Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

Variation across Scales: Measurement Fidelity under Twitter Data Sampling

Siqi Wu,[1,3] **Marian-Andrei Rizoiu,**[2,3] **Lexing Xie**[1,3]
[1]Australian National University, [2]University of Technology Sydney, [3]Data 61, CSIRO, Australia
{siqi.wu, lexing.xie}@anu.edu.au, marian-andrei.rizoiu@uts.edu.au

Abstract

A comprehensive understanding of data quality is the cornerstone of measurement studies in social media research. This paper presents in-depth measurements on the effects of Twitter data sampling across different timescales and different subjects (entities, networks, and cascades). By constructing complete tweet streams, we show that Twitter rate limit message is an accurate indicator for the volume of missing tweets. Sampling also differs significantly across timescales. While the hourly sampling rate is influenced by the diurnal rhythm in different time zones, the millisecond level sampling is heavily affected by the implementation choices. For Twitter entities such as users, we find the Bernoulli process with a uniform rate approximates the empirical distributions well. It also allows us to estimate the true ranking with the observed sample data. For networks on Twitter, their structures are altered significantly and some components are more likely to be preserved. For retweet cascades, we observe changes in distributions of tweet inter-arrival time and user influence, which will affect models that rely on these features. This work calls attention to noises and potential biases in social data, and provides a few tools to measure Twitter sampling effects.

1 Introduction

"Polls are just a collection of statistics that reflect what people are thinking in 'reality'. And reality has a well-known liberal bias." – Stephen Colbert[1]

Data quality is a timely topic that receives broad attention. The data noises and biases particularly affect data-driven studies in social media (Tufekci 2014; Olteanu et al. 2019). Overrepresented or underrepresented data may mislead researchers to spurious claims (Ruths and Pfeffer 2014). For example, opinion polls wrongly predicted the U.S. presidential election results in 1936 and 1948 because of unrepresentative samples (Mosteller 1949). In the era of machine learning, the data biases can be amplified by the subsequent models. For example, models overly classify agents doing cooking activity as female due to overrepresented correlations (Zhao et al. 2017), or lack the capacity to identify dark-skinned women due to underrepresented data (Buolamwini and Gebru 2018). Hence, researchers must be aware and take account of the hidden biases in their datasets for drawing rigorous scientific conclusions.

Twitter is the most prominent data source in ICWSM – 82 (31%) out of 265 full papers in the past 5 years (2015-2019) used Twitter data (listed in Section A of (Appendix 2020)), in part because Twitter has relatively open data policies, and in part because Twitter offers a range of public application programming interfaces (APIs). Researchers have used Twitter data as a lens to understand political elections (Bovet and Makse 2019), social movements (De Choudhury et al. 2016), information diffusion (Zhao et al. 2015), and many other social phenomena. Twitter offers two streaming APIs for free, namely *sampled* stream and *filtered* stream. The filtered stream tracks a set of keywords, users, languages, and locations. When the matched tweet volume is above a threshold, Twitter subsamples the stream, which compromises the completeness of the collected data. In this paper, we focus on empirically quantifying the data noises resulted from the sampling in the filtered stream and its impacts on common measurements.

This work addresses two open questions related to Twitter data sampling. Firstly, **how are the tweets missing in the filtered stream?** The sampling mechanism of the sampled stream has been extensively investigated (Kergl, Roedler, and Seeber 2014; Pfeffer, Mayer, and Morstatter 2018), but relatively little is said about the filtered stream. Since the two streaming APIs are designed to be used in different scenarios, it is pivotal for researchers who use the filtered stream to understand what, when, and how much data is missing. Secondly, **what are the sampling effects on common measurements?** Our work is inspired by Morstatter et al. (2013), who measured the discrepancies of topical, network, and geographic metrics. We extend the measurements to entity frequency, entity ranking, bipartite graph, retweet network, and retweet cascades. The answers to these questions not only help researchers shape appropriate questions, but also help platforms improve their data services.

We address the first question by curating two datasets that track suggested keywords in previous studies. Without leveraging the costly Twitter Firehose service, we construct the complete tweet streams by splitting the keywords and languages into multiple subcrawlers. We study the Twitter rate limit messages. Contradicting observations made by Sampson et al. (2015), our results show that the rate limit mes-

[1]At the 2006 White House Correspondents' Dinner.

sages closely approximate the volume of missing data. We also find that tweets are not missing at random since the sampling rates have distinct temporal variations across different timescales, especially at the level of hour and millisecond.

