suggesting that mitigating sampling bias when collecting social media data should be a priority for many studies analyzing online discourse.

1 Introduction

Social scientists studying contemporary public debates, discourses, and sentiments commonly turn to social media platforms for data. They see these data as offering insights into widespread thinking among members of some population of interest (Rafail 2018).¹ Unfortunately, using these data to make inferences about large-scale dialogue and discourse often comes with well-known problems (boyd and Crawford 2012; Marres 2017; Rafail 2018; Hargittai 2020; Özkula et al. 2023). For example, social media data are usually generated by certain kinds of people, who, by being, say, younger, wealthier, and more educated, are often unlike most other members of a population (Blank 2017; Malik et al. 2017). In addition, data from one social media platform capture only the socio-cultural dynamics on that platform, which may be strongly structured by that platform’s characteristics and, as a result, unlike related dynamics on other platforms (Venturini et al. 2018; Özkula et al. 2023).

Another significant problem with using social media to study a broad discourse stems from how researchers typically gather the data. One of the most common – if not the most common – collection techniques is “topic-based sampling” (Gerlitz and Rieder 2013). A researcher identifies target terms – including, in the case of Twitter (also known as “X”), hashtags – and then collects social media messages or posts (*e.g.*, tweets) containing those terms by querying the social media platform’s application program interface (API) (Burgess and Bruns 2015; Özkula et al. 2023). This strategy rests on the assumption that topics, themes, or content of interest to the researcher contain, and cluster around, the target terms (Gerlitz and Rieder 2013). With this technique, researchers can, for example, select tweets with hashtags like “clinton” and “trump” to examine mis/disinformation during the 2016 United States presidential election (Barberá 2018), use terms like “vaccineskill” and “vaccinesideeffects” to query Gab’s API and build a dataset for studying the radicalization of discourse about vaccines during the COVID-19 pandemic (Dehghan and Nagappa 2022), and construct a corpus of tweets about the Black Lives Matter movement by querying Twitter’s API with terms and phrases like “blm” and “black lives matter” (van der Veen 2022).

A problem arises because social media platforms’ APIs sample from the universe of observations using undisclosed methodologies, vary unpredictably across sampling draws, and are censored, sometimes in unclear ways (boyd and Crawford 2012; Lorentzen and Nolin 2017; McCormick et al. 2017; Rafail 2018). For example, despite Twitter’s API documentation describing the data it provides as a “random” sample, Pfeffer et al. (2018) used automated accounts to make particular tweets up to 84 times more likely to appear in their samples than expected.² Then, after demonstrating that data from Twitter are not in fact random, Pfeffer

¹ For example, in their review of research on digital activists, Özkula et al. (2023) report that 57 percent of articles used social media data (63 percent use digital data of some kind). Since this review only included articles published between 2011 and 2018, we suspect the proportion of recent articles drawing on social media data is greater.

² Gerlitz and Rieder (2013) offered some tentative – and early – evidence that Twitter’s APIs might have provided truly

et al.’s experiments additionally revealed how simple fluctuations in the (non-random) sampling process changed the results of downstream analyses (see also Lorentzen and Nolin 2017; Rafail 2018). Kim et al. (2020) conducted tests showing how data collected from Twitter’s three different APIs produced diverging datasets, even when using the same search parameters over the same data gathering time period. Sachs et al. (2021) found that TikTok’s API appeared to limit query results of songs’ usage in videos to the 4000 most popular videos. Given that TikTok has over a billion users, it is reasonable to expect that even moderately popular songs are used in many more videos. Thus topic-based sampling using query APIs risks inducing an unknown kind of bias when gathering social media data.

How have social scientists addressed the problem in research using these kinds of data to answer substantive questions? While some have uncritically repeated platforms’ descriptions of data handling (*e.g.*, Alshaabi et al. 2021), most take an understandable approach: they acknowledge the problem as a limitation, then, taking no or few mitigation steps, turn to the focus of their study – the substantive research questions (*e.g.*, Beauchamp 2017; Flores 2017; Steinert-Threlkeld 2017; Larson et al. 2019). Other than having limited remediation tools at their disposal, the reasoning seems to be that valuable insights can still be gained from any “large enough” dataset (see Salganik 2018). We agree with this perspective, but only when researchers have a sense of the potential bias (Freese and King 2018; Rafail 2018; Christensen et al. 2019). In the case of collecting social media data, the potential bias is unclear because APIs’ data handling and sampling strategies are unknown.

In this paper, we introduce an approach to topic-based sampling that minimizes selection bias and allows for reliable inferences about the population. Our approach builds on previous efforts to tackle the social media sampling problems by borrowing theory from animal ecology (Cooch and White 2002; Powell and Gale 2015) and tools from natural language processing. In doing so, we offer a method for topic-based sampling with logic that generalizes across social media platforms that are discursive environments. Practically, this approach can help researchers collect representative topic-based samples from which inferences about the abundance of the topic can be made.

Below, we first outline the logic of our approach, which we call an “orthogonal sampling approach”, then demonstrate the approach on two Twitter hashtags – **#golf** and **#work** – to show its ability to produce a representative sample from which inferences about the population can be made. Fundamental to the orthogonal sampling approach is the desire to eliminate sampling bias. As such, we also perform a placebo test on **#golf** to show how exaggerated population inferences become under sampling bias. Our empirical demonstration includes details of the methodology not presented in the initial outline.

random samples, but we place greater weight on the more recent (and copious) research identifying the biases in platform APIs’ samples, including Twitter’s APIs. We cite this work in this Introduction section.

Before continuing, we acknowledge that recent changes at Twitter – which are in the midst of happening as we write – may make it difficult to conduct some specific steps of our empirical exercise. Nevertheless, there is still value in using Twitter to demonstrate our approach. After all, it has been the prevailing source of data for social scientists studying social media³ and these researchers commonly gather data samples from the Twitter platform using topic-based sampling, or querying the API for hashtags or keywords related to particular topics of interest. As discussed, such topic-based API queries are often not free from sampling bias, as the API returns a subset of tweets under nebulous selection criteria.

More importantly, however, our main insight is unaffected by the changes at Twitter. Selection bias from API queries is not limited to Twitter; academic research based on collected data from an array of social media platform APIs faces the problem of biased samples from which we risk making questionable inferences about the population. Our approach, regardless of the platform being used, helps eliminate selection bias during topic-based sampling.

Specifically, we argue – once again, for any platform or API – that it is valuable to set the initial locations for collection to topics that are orthogonal (*i.e.*, independent) to the topic of interest. This circumvents much of the selection bias introduced by sampling directly on the topic of interest using an API. Then, it is important to use a multi-stage approach for gathering users, followers, and finally content (*e.g.*, tweets) to generate a large sample in which the topic of interest does not correlate with any bias introduced by the API. By sampling independently of the topic of interest (or with complete indifference), it is possible to calculate the “capture probability” of the topic of interest and produce unbiased estimates of the abundance of the topic in the population.

2 Related research

We are not the first to develop methods for handling unrepresentative, unstable, and opaque sampling from social media platforms’ APIs. Hino and Fahey (2019), for example, offer an elegant solution based on using Twitter account numbers to randomly sample from the entire population of accounts. However, since Twitter started assigning account numbers randomly in 2016 (rather than incrementally, as prior to 2016), this technique is not viable for studying recent content on Twitter. Lorentzen and Nolin (2017) highlight the benefits of multi-stage sampling: they first used topic-based sampling (*i.e.*, hashtag queries of Twitter’s API) to find documents about a particular topic, then used these documents’ IDs to collect related documents which were not in the initial sample, which primarily were replied-to tweets. This method, though, is most

³ For example, Özkula et al. (2023) note that “Twitter constituted 58% of all articles drawing on single-platform research ... Twitter [also] constituted a highly popular choice for the vast majority of articles drawing on multi-platform approaches (93.6%).”

appropriate for finding previously unknown related messages to reconstruct conversations on social media platforms. It is less useful for estimating characteristics of a population.

Rafail (2018) describes different sampling approaches for studying different types of social media user populations. The technique for sampling populations defined by a discourse – “the most difficult” type to sample (pg. 199), and the one we focus on – uses a dynamic set of keywords during topic-based sampling. That is, the researcher begins with a list of terms of interest, then iterates through a sequence of sampling, identifying related yet unanticipated terms, and resampling (see also King et al. 2017). The aim is to collect the documents that exhaustively capture the discourse of interest.