Addressing the second question, we measure the effects of Twitter data sampling across different subjects, e.g., the entity frequency, entity ranking, user-hashtag bipartite graph, retweet network, and retweet cascades. We find that (1) the Bernoulli process with a uniform rate can approximate the empirical entity distribution well; (2) the ranks of top entities are distorted; (3) the true entity frequency and ranking can be inferred based on sampled observations; (4) the network structures change significantly with some components more likely to be preserved; (5) sampling compromises the quality of diffusion models as the distributions of tweet inter-arrival time and user influence are substantially skewed. We remark that this work only studies the effects of Twitter sampling mechanism, but does not intend to reverse engineer it.

The main contributions of this work include:

- We show that Twitter rate limit message is an accurate indicator for the volume of missing tweets.
- A set of measurements on the Twitter data sampling effects across different timescales and different subjects.
- We show how to estimate the entity frequency and ranking of the complete data using only the sample data.
- We release a software package "Twitter-intact-stream" for constructing the complete data streams on Twitter[2].

2 Related work

Studies on Twitter APIs. Twitter has different levels of access (Firehose, Gardenhose, Spritzer) and different ways to access (search API, sampled stream, filtered stream). As the complete data service (Firehose) incurs excessive costs and requires severe storage loads, we only discuss the free APIs.

- Twitter search API returns relevant tweets for a given query, but it only fetches results published in the past 7 days (Twitter.com 2020d). The search API also bears the issue of data attrition. Research using this API to construct a "complete" dataset would inevitably miss parts of desired tweets (Wang, Callan, and Zheng 2015) since tweet creation and deletion are highly dynamic (Almuhimedi et al. 2013). To overcome this limitation, researchers can pivot to the streaming APIs.
- Twitter sampled streaming API returns roughly 1% of all public tweets in realtime (Twitter.com 2020c). Pfeffer, Mayer, and Morstatter (2018) detailed its sampling mechanism and identified potential tampering behaviors. González-Bailón et al. (2014) examined the biases in the retweet network from the 1% sample and the search API. While the 1% sample may be treated as a representative sample of overall Twitter activities (Morstatter, Pfeffer, and Liu 2014; Kergl, Roedler, and Seeber 2014), data filtering can only be conducted post data collection. Therefore, it is not suitable to create ad hoc datasets, e.g., tracking *all* tweets that contain the hashtag `#metoo`.
- Twitter filtered streaming API collects tweets matching a set of prescribed predicates in realtime (Twitter.com 2020a). Suppose that the streaming rate is below Twitter limit, the pre-filtering makes the filtered stream possible to construct the complete datasets without using the costly Firehose stream, e.g., on social movements (De Choudhury et al. 2016) and on news outlets (Mishra, Rizoiu, and Xie 2016). We focus on the scenes where the data streams are sampled. The most relevant work is done by Morstatter et al. (2013), in which they compared the filtered stream to the Firehose, and measured the discrepancies in various metrics. We extend the scope of measured subjects. Furthermore, we take a step to correct the sampling effects on entity measures.

Twitter sampling is deterministic (Joseph, Landwehr, and Carley 2014), therefore, simply stacking crawlers with the same predicates will not yield more data. However, users can improve the sample coverage by splitting the keyword set into multiple disjoint predicate sets, and monitoring each set with a distinct subcrawler (Sampson et al. 2015).

Effects of missing social data. Social data, which records ubiquitous human activities in digital form, plays a fundamental role in social media research. Researchers have pointed out the necessity to interrogate the assumptions and biases in data (Boyd and Crawford 2012; Ruths and Pfeffer 2014). Tufekci (2014) outlined four issues on data representativeness and validity. The hidden data biases may alter some research conclusions and even impact human decision making (Olteanu et al. 2019).

Gaffney and Matias (2018) identified gaps where data is unevenly missing in a widely used Reddit corpus. They suggested strong risks in research that concerns user history or network information, and moderate risks in research that uses aggregate counts. In this work, we use these qualitative observations as starting points and present a set of in-depth quantitative measurements. We corroborate the risks in user history study and network analysis. And we show how the complete counting statistics can be estimated.

Sampling from graphs and cascades. Leskovec and Faloutsos (2006) studied different graph sampling strategies for drawing representative samples. Wagner et al. (2017) considered how sampling impacts the relative ranking of groups in the attributed graphs. The effects of graph sampling has been extensively discussed by Kossinets (2006). In this work, the missing tweets can cause edge weights to decrease, and some edges to even disappear. On sampling a cascade, De Choudhury et al. (2010) found that combining network topology and contextual attributes distorts less the observed metrics. Sadikov et al. (2011) proposed a k-tree model to uncover some properties from the sampled data. They both sampled the cascades via different techniques (e.g., random, forest fire) and varying ratios. In contrast, the sampling in this work is an artifact of proprietary Twitter sampling mechanisms, and beyond the control of the users.

[2]The package, collected data, and analysis code are publicly available at https://github.com/avalanchesiqi/twitter-sampling

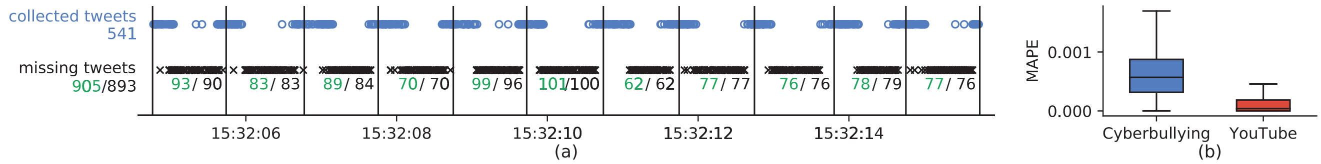

Figure 1: (a) Collected and missing tweets in an 11-second interval. blue circle: collected tweet; black cross: missing tweet; black vertical line: rate limit message. green number: estimated missing volume from rate limit messages; black number: count of missing tweets compared to the complete set. **(b)** MAPE of estimating the missing volumes in the rate limit segments.

3 Datasets and Twitter rate limit messages

We collect two datasets, using two sets of keywords employed in recent large-scale studies that use Twitter. We choose these works because they are high volume and informative for important social science problems (cyberbullying and online content sharing). We use ρ to denote the sampling rate – i.e., the probability that a tweet is present in the collected (sampled) dataset. We use subscripts to differentiate sampling rates that vary over time ρ_t, users ρ_u, networks ρ_n, and cascades ρ_c. The datasets are collected using the Twitter filtered streaming API and are summarized in Table 1.

- CYBERBULLYING (Nand, Perera, and Kasture 2016): This dataset tracks all tweets that mention any of the 25 recommended keywords from psychology literature (e.g., `gay`, `slut`, `loser`). The collection period is from 2019-10-13 to 2019-10-26.
- YOUTUBE (Rizoiu et al. 2017): This dataset tracks all tweets that contain at least one YouTube video URL by using the rule "`youtube`" `OR` ("`youtu`" `AND` "`be`"). The collection period is from 2019-11-06 to 2019-11-19.

The streaming client is a program that receives streaming data via Twitter API. The client will be rate limited if the number of matching tweets exceeds a preset threshold – 50 tweets per second as of 2020-03 (Twitter.com 2020b). When we use only one client to track all keywords, we find that both datasets trigger rate limiting. We refer to the crawling results from a single client as the *sample set*.

We develop a software package "Twitter-intact-stream"[3] for constructing the complete data streams on Twitter. The package splits the filtering predicates into multiple subsets, and tracks each set with a distinct streaming client. The CYBERBULLYING and YOUTUBE datasets are respectively crawled by 8 and 12 clients based on different combinations of keywords and languages. We remove the duplicate tweets and sort the distinct tweets chronologically. We refer to the crawling results from multiple clients as the *complete set*.

In very occasional cases, the complete sets also encounter rate limiting. Estimated from the rate limit messages (detailed next), 0.04% and 0.14% tweets in the complete sets are missing, which are negligible comparing to the volumes of missing tweets in the sample sets (47.28% and 8.47%, respectively). For rigorous comparison, we obtain a 30 minutes complete sample from Twitter Firehose and find the difference with our collected data is trivial (detailed in Section C of (Appendix 2020)). Hence, for the rest of this work, we treat the complete sets as if they contain no missing tweets.

[3] https://github.com/avalanchesiqi/twitter-intact-stream

	CYBERBULLYING		YOUTUBE	
	complete	sample	complete	sample
N_c	114,488,537	60,400,257	53,557,950	49,087,406
N_r	3,047	1,201,315	3,061	320,751
$\hat{N}_m$	42,623	54,175,503	77,055	4,542,397
$\bar{\rho}$	99.96%	52.72%	99.86%	91.53%

Table 1: Summary of two datasets. N_c: #collected tweets; N_r: #rate limit messages; $\hat{N}_m$: #estimated missing tweets; $\bar{\rho}$: mean sampling rate. Full specifications for all streaming clients are listed in Section B of (Appendix 2020).

Validating Twitter rate limit messages. When the streaming rate exceeds the threshold, Twitter API emits a rate limit message that consists of a timestamp and an integer. The integer is designed to indicate the cumulative number of missing tweets since the connection starts (Twitter.com 2020e). Therefore, the difference between 2 consecutive rate limit messages should estimate the missing volume in between.