While Rafail (2018) provides compelling evidence that dynamic topic-based sampling can approximate complete coverage of a discourse, the technique relies on the researcher selecting the “correct” new keywords during each iteration. In addition, the technique’s transparency and reproducibility are highly sensitive to how the researcher records the work (see Freese and King 2018; Christensen et al. 2019). Moreover, while the technique introduces a useful way to evaluate the coverage of a topic (when keywords are appropriately selected), it does not support the ability to make unbiased estimates of the topic size in the population.

3 Outline of the approach

We propose an orthogonal sampling approach based on a simple logic. Namely, submit API queries for channels of data that are independent from the channel of interest. This logic is akin to that used for animal population estimation (Cooch and White 2002; Powell and Gale 2015). When sampling a wild animal population, ecologists randomly distribute traps to a variety of places across the environment. This approach avoids biasing in favor of any particular location of the environment which could affect the representativeness of the sample. If, for instance, we sampled exclusively on locations in which we knew we would find a high abundance of tigers, using such locations would bias not only our estimates of the tiger population across the entire environment, as tigers would likely be less common in other locations, but high abundance locations might also present examples of tigers which are characteristically different from tigers captured in other locations. Analogously, our orthogonal sampling technique mitigates API-induced bias by randomly selecting a collection of data channels that are independent of the channel of interest. In animal research terms, the approach randomly distributes “traps” for our data of interest across the virtual environment.

In essence, this logic parallels conditional sampling. Because most social media platforms are a complex network of user relationships, the landscape of users on such platforms is difficult to sample from without bias. Rather than asking the API to sample a subset of data directly from the population which has an unknown amount of bias, conditional sampling is an approach to sample on the presence of another another

variable, such as a hashtag of interest. Given a population P and a subset S within P that has a specific condition C , we randomly draw samples from S (which has the condition C) and observe X (say, a target hashtag), given C . The condition C for our approach is that samples S be drawn independently of X , thus minimizing selection bias. Theoretically, if the samples S are independent of X , then we can estimate the population parameter from our samples, $E[X|C]$ or the expected value of X given C , which refers in this context to the capture probability of X . By focusing the sampling on this subset, S , our approach reduces the biases that might occur if we directly sampled tweets based on a hashtag of interest, leading to potentially more representative and unbiased estimates.

Our approach assumes that a researcher has a topic of interest in mind, and has decided how the topic is represented in social media content by one or more discursive elements. Examples of discursive elements can be a particular hashtag used in tweets, a specific song incorporated into TikToks, a phrase used in YouTube video snippets to describe the video content, or even an image or meme shared on Instagram. For the empirical demonstration below, we use hashtags appearing in tweets since, as earlier discussed, Twitter data are by far the most commonly used social media data in social science research.

The approach itself consists of four steps, or “layers of independence”. In the first step, we identify and collect a basket of “orthogonal” hashtags. A hashtag qualifies as orthogonal if its vector representation in an embedding space exhibits a satisfactory level of dissimilarity, or discursive independence, from our hashtag of interest. To calculate orthogonality, we use a GloVe embedding space that has been pre-trained on two billion tweets. GloVe (Global Vectors for Word Representation) is a model for learning vector representations of words. Developed by Pennington, Socher, and Manning (2014), the model aims to encode semantic meaning into a geometric space, typically of much lower dimensionality than the original vocabulary size. This is accomplished by leveraging the global statistical information of a corpus. By using GloVe embeddings, our approach effectively transforms the problem of sampling from a large, sparse, and high-dimensional space into a problem of sampling from a well-behaved, continuous, and semantically-rich space.

Using the embedding space, we vectorize our target hashtags and calculate which other vectors (and the hashtags those vectors represent) are independent using cosine similarity. Cosine similarity ranges between -1, exact opposite, and 1, exact same, with 0 indicating that two vectors are orthogonal, which is a strict mathematical criterion for independence in vector space.⁴ In other words, a cosine similarity of zero in this space is a strong indicator that two hashtags (or other kinds of discursive elements) are unrelated or independent in their contextual usage.

The second step involves collecting the orthogonal discursive elements after identifying them, then gath-

⁴ Cosine similarity is a geometrically robust metric that measures the cosine of the angle between two vectors. We define it formally below.

ering a random sample of users from each orthogonal hashtag location. Using Twitter API endpoints, we identify a large collection of orthogonal hashtags. These effectively represent a pool of locations independent of our target hashtag, across the twitter environment. Sampling from these orthogonal locations avoids biasing the search for the target hashtag in any particular location.

Next, we construct a sample of “first-order” users through a random search of the Twitter landscape from the starting points of our orthogonal hashtags. At this stage, researchers may select to purposefully eliminate some users from the sample, such as users who are bots or so-called sockpuppets.⁵ While the sample of users gathered at this step might qualify as random and unbiased in most cases, the sample remains susceptible to API filtering. Namely, if an orthogonal hashtag is very popular, the API requests could select certain types of users who are not as representative of the population, such as popular users.⁶

The third step helps circumvent possible selection bias introduced during the search for “first-order” users. We collect a random sample of “second-order” users from first-order users’ social networks (*e.g.*, their followers). Researchers can again eliminate undesirable second-order users, such as bots or corporate accounts.⁷ A key feature of this step is taking advantage of the transparency that *does* exist: social media platforms’ APIs are usually more transparent about the sampling procedure when specific user accounts are queried. For example, the Twitter API provides followers in groups of 5000 and they are ordered from most recent to oldest with respect to following. A second feature is that collecting a random sample of followers also allows us to collect tweets from a wider segment of the environment.

[Figure 1 about here]

After completing the first three steps, we have a large random sample of (second-order) users distributed widely across the platform environment, and who are discursively independent of the discursive element of interest. These users represent the “traps” from which we look for online content featuring the target discursive element. In the fourth and final step, we collect these users’ content, which, in our implementation, are tweets. For Twitter, each user’s most recent 3,200 tweets can be retrieved.⁸ In addition, user timeline requests return tweets produced in chronological order, so we gain further transparency regarding where censoring in our requests might occur, *i.e.*, tweets older than a user’s most recent 3,200. The prevalence of such censoring should be relatively rare, especially if the researcher is interested in content occurring within a specific and more recent time interval.

Figure 1 presents a schema of our sampling approach. For illustrative purposes, and for consistency with

⁵ In addition, note that this search is limited to public user profiles.

⁶ Another method also exists to circumvent this bias, addressed in the Methods section.

⁷ And, again, the search is limited to public accounts.

⁸ Sysomos, a social media platform research firm, estimates that about 88% of users have produced fewer than 999 Tweets and around 97% have produced fewer than 4999.

our empirical demonstration, it uses the Twitter hashtag **#interest** as the target discursive element.

4 Empirical demonstration

We now turn to demonstrating our approach with an empirical analysis of two target Twitter hashtags, **#golf** and **#work**.⁹ Our demonstration is grounded in a practical scenario. Imagine that a researcher is interested in understanding online discourse about a topic. The researcher, like many others, turns to a social media platform for data. They would like to gather a representative sample of the discourse, not just collect a sample heavy with the most popular or extreme content. They would also like to get a sense of how pervasive discourse is about the topic on the social media platform. That is, they want to know how abundant the discourse is or how likely it is to come upon the discourse one uses the platform.

We begin our demonstration by elaborating each step of our approach for collecting samples of **#golf** and **#work**, including formal definitions and technical details. Then, we explain how we evaluated the quality of the samples, and, by implication, our approach. That is, because social media researchers using topic-based sampling are often interested in their target discourse’s pervasiveness, it is critical that we provide evidence that our approach produces samples that reflect the characteristics of the population from which they are drawn. Thus, in order to validate that our approach does provide such a representative sample, the empirical demonstration involves extensive evaluation of the parameters of interest.

The evaluation required the collection of two additional kinds of data. First, we obtained the true population of tweets containing our target hashtags, or the “ground truth population”, for a given period of time. Collecting the ground truth population is possible because the Twitter API allows users to collect all of a particular hashtag if the abundance of that hashtag amounts to less than 1% of the tweet population (Pfeffer et al. 2018). This is why we chose **#golf** and **#work** as our target hashtags; they were sufficiently small to collect the complete population for during the time that we conducted data collection.

The second kind of data were a purposefully biased sample of **#golf**, or a “placebo” sample. Collecting this sample entailed following a common approach in the scholarship: we sampled from a basket of hashtags that are similar to our hashtag of interest. Since this (deliberately) introduced selection bias into the sample, we refer to this sample as “Biased **#golf**”.

With our approach’s samples of **#golf** and **#work**, the ground truth of **#golf** and **#work**, and the Biased **#golf** sample, we conducted our assessments of quality. In the first assessment, we examined the bias of

⁹ For readers unfamiliar with Twitter, Twitter is a prominent social media platform introduced in 2006, enabling users to share real-time updates in short messages called “tweets”. Tweets are concise messages limited to 280 characters, facilitating quick communication and engagement. Twitter’s asymmetrical model allows users to follow others, creating a diverse online community connecting individuals, organizations, and public figures. Hashtags are keywords preceded by “#” used to categorize and group tweets around specific topics, fostering discussions and trend discovery.

our population estimate with respect to the true population parameter. Practically, this involved observing how close our estimated parameter was to the true population parameter, and comparing these results to estimates based on Biased `#golf`.

Next, we performed a coverage assessment of the three samples, *i.e.*, `#golf`, `#work`, and Biased `#golf`. This assessment entailed checking whether the true parameter falls within the confidence interval within a particular frequency. Ideally, for a 95% confidence interval, the true parameter should fall within the interval in 95 out of 100 samples. Then, we assessed the consistency of our estimated parameter for all three samples. An estimated parameter is consistent if it converges in probability to the true parameter as the sample size tends to infinity. Finally, we compared the efficiency between our sampling approach as implemented for `#golf` and the sampling method implemented for Biased `#golf`. The more efficient population estimate should have a smaller and narrower variance.

If the estimator performs well across these assessments, it strongly suggests that the sample from which it is derived is a good representation of the broader population. When the coverage rate of an estimator consistently matches the intended level (say, 95% for a 95% confidence interval), it suggests that the sample data is not systematically skewed or biased. A consistent estimator derived from larger samples reinforces the idea that we are capturing the genuine underlying trend or characteristic of the entire population. Finally, efficiency ensures that the estimator not only targets the true parameter but does so with the minimum possible variance given the data. We explain the assessments in greater detail below and in appendices.

4.1 Methods

4.1.1 Ground truth

Before implementing our approach, we collected and verified ground truth populations of `#golf` and `#work` for a given period of time. With the true populations, we can evaluate the quality of our sampling approach by estimating the abundance of `#golf` and `#work` from our sample and comparing these estimates to the ground truth for each.

To collect the ground truth population of our target hashtag, we used the `search_tweets()` of the `rtweets` R package to request tweets from both “verified” and “unverified” users. The parameter was set to collect only English tweets to make management of the data more convenient. We set the “type” parameter to “recent”, which allows us to collect all tweets from most recent to an end date and time chosen by the API. This ensures that we have all tweets over a particular time interval – typically between 6 and 12 days. We also included retweets and asked the API to retry whenever it reached the rate limit.

We then verified our ground truth populations for the study period by collecting tweets for each hashtag

twice (an A sample and B sample). Because these samples were collected in sequence (A first, then B), we subset the B sample to only include tweets that were older than the most recent tweet of the A sample, and we subset the A sample to include only tweets that were more recent than the oldest tweet in the B sample. Through this sub-setting, sample A and B were given consistent time intervals. We trust the API documentation that we could gather the complete population of hashtags with small enough populations, but we tested the validity of this claim by verifying that sample A and sample B were the exact same set of tweets by taking the set difference between the two.

For `#golf`, the ground truth population consisted of 11,319 tweets between October 10, 2022 and October 21, 2022. For `#work`, the ground truth population consisted of 16,676 tweets between December 28, 2022 and January 7, 2023.¹⁰

4.1.2 Step 1: Orthogonal locations

The first step of the orthogonal sampling approach involves creating a basket of hashtags that are orthogonal or independent of our hashtag of interest. To assemble a basket of orthogonal hashtags for `#golf` and `#work`, we use a pre-trained GloVe embedding space of 200 dimensions from the Stanford NLP Group (Pennington et al. 2014). The particular GloVe model used was trained on 2B tweets. We then trimmed the model by removing non-English terms and any words that did not exist in the Augmented List of Grady Ward’s English Words and Kantrowitz’s Names List from the `qdap` R package (Rinker 2023), which together sum to 122,806 words, and a slang dictionary created using dictionary.com, totalling 943 words. We used these dictionaries to limit the modeling space to ensure that the orthogonal hashtags we were about to detect – the hashtags which would serve as initial locations for sampling – were substantive and easy to interpret.

The pre-trained GloVe model represents each word as a 200 dimension vector in the embedding space. This allows us to discern words’ similarity by comparing vectors within the embedding space. Embedding spaces are frequently used for such purposes, especially because empirically studies have shown that the vectors often represent meaningful relationships. For example, the vector difference between “man” and “woman” in the GloVe model should be relatively equal to the vector difference between “king” and “queen.” Our use of the embedding space is slightly different, however. Instead of seeking out meaningful differences between words in the embedding space, we simply aim to identify orthogonal or independent relationships between words in the embedding space. By identifying orthogonal words in the embedding space, we can gather a set of locations to sample from that are independent of our hashtags of interest. Appendix A provides details of the GloVe model.

¹⁰Because the first and last day of the ground truth consist of partial days, both samples are subset again to eliminate the partial first and last days for convenience when comparing population estimates.

We implemented our search for orthogonal hashtags for both `#golf` and `#work` using the `sim2()` function of the `text2vec` package in R. The `sim2()` function takes as inputs the vector for our hashtag of interest in the GloVe embedding space, the embedding space object, and a method for calculating similarity. The method we used, as earlier mentioned, was cosine similarity (see Appendix B for further details about computing cosine similarity).

The cosine similarity calculation is expressed to five decimal places and therefore rarely returns a value of exactly 0 to indicate perfect orthogonality. For this reason, we used a rounding method to qualify orthogonal hashtags. We rounded cosine similarity calculations to three decimal places, transforming any cosine similarity below 0.005 and above -0.005 to 0, satisfying a sufficient level of independence.

For both `#golf` and `#work`, we assembled baskets of orthogonal hashtags with cosine similarities in the embedding space that fell within our rounding interval for independence. We identified a basket of 2,244 orthogonal hashtags for `#golf` and 1,695 orthogonal hashtags for `#work`.

4.1.3 Step 2: Sampling tweets and users

In the second step of the approach, we collected tweets from each orthogonal hashtag and then extract information about the Twitter users who produced those tweets. For each orthogonal hashtag, we made an API query using the `search_tweets()` function of the `rtweets` R package. During these requests, we used similar parameters as we did in collecting our ground truth populations, including requesting tweets from “verified” and “unverified” users and retweets. In each request, we set a limit of 18,000 tweets to cut down on rate limit restrictions. To help limit the amount of filtering that might occur from only 18,000 tweets from each hashtag, we requested tweets from most recent to older, as the ground truth populations consisted of tweets that were no more than 12 days old.

At this stage, we do not use the tweet content, we only save the user ids associated with the tweets at each orthogonal hashtag. We then take a random 10% sample the users collected at each orthogonal hashtag. Collecting only 10% helped us limit the size of sample for subsequent steps. In total, we collected 4,174 users from hashtags orthogonal to `#golf` and 8,347 users from hashtags orthogonal to `#work`. `#golf` included more orthogonal hashtags than `#work` in its basket, but we sampled less than half the number of users from these hashtags as we did from `#work`. This indicates that at least some of the orthogonal hashtags for `#work` were more common on Twitter.

4.1.4 Step 3: Sampling followers

In the previous step, we limited the number of tweets requested from each orthogonal hashtag in implementation to 18,000. We did this to avoid incurring time costs due to rate limit restrictions. We acknowledge,

however, that this choice could bias the sampling toward tweets that are possibly more popular or highlighted by the API with some other characteristic that might affect independence in our sampling procedure. We circumvent this by collecting followers from each user sampled in the previous step.

Sampling followers has other advantages. First, API requests for users' social networks, such as Twitter users' followers, tend to be more transparent. For example, Twitter API requests for followers return followers in batches of 75,000 in chronological order of following (in order of most recent). More than 75,000 followers can be gathered from a users, but not without incurring rate limit restrictions. However, because the average Twitter user in 2023 is estimated to be 707 (Bagadiya 2023), we believe a request for 75,000 gathers all followers from most users. Second, this subsequent round of sampling helps increase the size of our sampling approach to much larger set of users across Twitter. We collect followers from specific users using the `get_followers()` function in `rtweets`.

In our empirical exercise, we limited the number of followers actually used in subsequent steps of the approach by randomly sampling 10% of the followers that we gathered from each user. Implementing this random sampling further reduced possible selection bias from the API and shorten the duration of the sampling approach. For `#golf` we sampled 1,284,600 followers. For `#work` we sampled 1,143,700 followers.