We empirically validate the rate limit messages. We divide the datasets into a list of segments where (a) they contain no rate limit message in the complete set; (b) they are bounded by 2 rate limit messages in the sample set. This yields 1,871 and 253 segments in the CYBERBULLYING and YOUTUBE datasets, respectively. The lengths of segments range from a few seconds to several hours, and collectively cover 13.5 days out of the 14-day crawling windows. In this way, we assure that the segments in the complete set have no tweet missing since no rate limit message is received. Consequently, for each segment we can compute the volume of missing tweets in the sample set by either computing the difference of the two rate limit messages bordering the segment, or by comparing the collected tweets with the complete set. Figure 1(a) illustrates the collected and missing tweets in an 11-second interval. The estimated missing volumes from rate limit messages closely match the counts of the missing tweets in the complete set. Overall, the median error in estimating the missing volume using rate limit messages is less than 0.0005, measured by mean absolute percentage error (MAPE) and shown in Figure 1(b). We thus conclude that the rate limit message is an accurate indicator for the number of missing tweets. Note that it only approximates the volume of missing tweets, but not the content.

Our observations contradict those from Sampson et

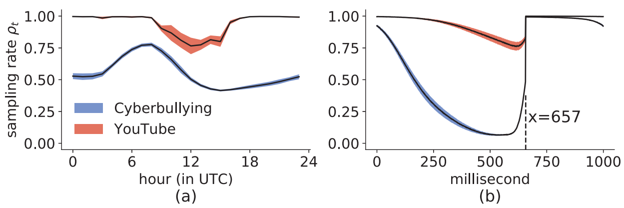

Figure 2: Sampling rates are uneven **(a)** in different hours or **(b)** in different milliseconds. black line: temporal mean sampling rates; color shades: 95% confidence interval.

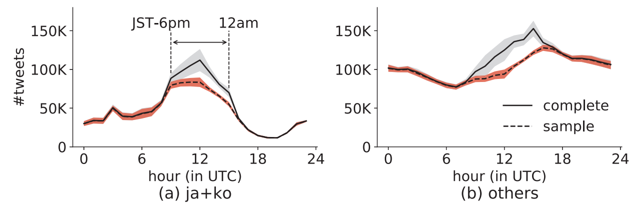

Figure 3: Hourly tweet volumes in YOUTUBE dataset. **(a)** Japanese+Korean; **(b)** other languages. black line: temporal mean tweet volumes; color shades: 95% confidence interval.

al. (2015), who used the same keyword-splitting approach, yet found that the rate limit messages give inaccurate estimations. They consistently retrieved more distinct tweets (up to 2 times) than the estimated total volume, i.e., the number of collected tweets plus the estimated missing tweets. In contrast, our datasets only have a small deviation (0.08% and 0.13%, comparing N_c of the complete set to $N_c+\hat{N}_m$ of the sample set in Table 1). This discrepancy is due to a different implementation choice back in 2015 – instead of having 1 rate limit message for each second, the rate limit messages were spawned across 4 threads, resulting in up to 4 messages per second. We provide a detailed analysis in Section C of (Appendix 2020).

4 Are tweets missing at random?

In this section, we study the randomness of Twitter sampling – do all tweets share the same probability of missing? This is relevant because uniform random sampling creates representative samples. When the sampling is not uniform, the sampled set may suffer from systematic biases, e.g., some tweets have a higher chance of being observed. Consequently, some users or hashtags may appear more often than their cohorts. We tackle the uniformity of the sampling when accounting for the tweet timestamp, language, and type.

Tweet timestamps. Figure 2(a) plots the hourly sampling rates. CYBERBULLYING dataset has the highest sampling rate (ρ_t=78%) at UTC-8. The lowest sampling rate (ρ_t=41%) occurs at UTC-15, about half of the highest value. YOUTUBE dataset is almost complete (ρ_t=100%) apart from UTC-8 to UTC-17. The lowest sampling rate is 76% at UTC-12. We posit that the hourly variation is related to the overall tweeting dynamics and the rate limit threshold (i.e., 50 tweets per second): higher tweet volumes yield lower sampling rates. Figure 2(b) shows the sampling rate at the millisecond level, which curiously exhibits a periodicity of one second. In CYBERBULLYING dataset, the sampling rate peaks at millisecond 657 (ρ_t=100%) and drops monotonically till millisecond 550 (ρ_t=6%) before bouncing back. YOUTUBE dataset follows a similar trend with the lowest value (ρ_t=76%) at millisecond 615. This artifact leaves the sample set vulnerable to automation tools. Users can deliberately schedule tweet posting time within the high sampling rate period for inflating their representativeness, or within the low sampling rate period for masking their content in the public API. The minutely and secondly sampling rates are included in Section D of (Appendix 2020).