4.1.5 Step 4: Sampling tweets

The samples of followers gathered in the preceding step do not have selection bias with regard to our target hashtags. We do not claim that these samples are entirely representative of the Twitter user base; we only claim that they have been selected independently of a particular hashtag of interest. As a result, the tweets found among these users should be representative of the population of tweets featuring the hashtag of interest.

In the final step of our approach, we collect tweets from the samples of followers. To do so, we used the `user_timeline()` endpoint of the Twitter API. This was easily implemented using the `get_timeline()` function from the `rtweets` library. Tweets are returned by the API in order of their occurrence, starting with the most recent, and up to 3,200 tweets from each Twitter user. Another rate limit restriction limits us to collecting tweets from 900 users at a time. This slows the collection of data, but does not introduce any API related bias.

We sampled in chunks of 100 followers, which allowed us to save and remove these chunks from the environment and maintain processing power throughout the collection procedure. We also only collected tweets from a 5% random sample of these follower chunks. For our results, we only needed a sample size large enough to test the quality of our sampling approach by way of estimating the population size. A smaller sample typically provides a less accurate estimate of the population size, as the estimate will fluctuate more

with small amounts variation in the number of tweets captured, so the choice to estimate the population size with the smaller sample size served as a robustness test on the quality of the sample. In total, we collected 4,301,214 tweets in implementing the sampling approach for `#golf` and 2,542,972 tweets for `#work`.

4.1.6 Capture probability and population inference

After completing our proposed sampling approach, we estimated the true population for both `#golf` and `#work`. This was necessary to subsequently evaluate the quality of the orthogonal sampling approach. In other words, we sought to answer the question, how good are the estimates based on our samples?

Our approach to estimation relies on the capture probability of our hashtags of interest in their respective samples. The underlying logic is that if the orthogonal sampling approach generates an accurate estimate of the population as derived from its capture probability in our sample, then we have produced a representative sample (at least from the perspective of its prevalence in the population).

To estimate the population, we first calculated the capture probability for each of the two target hashtags in their respective samples. The capture probability is simply the number of tweets featuring the hashtag of interest divided by the total number of tweets in the population. It is defined by the following equation:

$$\hat{p} = \frac{|X|}{|S|} \quad (1)$$

where $|X|$ is the count of tweets in the subset featuring the hashtag of interest and $|S|$ is the total count of tweets in the sample. That is, $X = \{all x : x \in X, x \in S\}$. Since we have sampled independently of our hashtag of interest, the capture probability is thus the probability of randomly sampling a tweet featuring our hashtag of interest in the wider Twitter environment.

In addition to estimating the the capture probability, we computed confidence intervals by bootstrapping our sample using the `boot` library in R (Canty and Ripley 2022). Following Keener (2010), the bootstrap method used here and in a number of the following tests can be defined as such : Let X_1, X_2, \dots, X_n be the observed dataset, where n is the size of the dataset. Then, to bootstrap, we create B new datasets D_1, D_2, \dots, D_B by sampling n times with replacement from X_1, X_2, \dots, X_n . Each dataset D_b , for $b = 1, 2, \dots, B$, is a random sample with replacement and is defined as:

$$D_b = \{X_{b1}, X_{b2}, \dots, X_{bn}\}$$

Each X_{bi} is a random variable such that

$$X_{bi} \sim Uniform(\{X_1, X_2, \dots, X_n\})$$

We use the bootstrap approach for estimation because of its ability to approximate a true unknown distribution from the empirical distribution of the observed data. By the law of large numbers, as n goes to infinity, the empirical distribution $F_n(x)$ should converge to the true distribution $F(x)$:

$$F_n(x) \xrightarrow{a.s.} F(x) \text{ as } n \rightarrow \infty$$

The bootstrap approximates the sampling distribution of a statistic T calculated on the sample. As the sample size n grows large, the bootstrap distribution should become an increasingly accurate approximation of the sampling distribution of T .

The statistic used in our case is very simple. We generate an indicator variable that evaluates to 1 if an observation in the sample includes our hashtag of interest and 0 if it does not. Each row is then given a uniform weight, equal to the $\frac{1}{n}$, where n equals the total number of observations in the data. Our bootstrapping consists of re-sampling from the data with replacement the same number of observations present in the data, generating a weighted sum of the indicator variable. The equation for this sum is given below:

$$\hat{p} = \sum_{i=1}^n I_i \in \{0, 1\} \times \frac{1}{n} \quad (2)$$

Where $I_i \in \{0, 1\}$ is the indicator variable that evaluates to 1 if the observation includes our hashtag of interest and 0 if it does not, and $\frac{1}{n}$ is the uniform weight applied to each observation in the data. The summation effectively calculates the fraction of observations that include our hashtag of interest. We iterate this bootstrapping approach 1000 times and generate an average capture probability over those iterations along with the confidence interval for the capture probability. In our Results section, we treat the average capture probability as a more robust capture probability when the hashtag of interest population is small.¹¹

We estimate the population size of tweets featuring our hashtag of interest by multiplying the capture probability by an estimate for the total number of tweets generated over our given time period on Twitter. We created this multiplier using data from Story Wrangler (Alshaabi et al. 2021). Story Wrangler uses the Twitters Decahose API to gather a 10% sample of publicly available tweets daily. Using this data, we can produce an estimate for the number of English language tweets generated on Twitter over our time period of interest. In order to use this estimate, we censor our sample to dates in which we have the complete day, eliminating partial days at the beginning and end of the sample.

The estimate involves taking the sum of the number of English tweets for each day over our the time

¹¹In our analysis, the parameter estimate is capture probability. The population inference stems from this, but could be slightly biased due to the fact that tweet volume estimates are (a) estimates, and (b) supplied by Twitter. For other platforms, capture probability would likely be the parameter of interest.

period of interest, and multiplying this sum by 10. The equation for estimating the population is given below:

$$\hat{P} = \hat{p} \times \hat{T} \quad (3)$$

where \hat{P} is our population estimate for our hashtag of interest, \hat{p} is the capture probability of our hashtag of interest, and \hat{T} is the estimate for the total number of English tweets derived from Story Wrangler data.¹² In our application, the \hat{p} used for the population estimate is the average capture probability derived from bootstrap sampling. We also apply the estimated total population multiplier to the confidence intervals generated for the capture probability.

4.1.7 Biased placebo sample

We next conducted a second sampling approach to help in our assessment of our orthogonal approach’s quality. Specifically, we examined the effect of eliminating selection bias using orthogonal hashtags by implementing the approach with deliberate sampling bias – collecting data using correlated hashtags, which is the common approach – and making a population inference from this biased sample. We could not directly test how biased the Twitter API is because we deliberately chose hashtags for which we could collect all tweets over a particular period of time. Therefore, we simulated selection bias by choosing hashtags from the GloVe embedding space with a cosine similarity beyond a chosen similarity threshold (*i.e.*, similar, rather than orthogonal, hashtags).

We ran this biased sampling approach on **#golf** as our hashtag of interest. We implemented the sampling approach using hashtags that had a cosine similarity of 0.3 or greater with **#golf**. We chose this similarity threshold because it is a conservative level of similarity, and estimation problems with a low level of bias will foreground the problem of not eliminating sampling bias.

We additionally searched across our collected sample of users and found 312 that also existed in the ground truth population. We then marked these users and their followers and removed any tweets created by these individuals from the sample — about 1.2 million tweets. This step makes for a very conservative estimate of the bias because it assumes that none of the 1.2 million tweets removed could have been associated with hashtags other than **#golf**.

¹²For other platforms it may be difficult to get an estimate of the amount of content generated over a period of time. In most cases, however, capture probability should be sufficient to interpret the prevalence of a topic on the social media platform.

4.1.8 Evaluating results

We evaluate the results of the orthogonal sampling approach with a set of assessments. First, we examine the practical performance of the approach. Namely, we compare how close the estimates generated using the orthogonal approach are to the true population, as well as how close the estimates generated with the biased sample are to the true population. To quantify the comparison, we use relative error, the proportional difference between an estimated parameter and a true parameter. Appendix C presents details of this assessment.

Second, we implement a number of assessments to statistically evaluate the quality of the estimators given our sampling approach. Since we have ground truth, we can easily assess our population inferences for *coverage*, *consistency*, and *efficiency*.

As mentioned, The quality of the estimates depends, crucially, on the representativeness of the sample from which we are drawing. *Coverage* refers to the frequency with which the confidence interval captures the true parameter value. If the confidence intervals produced by the estimator frequently capture the true parameter value, this suggests that the sample provides a good representation of the population. This would indicate good coverage. Poor coverage would mean that the sample is not reliably capturing the population characteristics. Appendix D elaborates on how we measure consistency, including formal definitions.

Consistency refers to the property that, as the sample size increases, the estimator converges to the true parameter value. If an estimator is consistent, it provides assurance that given a large enough sample size, the estimator will approach the true population parameter. In essence, a consistent estimator “self-corrects” as the sample size increases, which indicates that the sampling process is capturing the full breadth and characteristics of the population over time. Together, these properties provide a strong theoretical backbone for the representativeness of the sample (Keener 2010). Appendix E details our approach to assessing consistency, and offers formal definitions.

Finally, *efficiency* refers to whether an estimator makes good use of the data to produce estimates with low variance. An efficient estimator extracts more information from a given sample size than an inefficient one. If the estimator is efficient, it is a sign that the sample is “information-rich” and likely well-representative of the population, allowing for more reliable inferences. Appendix F explains our approach to measuring efficiency and includes formal definitions.

4.2 Results

4.2.1 Population estimates

Figure 2 presents the results of the orthogonal sampling approach for **#golf** and **#work**, as well as the results for the biased sampling approach on **#golf**. The horizontal axis plots population, and the vertical axis is the density with respect to where population estimates from bootstrap sampling. For each plot, the dashed yellow line indicates the population estimate computed from the sampling approach, the solid red line indicates the ground truth, and the magenta dotted lines indicate the 95% confidence intervals for the population estimate. Below each plot is a small table presenting the specific values for each of the vertical lines.

[Figure 2 about here]

For both orthogonal samples, **#golf** and **#work**, the true population parameter lies within the 95% confidence intervals for the population estimate. The estimate for **#golf** is 8,368, while the ground truth is 10,562, which shows an underestimated bias of 2,194. The population estimate for **#work** is 12,826, while the ground truth is 14,714. This is an underestimated bias of 1,888. We can see that the ground truth for **#work** is closer to the center of the distribution of population estimates, indicating that the true population is more frequently represented by the population estimates. The true population for **#golf** is closer to the upper bound of the 95% confidence interval, but still well within the interval.

For Biased **#golf** – generated through a sampling approach common in the literature and, as a result, having selection bias at the hashtag level – we observe poorer quality estimates. The population estimate for **#golf** here is 24,335, which is an overestimated bias of 13,773. We can also see that the true population parameter is outside the 95% confidence interval, suggesting that the sample is significantly different from the true population at $\alpha = 0.05$.

We additionally see in Figure 2 that the orthogonal samples for **#golf** and **#work** have relatively narrow 95% confidence intervals compared to the biased sample for **#golf**. The 95% confidence intervals for the orthogonal sample of **#golf** lies between 5,498 and 11,245, a span of 5,747. Note that this interval is 0.544 or 54.4% of the true parameter. For **#work** the 95% confidence interval has a lower bound of 7,787 and an upper bound of 17,407, a span of 9,620. This interval is 0.654 or 65.4% of the true parameter. The biased sample on the other hand shows a much wider 95% confidence interval for **#golf**. The lower bound is 11,452 and the upper bound is 34,356, a span of 22,904. This interval is 2.16 or 216% of the true population parameter. The wider confidence interval of the biased sample suggests that the selection bias produces a less efficient estimator. This will be tested more rigorously in the results section on efficiency.

Figure 3 shows the relative error or relative bias between population estimates (as defined in Appendix C). We see that both orthogonal samples present much smaller relative error values than the biased sampling approach. The orthogonal samples for `#golf` and `#work` show relative errors of 0.208 or 0.128, respectively. Estimates for `#golf` and `#work` were underestimates of the true population by 20.8% and 12.8%, respectively. The biased sample, on the other hand, shows a relative error of 1.304, which means that the error or bias of the estimate is larger than the true population parameter.

[Figure 3 about here]

4.2.2 Coverage rate

The coverage rate represents the probability that the true parameter falls within the 95% confidence interval of the estimated parameter (see Appendix D). We assess this by deriving 100 population estimates from bootstrap sampling, computing confidence intervals for each estimate, and counting how many times the true population parameter falls within the 95% confidence interval. In essence, the coverage rate observes a likelihood that the underlying sample is representative of the true population.

Figure 4 shows the results of the coverage assessment for both orthogonal samples and the biased sample. Both `#golf` and `#work` obtain a coverage rate of 1. These results suggest that the orthogonal samples for `#golf` and `#work` present good representation of the true population. For the biased sample of `#golf`, we see a coverage rate of 0.02, meaning we found only a probability of 0.02 that the true parameter fell within the 95% confidence interval of the estimated parameter. This suggests a very low probability that the biased sample is representative of the true population.

[Figure 4 about here]

4.2.3 Consistency

The consistency of the estimator given the sample evaluates whether the bias between the estimated parameter and the true parameter decreases as the sample size decreases (see Appendix E). This follows from the Law of Large Numbers and the Central Limit Theorem, that, if the sample is representative of the true population, then the estimated parameter should converge on the true parameter as the sample size increases toward infinity.

Figure 5 presents plots for both orthogonal samples and the biased sample. The horizontal axis indicates the size of the sample and the vertical axis indicates the relative error (see Appendix C) between the parameter estimate and the true parameter. For both orthogonal samples, the relative error between

the population estimate and the true parameter approaches 0 as sample size increases, with both curves monotonically decreasing.

For the biased sample, we also observe an improvement in relative error as sample size increases, but there are signs for concern. The curve is not monotonically decreasing. At size = 0.01 and 0.2 the relative errors *increase*. It is also important to observe that the relative errors are several multiples of the true parameter value (between 3.5 and 1.4). The relative errors also flatten out around 1.4, suggesting that the values are not approaching 0. The results confirm notably more consistency in estimates derived from the orthogonal samples than from the biased samples.

[Figure 5 about here]

4.2.4 Efficiency

Efficiency is an assessment of how small and stable the variance of one estimator is relative to another (see Appendix F). If one estimator has a smaller and more stable variance then it is considered more efficient. In our assessment of efficiency, we compare the efficiency of the population estimate for `#golf` under the orthogonal sampling approach to the biased sampling approach. We construct a statistical test for efficiency, by computing a 95% confidence interval of the variance in the population estimate under each sampling approach. By doing this, we observe not only how large the variance tends for each parameter estimate, but also, by looking at the confidence intervals, how stable these variances are. Our statistical assessment involves evaluating the overlap between the 95% confidence intervals of the orthogonal estimator and the biased estimator. If the confidence intervals do not overlap, then we conclude that the variances of the estimates are significantly different (at $\alpha = 0.05$).

Figure 6 shows 95% confidence intervals for `#golf` and Biased `#golf`. We immediately see the orthogonal sample produces much smaller variances, with the confidence interval falling between 2,172,313 and 2,212,706. The span of the confidence interval for the orthogonal sample is only 40,393, suggesting very stable variances for the population estimate. The biased sample presents much larger variance values. The lower bound of the 95% confidence interval is 34,464,304 and the upper bound is 35,143,436. This shows the variance of the population estimate under the biased sampling approach to be much larger. The span of the confidence interval for the biased sample is 679,132. The orthogonal sample produces a much smaller span, indicating that the estimate is more stable. We observe here, conclusively, that the orthogonal sample produces a much more efficient population estimate than the biased sample.

[Figure 6 about here]

5 Discussion and Conclusion

This paper is motivated by a problem affecting a segment of social science research. Studies that use social media data collected via platforms’ APIs are typically either limited to treating such data as non-representative samples or inappropriately as representative, likely resulting in misleading conclusions. This problem is due to the unknown ways in which APIs subset and filter data when using traditional requests, such as collecting data by asking for content based on a specific hashtag, a kind of topic-based sampling (Rafail 2018).

We have proposed an approach for gathering representative samples of social media content using traditional APIs that leverages the discursive space of the social media platform as a opportunity to reduce selection bias. Specifically, we reduce selection bias by gathering a sample of users and user content that has been captured independently of our topic of interest. This “orthogonal sampling approach” uses word embeddings to collect data from a large basket of hashtags or locations that have no salient similarity or dissimilarity to the hashtag of interest.