	CYBERBULLYING		YOUTUBE	
	complete	sample	complete	sample
%root tweets	14.28%	14.26%	25.90%	26.19%
%retweets	64.40%	64.80%	62.92%	62.51%
%quotes	7.37%	7.18%	3.44%	3.40%
%replies	13.94%	13.76%	7.74%	7.90%

Table 2: The ratios of the 4 tweet types (root tweet, retweet, quote, and reply) in the complete and the sample sets.

Tweet languages. Some languages are mostly used within one particular timezone, e.g., Japanese and Korean[4]. The temporal tweet volumes for these languages are related to the daily activity patterns in the corresponding countries. We break down the hourly tweet volumes of YOUTUBE dataset into Japanese+Korean and other languages. The results are shown in Figure 3. Altogether, Japanese and Korean account for 31.4% tweets mentioning YouTube URLs. The temporal variations are visually different – 48.3% of Japanese and Korean tweets are posted in the evening of local time (JST-6pm to 12am), while tweets in other languages disperse more evenly. Because of the high volume of tweets in this period, sampling rates within UTC-9 to UTC-15 are lower (see Figure 2(a)). Consequently, "ja+ko" tweets are less likely to be observed (89.0% in average, 80.9% between JST-6pm and 12am) than others (92.9% in average).

Tweet types. Twitter allows the creation of 4 types of tweets. The users create a *root tweet* when they post new content from their home timelines. The other 3 types are interactions with existing tweets: *retweets* (when users click on the "Retweet" button); *quotes* (when users click on the "Retweet with comment" button); *replies* (when users click on the "Reply" button). The relative ratios of different types of tweets are distinct for the two datasets (see Table 2). CYBERBULLYING has higher ratios of retweets, quotes, and replies than YOUTUBE, implying more interactions among users. However, the ratios of different types are very similar in the sampled versions of both datasets (max deviation=0.41%, retweets in YOUTUBE dataset). We conclude that Twitter data sampling is not biased towards any tweet type.

[4]Japanese Standard Time (JST) and Korean Standard Time (KST) are the same.

	complete	sample	%miss	est. %miss
#users	19,802,506	14,649,558	26.02%	26.12%
#hashtags	1,166,483	880,096	24.55%	24.31%
#URLs	467,941	283,729	39.37%	38.99%

Table 3: Statistics of entities in CYBERBULLYING dataset, mean sampling rate $\bar{\rho}$=52.72%.

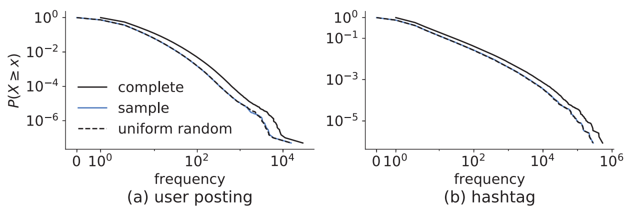

Figure 4: The frequency distributions of **(a)** user posting and **(b)** hashtag. The x-axis starts at 0 rather than 1, as the sample set and uniform random sample both have missing entities.

5 Impacts on Twitter entities

In this section, we study how the data sampling affects the observed frequency and relative ranking of Twitter entities, e.g., users, hashtags, and URLs. We first use a Bernoulli process to model the Twitter data sampling (Section 5.1). Next, we show how the entity statistics for one set (e.g., the complete) can be estimated using the other set (the sample, Section 5.2). Finally, we measure the distortions introduced in entity ranking by sampling and how to correct them (Section 5.3). The basic statistics of entities are listed in Table 3. The analyses in this section, Section 6, and Section 7, are done with CYBERBULLYING dataset since its sampling effects are more prominent.

5.1 Twitter sampling as a Bernoulli process

We examine how well we can use a Bernoulli process to approximate the Twitter sampling process. Assuming that tweets are sampled identically and independently, the Twitter sampling can be be seen as a simple Bernoulli process with the mean sampling rate $\bar{\rho}$. We empirically validate this assumption by plotting the complementary cumulative density functions (CCDFs) of user posting frequency (the number of times a user posts) and hashtag frequency (the number of times a hashtag appears) in Figure 4. The black and blue solid lines respectively show the CCDFs of the complete and the sample sets, while the black dashed line shows the CCDF in a synthetic dataset constructed from the complete set using a Bernoulli process with rate $\bar{\rho}$=52.72%. Firstly, we observe that the CCDF of the sample set is shifted left, towards the lower frequency end. Visually, the distributions for the synthetic (black dashed line) and for the observed sample set (blue solid line) overlap each other. Furthermore, following the practices in (Leskovec and Faloutsos 2006), we measure the agreement between these distributions with Kolmogorov-Smirnov D-statistic, which is defined as

$$D(G, G') = \max_x\{|G(x) - G'(x)|\} \quad (1)$$

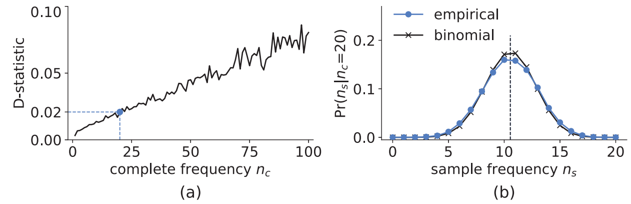

Figure 5: **(a)** D-statistic between empirical distribution and binomial distribution for the number of tweets a user posts. **(b)** The probability distribution of observing n_s times in the sample set when n_c=20.

where G and G' are the cumulative distribution functions (CDFs) of two distributions. With a value between 0 and 1, a smaller D-statistic implies more agreement between two measured distributions. The results show high agreement between entity distributions in the synthetic and the observed sample sets (0.0006 for user posting and 0.002 for hashtag). This suggests that despite the empirical sampling rates not being unique over time, a Bernoulli process of constant rate can model the observed entity frequency distribution well[5].

5.2 Entity frequency

We investigate whether the statistics on one set (complete or sample) can be estimated using only the statistics of the other set and the Bernoulli process model. We use n_c to denote the frequency in the complete set, and n_s the frequency in the sample set ($n_c \geq n_s$). More precisely, we ask these three questions: What is the distribution of n_s given n_c=k? What is the distribution of n_c given n_s=k? How many entities are missing altogether given the distribution of n_s?

Modeling sample frequency from the complete set. For a user who posts n_c times in the complete set, their sample frequency under the Bernoulli process follows a binomial distribution $B(n_c, \bar{\rho})$. Specifically, the probability of observing the user n_s times in the sample set is

$$\Pr(n_s|n_c, \bar{\rho}) = \binom{n_c}{n_s}\bar{\rho}^{n_s}(1-\bar{\rho})^{n_c-n_s} \quad (2)$$

We compute the empirical distribution and binomial distribution for n_c from 1 to 100. This covers more than 99% users in our dataset. Figure 5(a) shows the D-statistic between two distributions as a function of complete frequency n_c. The binomial distribution models the empirical data better when n_c is smaller. Figure 5(b) illustrates an example of n_c=20. The binomial distribution closely approximates the empirical distribution. Their mean sample frequencies (dashed vertical lines) are also identical (10.54).

Inferring complete frequency from the sample set. Under the Bernoulli process, for users who are observed n_s times

[5]We do not choose the goodness of fit test (e.g., Kolmogorov-Smirnov test) because our sample sizes are in the order of millions. And trivial effects can be found to be significant with very large sample sizes. Instead we report the effect sizes (e.g., D-statistic). Alternative distance metrics (e.g., Bhattacharyya distance or Hellinger distance) yield qualitatively similar results.

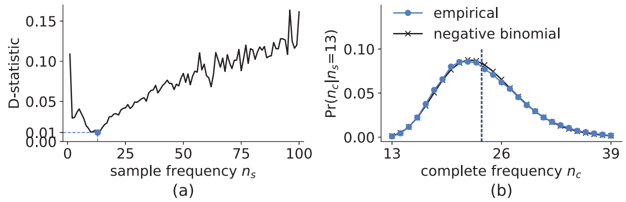

Figure 6: **(a)** D-statistic between empirical distribution and negative binomial distribution for the number of tweets a user posts. **(b)** The probability distribution of posting n_c times in the complete set when $n_s{=}13$.

in the sample set, their complete frequencies follows a negative binomial distribution $NB(n_s, \bar{\rho})$. The negative binomial distribution models the discrete probability distribution of the number of Bernoulli trials before a predefined number of successes occurs. In our context, given n_s tweets ($n_s{\geq}1$) are successfully sampled, the probability of having n_c tweets in the complete set is

$$\Pr(n_c|n_s, \bar{\rho}) = \binom{n_c{-}1}{n_s{-}1} \bar{\rho}^{n_s} (1{-}\bar{\rho})^{n_c-n_s} \tag{3}$$

We compute the empirical distribution and negative binomial distribution for n_s from 1 to 100. Figure 6(a) shows the D-statistic as a function of sample frequency n_s. Negative binomial distributions models the best when the number of observed tweets is between 9 and 15 (D-statistic$<$0.02). Figure 6(b) shows both distributions for $n_s{=}13$, where the minimal D-statistic is reached. The negative binomial distribution closely resembles the empirical distribution. Their estimated mean complete frequencies are very similar (23.60 vs. 23.72, shown as dashed vertical lines).