In the empirical demonstration, we implemented the orthogonal sampling approach on **#golf** and **#work** – two hashtags for which we could obtain a true population parameter, the abundance of the topic on a platform. From the orthogonal samples, we estimated the population parameter and compared it to the true population. We also conducted a comparison approach in which we implemented the sampling approach with deliberate selection bias on **#golf**. The intention of this placebo approach was to compare the effects of incurring selection bias to the orthogonal samples. When we compared the orthogonal sampling approach to the biased sampling approach, we found that not only do the orthogonal samples have lower relative error – bias relative to the true parameter – but that they satisfied three other assessments that provide strong justification for the orthogonal samples being representative of the true population. When measuring the coverage of estimates using both the **#golf** and **#work** samples, we observed a probability of 1 that the true parameter fell within the 95% confidence intervals of the estimated parameter, while for the biased golf sample we observed a probability of only 0.02. We also found that both orthogonal samples provided relatively consistent estimators, in that we saw a reduction in the bias between the estimated parameter and the true parameter as sample sizes grew. We did not observe much consistency for the biased sampling approach. Lastly, we observed that the orthogonal **#golf** provided population estimates that were significantly more efficient than the biased sampling approach.

The empirical demonstration provided a rigorous justification that the samples collected using the orthogonal approach are representative of the true population. Further, they suggest that as our sample size increases, so should the representativeness. The demonstration additionally showed, importantly, that selec-

tion bias has a very strong negative effect on properties that are characteristics of a representative sample, suggesting that mitigating sampling bias when collecting social media data should be a priority for any study that aims to analyze content as an extension of a larger group of individuals.

5.1 Limitations

Nearing the end of our data collection efforts in March and April 2023, Twitter changed much of its API framework. For our sampling approach, we relied on an Academic Twitter account, which allowed us to make thousands of API requests. As of writing this paper, comparable access to the Twitter API requires an enterprise account which can cost thousands per month. Most API endpoints have also been throttled down to limit the number of requests and to introduce more rate-limit delays, which can only be circumvented with account upgrades. Despite this, the implementation of the orthogonal approach would be similar on other platforms. Our approach is designed around hashtags or keywords, which are used on many other social media platforms with far larger numbers of users, such as Youtube and Reddit (Hosseinmardi et al. 2021).

The empirical demonstration itself faced some limitations related to the Twitter API. Because we could only get ground truth populations of hashtags that had a somewhat low abundance on Twitter, we were required to collect very large orthogonal samples because the capture probability for these hashtags were low. However, to collect a truly large enough sample would take too long due to rate-limit wait-times, so we gathered a small subset at each step (10% random samples), which we believe limited the accuracy of our parameter estimates (despite still being respectable). For typical use of the approach, we believe the topics of interest will likely be those that are significantly more prevalent, which should allow researchers to both collect less data and achieve more accurate results.

5.2 Extensions

Our method can be developed in several directions. Here, we highlight two which are likely to be among the most useful for future social science research. First, the method can be adapted to other social media platforms. After all, online discourse usually spans multiple platforms (Venturini et al. 2018; Özkula et al. 2023). During adaptation, the technical details will have to change – for example, sampling “repliers” to YouTube commenters rather than Twitter followers – but the core insights of our approach will remain the same. Chief among them is that it is valuable to set the initial locations for collection to topics that are orthogonal to the topic of interest.

A second direction for future development is to adjust our method to estimate characteristics of social

media users, such as behavior and location (rather than estimates of word use). This could be as simple as obtaining a sample of social media posts with certain words – say, tweets containing a hashtag of interest – using our method, then analyzing the behavior of the authors of these posts (*e.g.*, Barberá 2018). Or, it could entail a combination with other social media analysis techniques, such as using our approach to estimate the use of certain place names, then using methods for inferring users’ locations and spatial mobility from these names (see Zheng et al. 2018). These kinds of extensions – and most social science research using social media data to study the extent of a discourse – would benefit from treating the discursive space of interest (*e.g.*, a social media platform) geometrically and then using our approach, or similar ones, to sample the space with reduced bias.

Data availability statement

All data and code supporting the results presented in this paper will be made available on Open Science Foundation upon acceptance for publication.

References

- Alshaabi, Thayer, Jane L. Adams, Michael V. Arnold, Joshua R. Minot, David R. Dewhurst, Andrew J. Reagan, and Peter Sheridan Danforth, Christopher M. and Dodds. 2021. “Storywrangler: A Massive Exploratorium for Sociolinguistic, Cultural, Socioeconomic, and Political Timelines Using Twitter.” *Science Advances* 7.
- Bagadiya, Jimit. 2023. “500+ Social Media Statistics You Must Know in 2023.”
- Barberá, Pablo. 2018. “Explaining the Spread of Misinformation on Social Media: Evidence from the 2016 U.S. Presidential Election.” *APSA Comparative Politics Newsletter* .
- Beauchamp, Nicholas. 2017. “Predicting and Interpolating State-Level Polls Using Twitter Textual Data.” *American Journal of Political Science* 61:490–503.
- Blank, Grant. 2017. “The Digital Divide Among Twitter Users and Its Implications for Social Research.” *Social Science Computer Review* 35:679–697.
- boyd, danah and Kate Crawford. 2012. “Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon.” *Information, Communication & Society* 15:662–679.
- Burgess, Jean and Axel Bruns. 2015. *Compromised Data: From Social Media to Big Data*, chapter Easy Data, Hard Data: The Politics and Pragmatics of Twitter Research After the Computational Turn, pp. 93–111. Bloomsbury Publishing.
- Canty, Angelo and B. D. Ripley. 2022. *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-28.1.
- Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. University of California Press.
- Cooch, E.G. and Gary White. 2002. *Progam MARK: a gentle introduction*. <http://www.nativefishlab.net/library/internalpdf/21294.pdf>.
- Dehghan, Ehsan and Ashwin Nagappa. 2022. “Politicization and Radicalization of Discourses in the Alt-Tech Ecosystem: A Case Study on Gab Social.” *Social Media Society* 8:1–12.
- Flores, René D. 2017. “Do Anti-Immigrant Laws Shape Public Sentiment? A Study of Arizona’s SB 1070 Using Twitter Data.” *American Journal of Sociology* 123:333–384.
- Freese, Jeremy and Molly M. King. 2018. “Institutionalizing Transparency.” *Socius* 4:<https://doi.org/10.1177/2378023117739216>.

- Gerlitz, Carolin and Bernhard Rieder. 2013. "Mining One Percent of Twitter: Collections, Baselines, Sampling." *M/C Journal* 16.
- Hargittai, Eszter. 2020. "Potential Biases in Big data: Omitted Voices on Social Media." *Social Science Computer Review* 38:10–24.
- Hino, Airo and Robert A. Fahey. 2019. "Representing the Twittersphere: Archiving a Representative Sample of Twitter Data Under Resource Constraints." *International Journal of Information Management* 48:175–184.
- Hosseinmardi, Homa, Amir Chasemian, Aaron Clauset, Markus Mobius, David M. Rothschild, and Duncan J. Watts. 2021. "Examining the Consumption of Radical Content on YouTube." *Proceedings of the National Academy of Sciences* 118:e2101967118.
- Keener, Robert W. 2010. *Theoretical Statistics: Topics for a Core Course*. Springer, 2010th edition edition.
- Kim, Yoonsang, Rachel Nordgren, and Sherry Emery. 2020. "The Story of Goldilocks and Three Twitter APIs: A Pilot Study on Twitter Data Sources and Disclosure." *International Journal of Environmental Research and Public Health* 17:<https://doi.org/10.3390/ijerph17030864>.
- King, Gary, Patrik Lam, and Margaret E. Roberts. 2017. "Computer-Assisted Keyword and Document Set Discovery from Unstructured Text." *American Journal of Political Science* 61:971–988.
- Larson, Jennifer M., Jonathan Nagler, Jonathan Ronen, and Joshua A. Tucker. 2019. "Social Networks and Protest Participation: Evidence from 130 Million Twitter Users." *American Journal of Political Science* 63:690–705.
- Lorentzen, David Gunnarsson and Jan Nolin. 2017. "Approaching Completeness: Capturing a Hashtagged Twitter Conversation and Its Follow-On Conversation." *Social Science Computer Review* 35:277–286.
- Malik, Momin M., Hemank Lamba, Constantine Nakos, and Jurgen Pfeffer. 2017. "Population Bias in Geotagged Tweets." *9th International AAAI Conference on Weblogs and Social Media* 35:18–27.
- Marres, Noortje. 2017. *Digital Sociology: The Reinvention of Social Research*. Polity Press.
- McCormick, Tyler H., Hedwig Lee, Nina Cesare, Ali Shojaie, and Emma S. Spiro. 2017. "Using Twitter for Demographic and Social Science Research: Tools for Data Collection and Processing." *Sociological Methods & Research* 46:390–421.
- Özkula, Suay M., Paul J. Reilly, and Jenny Hayes. 2023. "Easy data, same old platforms? A Systematic Review of Digital Activism Methodologies." *Information, Communication & Society* 26:1470–1489.

- Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. “GloVe: Global Vectors for Word Representation.” In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Pfeffer, Jürgen, Katja Mayer, and Fred Morstatter. 2018. “Tampering with Twitter’s Sample API.” *EPJ Data Science* 7:<https://doi.org/10.1140/epjds/s13688-018-0178-0>.
- Powell, Larkin and George Gale. 2015. *Estimation of Parameters for Animal Populations: a primer for the rest of us*. Caught Napping Publications.
- Rafail, Patrick. 2018. “Nonprobability Sampling and Twitter: Strategies for Semibounded and Bounded Populations.” *Social Science Computer Review* 36:195–211.
- Rinker, Tyler W. 2023. *qdap: Quantitative Discourse Analysis Package*. Buffalo, New York. 2.4.6.
- Sachs, Jeffrey, Rahshemah Wise, and Daniel Karell. 2021. “The TikTok Self: Music, Signaling, and Identity on Social Media.” *Open Science Foundation* DOI:10.31235/osf.io/2rx46.
- Salganik, Matthew J. 2018. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.
- Steinert-Threlkeld, Zachary C. 2017. “Spontaneous Collective Action: Peripheral Mobilization During the Arab Spring.” *American Political Science Review* 111:379–403.
- van der Veen, Maurits. 2022. “Blmtwitter: The Black Lives Matter (BLM) Twitter Corpus.” *SocArXiv* p. doi:10.31235/osf.io/kna9s.
- Venturini, Tommaso, Liliana Bounegru, Jonathan Gray, and Richard Rogers. 2018. “A Reality Check(list) for Digital Methods.” *New Media & Society* 20:4195–4217.
- Zheng, Xin, Jialong Han, and Aixin Sun. 2018. “A Survey of Location Prediction on Twitter.” *IEEE Transactions on Knowledge and Data Engineering* 30:1652–1671.

Appendix A Details of the GloVe model

Mathematically, GloVe aims to solve the following optimization problem:

$$J = \sum_{i,j=1}^V f(X_{ij}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log(X_{ij}) \right)^2$$

Here, J is the objective function to minimize, w_i and \tilde{w}_j are the word vectors for words i and j , b_i and \tilde{b}_j are the corresponding biases, X_{ij} is the number of times word j appears in the context of word i , given a context window parameter, and f is a weighting function to prevent common pairs from heavily leveraging training. In this equation, V is the vocabulary size.

GloVe embeddings provide a continuous, multi-dimensional space where semantic relationships between words (or hashtags, in your case) are encoded as geometric relationships. GloVe embeddings are trained on a massive corpus (in our case, 2 billion tweets), which allows them to capture complex relationships that simple methods like co-occurrence counts or Jaccard similarity cannot. They can pick up on syntactic and semantic nuances that make them a better gauge for true independence. Since GloVe is trained on a large corpus, it is likely to provide a more generalizable form of independence that is more reflective of the broader usage of the hashtag, not just a local or community-specific usage.

Appendix B Details of cosine similarity

We used cosine similarity to measure orthogonality between our hashtag of interest and other possible hashtags. Cosine similarity measures the cosine of the angle between two non-zero vectors in an inner product space. It is a mathematically robust measure for independence in a vector space. Cosine similarity is defined by the equation below:

$$\text{cosine similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

In the preceding equation, A and B represent two non-zero vectors for which you want to calculate cosine similarity. The equation evaluates the dot product of the two vectors divided by the Euclidean norms or magnitudes of the two vectors. Conventionally, cosine similarity returns a value between -1 and 1, with 0 defined as orthogonality. Cosine similarity values closer to 1 indicate words that are more similar, and values closer to -1 indicate words that are opposing. In an embedding space, words that are similar or opposite correlate in either a positive or negative direction. Our interest is in words that have a cosine similarity with our hashtag of interest of close to 0, indicating independence or a lack of an association.

Appendix C Measuring practical performance

Evaluating practical performance involves comparing the estimates derived from the orthogonal sampling approach to the true parameter values. To do this, we provide numerical values for the population estimates, as well as the true population value. We estimate the population by bootstrapping our sample, and in the results section provide density plots for the 1000 iterations of the bootstrapped parameter estimate. We also provide confidence intervals for the estimated population parameter, as well as the relative error, defined below:

$$\text{relative error} = \frac{|\hat{\theta} - \theta|}{|\theta|}$$

where $\hat{\theta}$ denotes our population estimate and θ is the true population parameter.

The relative error calculates the proportional difference between the estimated parameter and the true parameter. Evaluating the quality of the parameter estimate using relative error is challenging, however, in that benchmarks for population estimates on social media content, to our knowledge, do not exist. For this reason, we compare the relative error of the estimates derived from the orthogonal sampling approach to that of the biased sampling approach. We also acknowledge that the relative error of our implementation of the sampling approach could be larger due to the fact that our sample is a very small subsample of the what could have been gathered. This choice was made to shorten the duration of data collection and to collect within certain limits on computing power. For this reason, the three other evaluation strategies (*i.e.*, coverage, efficiency, consistency) serve as more objective measures of the quality of the sample.

Appendix D Measuring coverage

The coverage assessment involves checking whether the true parameter falls within the confidence interval with a particular frequency. Ideally, for a 95% confidence interval, the true parameter should fall within the interval in 95 out of 100 samples in which we estimate the population parameter $\hat{\theta}$. In our case, we have the population estimate \hat{P} , defined in Equation (3). The coverage assessment is defined as follows:

Let S_1, S_2, \dots, S_M be samples from M bootstraps of the original sample. For each S_i , we compute \hat{P} and confidence interval CI_i , where $CI_i = [\hat{P}_{i,lower}, \hat{P}_{i,upper}]$. With estimated confidence intervals, we then count the number of times the true parameter falls within the estimated confidence intervals as follows:

$$C = \sum_{i=1}^M I \in \{0, 1\},$$

where $I = 1$ if $P \in CI_i$ and $I = 0$ otherwise. We then calculate the coverage rate, defined below:

$$\text{coverage rate} = \frac{C}{M}$$

Here, C is the number of times the true parameter, P , falls within the confidence intervals for the estimated parameter, \hat{P} , and M equals the total number of bootstrapped samples. In our implementation, we estimate confidence intervals for our parameter, \hat{P} , by performing bootstrap sampling with 1000 iterations. We then implement this calculation $M = 100$ times to calculate the coverage rate defined above.

Appendix E Measuring consistency

An estimator is consistent if it converges in probability to the parameter it estimates as the sample size approaches infinity. This is in line with the Law of Large Numbers, which states that as the sample size increases to infinity, the estimated parameter will converge to the true population parameter. Empirically testing the consistency of our parameter estimate \hat{P} involves looking at its convergence with the true parameter value as sample sizes under the sampling approach increase. Mathematically, the estimator \hat{P}_n is consistent with P if:

$$\lim_{n \rightarrow \infty} \text{Bias}(\hat{P}_n) = \lim_{n \rightarrow \infty} (\hat{P}_n - P) = 0.$$

For practical purposes, the consistency of our estimator \hat{P}_n is defined as:

$$\text{If } n \rightarrow \infty, \text{ then } \hat{P}_n \xrightarrow{P} P.$$

We implement a test for consistency by taking K random subsets of the original sample, S of increasing size and measuring the bias of our population estimate with respect to the true parameter value. The test is defined as follows: Let the original sample S consist of N observations. We take K random subsets of the original sample S , where $|K| = 10$ and $K = \{f_1, f_2, \dots, f_K\}$. Here, f_i denotes a fraction, where $N \times f_i$ denotes a random subset of S . The set $K = \{f_1, f_2, \dots, f_K\}$ are fractions of increasing size and are spaced logarithmically. This allows us to see how bias changes not just for small to moderate sample sizes but also for very large ones. In essence, we estimate the population parameter \hat{P} from across random subsets of the original sample S . These subsets increase in size, and thus allow us to empirically observe the difference between the estimated parameter and the true parameter as the sample size increases. If the estimated parameter is consistent, we should see bias diminish as the sample size increases. In our implementation, we

actually take 100 samples at each sample size f_i and compute average measures of the population estimate \hat{P} to reduce the sensitivity of the population estimate to specific random samples. Importantly, a consistent estimator should help justify the quality of the estimate despite a potentially poor relative error calculation. This is because, if the estimator is consistent, then the estimator should converge on the true parameter, and thus reduce the relative error, with a larger sample.