Estimating missing volume from the sample set. In data collection pipelines, the obtained entities from the filtered stream are sometimes used as seeds for the second step crawling, such as constructing user timelines based on user IDs (Wang, Callan, and Zheng 2015), or querying YouTube statistics based on video URLs (Wu, Rizoiu, and Xie 2018). However, some entities may be completely missing due to Twitter sampling. We thus ask: can we estimate the total number of missing entities given the entity frequency distribution of the sample set?

We formulate the problem as solving a matrix equation with constraints. We use the symbol $\mathbf{F}$ to denote the entity frequency vector. $\mathbf{F}[n_s]$ represents the number of entities that occurs n_s times in the sample set. We want to estimate the frequency vector $\hat{\mathbf{F}}$ of the complete set. For any n_s, its sample frequency $\mathbf{F}[n_s]$ satisfies

$$\mathbf{F}[n_s] = \sum_{k=n_s}^{\infty} \Pr(n_s|k, \bar{\rho}) * \hat{\mathbf{F}}[k] \tag{4}$$

We constrain $\hat{\mathbf{F}}$ to be non-negative numbers and decrease monotonically since the frequency distribution is usually heavy-tailed in practice (see Figure 4). We use the frequency vector for $n_s{\in}[1, 100]$. The above matrix equation can be solved as a constrained optimization task. For users who post n_c times in the complete set, the probability of their tweets completely missing is $\Pr(n_s{=}0; n_c, \bar{\rho}){=}(1{-}\bar{\rho})^{n_c}$. Altogether, the estimated missing volume is $\sum_{n_c=1}^{\infty} (1{-}\bar{\rho})^{n_c}\hat{\mathbf{F}}[n_c]$ for the whole dataset. We show the estimated results in the rightmost column of Table 3. The relative errors (MAPE) are smaller than 0.5% for all entities. This suggests that the volume of missing entities can be accurately estimated if the frequency distribution of the sample set is observed.

Summary. Although the empirical sample rates have clear temporal variations, we show that we can use the mean sampling rate to estimate some entity statistics, including the frequency distribution and the missing volume. This reduces the concerns on assuming the observed data stream is a result of uniform random sampling (Joseph, Landwehr, and Carley 2014; Morstatter, Pfeffer, and Liu 2014; Pfeffer, Mayer, and Morstatter 2018).

5.3 Entity ranking

Entity ranking is important for many social media studies. One of the most common strategies in data filtering is to keep entities that rank within the top x, e.g., most active users or most mentioned hashtags (Morstatter et al. 2013; González-Bailón et al. 2014). We measure how the Twitter data sampling distorts entity ranking for the most active users, and whether the ground-truth ranking in the complete set can be inferred from the sample ranking. Note that in this subsection, we allow the sampling rates to be time-dependent ρ_t and user-dependent ρ_u – as the sampling with a constant rate would preserve the ranking between the complete and the sample sets. For the universal ranking (considering all entities), we use percentile to measure it and find the higher ranked entities have smaller deviations (detailed in Section E of (Appendix 2020)).

Detecting rank distortion. Figure 7(a) plots the most active 100 users in the sample set on the x-axis, and their ranks in the complete set on the y-axis. Each circle is colored based on the corresponding user sampling rate ρ_u. The diagonal line indicates uniform random sampling, in which the two sets of ranks should be preserved. The users above the diagonal line improve their ranks in the sample set, while the ones below lose their positions. Figure 7(c) highlights a user *WeltRadio*, who benefits the most from the sampling: it ranks 50th in the complete set, but it is boosted to 15th place in the sample set. Comparing the complete tweet volume, its volume (4,529) is only 67% relative to the user who actually ranks 15th in the complete set (6,728, user *thirdbrainfx*). We also find that *WeltRadio* tweets mostly in the very high sampling rate secondly period (millisecond 657 to 1,000), resulting in a high user sampling rate ($\rho_u{=}79.1\%$). On the contrary, Figure 7(d) shows a user *bensonbersk* with decreased rank in the sample set and low sampling rate ($\rho_u{=}36.5\%$). Examining his posting pattern, this user mainly tweets in the low sampling rate hours (UTC-12 to 19).