To conduct this assessment, we use 10 fractions of the sample between 0.01 and 0.5, spaced logarithmically $K = \{0.010, 0.015, 0.024, 0.037, 0.057, 0.088, 0.136, 0.210, 0.324, 0.500\}$, and bias is computed using relative error (see Appendix C).

Appendix F Measuring efficiency

Efficiency refers to the variance of an estimator, in our case, \hat{P} . A more efficient estimator will have a smaller variance, which in turn translates to narrow confidence intervals. Assessing efficiency involves a comparison, often between estimators. However, in our case, we measure efficiency between two sampling approaches: the orthogonal approach implemented on `#golf` as well as the biased approach implemented on `#golf`. Commonly, relative efficiency might be used as a test of efficiency, defined below:

$$\text{relative efficiency} = \frac{Var(\hat{P}_1)}{Var(\hat{P}_2)},$$

where $Var(\hat{P}_1)$ is the variance of the population estimate under one sampling approach and $Var(\hat{P}_2)$ is the variance of the population estimate under the second approach. We, however, view this approach as too interpretive, and too sensitive to the specific samples underlying each estimate. We therefore introduce an approach to testing efficiency that accounts for sampling variation under the orthogonal approach and the biased approach.

In this approach, we calculate multiple variances for the estimated the population parameter, \hat{P} and compute confidence intervals for variances under each sampling approach. Let M denote a set of samples $M = \{B_1, B_2, \dots, B_M\}$, where each sample B_i is a bootstrapped set of n samples, K , where $k \in \{k_1, k_2, \dots, k_n\}$. For each set of bootstrapped samples, B_i , we compute the variance of the n population estimates, $Var(\hat{P}_{B_i})$. From the M variances of the population estimate, $Var(\hat{P}_{B_i})$, we compute 95% confidence intervals, $CI_{var(\hat{P})} = [CI_{var(\hat{P}),lower}, CI_{var(\hat{P}),upper}]$.

We then construct a statistical test to compare these confidence intervals for the orthogonal sampling approach to the biased sampling approach. Let $CI_{Var,Ortho}$ denote the 95% confidence intervals for the variance of the population estimate under the orthogonal sampling approach, and $CI_{Var,Biased}$ denote the 95%

confidence intervals for the variance of the population estimate under the biased sampling approach. We observe the differences between the variance confidence intervals for each approach. The statistical test involves observing possible overlap between the two confidence intervals. If the 95% confidence intervals for the population estimates from the orthogonal sample and the biased sample do not overlap, then we conclude that the difference between the variances is statistically significant. Given $CI_{Var,Ortho} = [L_{Var,Ortho}, U_{Var,Ortho}]$ and $CI_{Var,Biased} = [L_{Var,Biased}, U_{Var,Biased}]$, where L and U denote the lower and upper boundary of the 95% confidence interval, our test for overlap can be defined as:

$$\text{overlap} = \max(0, \min(U_{Var,Ortho}, U_{Var,Biased}) - \max(L_{Var,Ortho}, L_{Var,Biased}))$$

If there is no overlap between the confidence intervals of the variances of the two sampling approaches, then the test results in 0. In this case, the lower bound CI of one sampling approach is higher than the upper bound CI of the other approach, indicating statistical significance at $\alpha = 0.05$. We use this assessment as a more rigorous test of efficiency for each sampling approach. In our implementation, we compute the 95% confidence intervals by computing the variance for each sampling approach using 100 bootstrap samples of variances. Each variance of the population estimate is computed from 1000 bootstrap samples of the original sample.

Orthogonal Sampling Approach

1. Identify hashtag(s) of interest

#interest

Hashtag(s) of Interest

2. Using a GloVe embedding space pre-trained on 2B tweets, create large basket of "orthogonal" hashtags that show independence from hashtag of interest.

$\perp \#_1$ $\perp \#_2$... $\perp \#_i$
Orthogonal Hashtag Orthogonal Hashtag ... Orthogonal Hashtag




Reasoning: we want to eliminate selection bias from looking at hashtags that in some way correlate with our hashtag of interest.

3. From EACH orthogonal hashtag, take a small random sample of users.

Note: we may deem some users to be of poor quality -- bots, advertisements, etc. -- and we are limited to public accounts.

 ₁  ₂ ...  _i
User User ... User

4. From EACH sampled user (for applicable users), collect a small random sample of their followers.

 ₁  ₂ ...  _i
Follower Follower ... Follower

5. From EACH sampled user (for applicable users), collect tweets for given day.

 ₁  ₂ ...  _i

Figure 1: Schema of the sampling approach

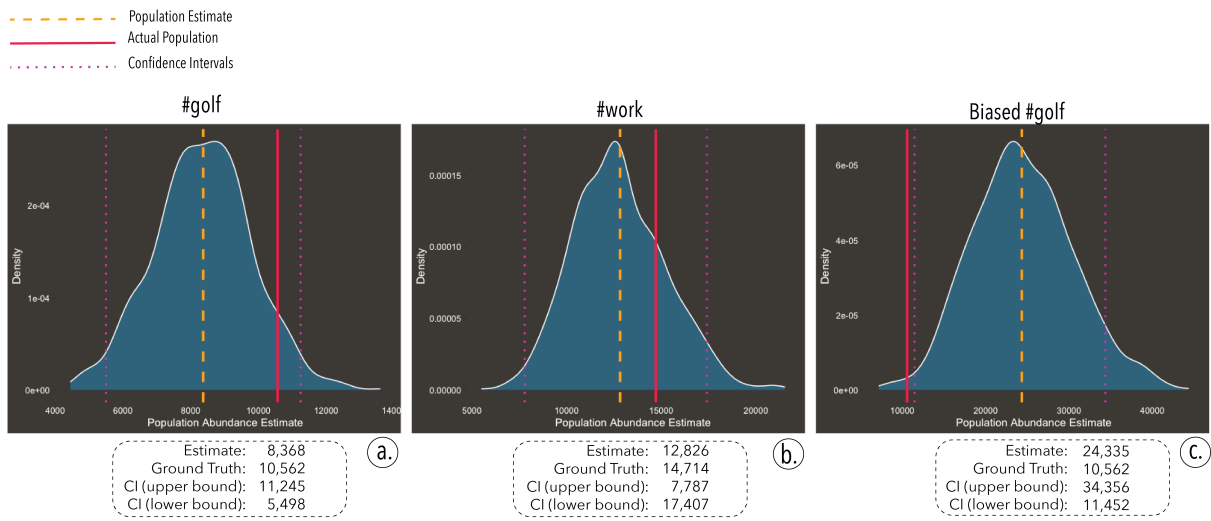


Figure 2: Results of the sampling approach

#golf	#work	Biased #golf
<i>0.208</i>	<i>0.128</i>	<i>1.304</i>

Figure 3: Relative error (bias)

#golf	#work	Biased #golf
<i>1.00</i>	<i>1.00</i>	<i>0.02</i>

Figure 4: Coverage assessment

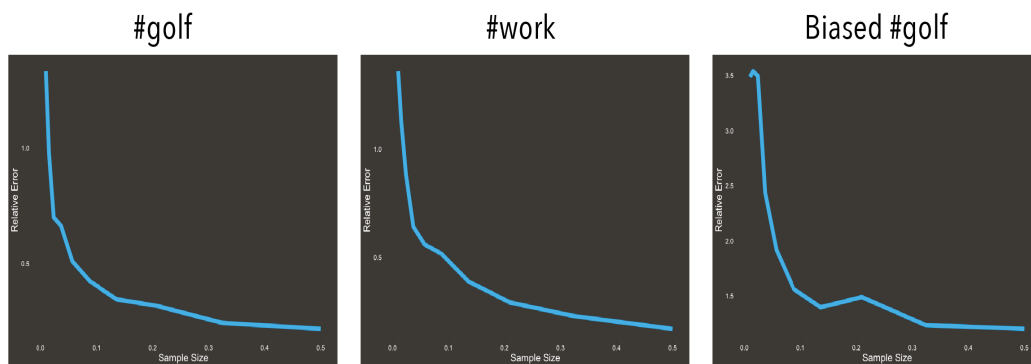


Figure 5: Consistency assessment

#golf	:	Biased #golf	:	Overlap $\alpha = 0.05$
$Cl_{lower}: 2,172,313$:	$Cl_{lower}: 34,464,304$:	0
$Cl_{upper}: 2,212,706$:	$Cl_{upper}: 35,143,436$:	

Figure 6: Efficiency assessment