Estimating true ranking from the sample set. Apart from measuring the rank distortion between the complete and the sample sets, we investigate the possibility of estimating the ground-truth ranks by using the observations from the sample set. From the rate limit messages, we extract

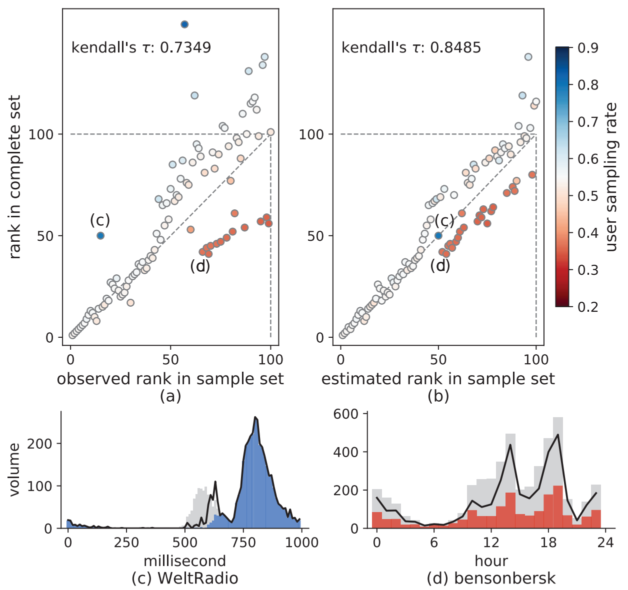

Figure 7: **(a)** Observed ranks in the sample set (x-axis) vs. true ranks in the complete set (y-axis). **(b)** Estimated ranks improve the agreement with the ground-truth ranks. **(c)** user *WeltRadio*, observed/true/estimated ranks: 15/50/50. **(d)** user *bensonbersk*, observed/true/estimated ranks: 66/42/52. blue/red shades: sample tweet volume; grey shades: complete tweet volume; black line: estimated tweet volume.

the temporal sampling rates that are associated with different timescales (hour, minute, second, and millisecond), i.e., $\rho_t(h, m, s, ms)$. Based on the negative binomial distribution, for a user who we observe n_s times at timestamp $\kappa{=}(h, m, s, ms)$, the expected volume is $n_s/\rho_t(\kappa)$. We compute the estimated tweet volumes for all users and select the most active 100 users. Figure 7(b) shows the estimated ranks on the x-axis and the true ranks on the y-axis. We quantify the degree of agreement using Kendall's τ, which computes the difference of concordant and discordant pairs between two ranked lists. With value between 0 and 1, a larger value implies more agreement. The Kendall's τ is improved from 0.7349 to 0.8485 with our estimated ranks. The rank correction is important since it allows researchers to mitigate the rank distortion without constructing a complete data stream.

6 Impacts on networks

In this section, we measure the effects of data sampling on two commonly studied networks on Twitter: the user-hashtag bipartite graph, and the user-user retweet network.

6.1 User-hashtag bipartite graph

The bipartite graph maps the affiliation between two disjoint sets of entities. No two entities within the same set are linked. Bipartite graphs have been used in many social applications, e.g., mining the relation between scholars and published papers (Newman 2001), or between artists and concert venues (Arakelyan et al. 2018). Here we construct the

	complete	sample	ratio
#tweets with hashtags	24,539,003	13,149,980	53.59%
#users with hashtags	6,964,076	4,758,161	68.32%
avg. hashtags per user	9.23	7.29	78.97%
#hashtags	1,166,483	880,096	75.45%
avg. users per hashtags	55.09	39.40	71.51%

Table 4: Statistics of user-hashtag bipartite graph in CYBERBULLYING dataset. Ratio (rightmost column) compares the value of the sample set against that of the complete set, mean sampling rate $\bar{\rho}$=52.72%.

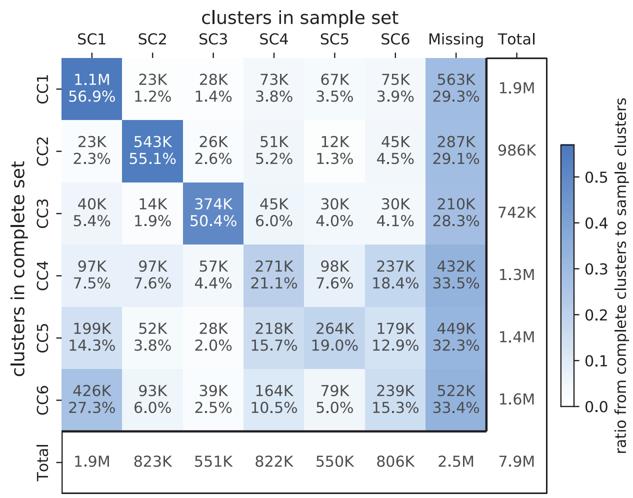

Figure 8: The change of clusters from complete set to sample set. Each cell denotes the volume (top number) and the ratio (bottom percentage) of entities (users and hashtags) that traverse from a complete cluster to a sample cluster. Clusters are ordered to achieve maximal ratios along the diagonal.

user-hashtag bipartite graphs for both the complete and the sample sets. This graph links users to their used hashtags. Each edge has a weight – the number of tweets between its associated user and hashtag. The basic statistics for the bipartite graphs are summarized in Table 4.

Clustering techniques are often used to detect communities in such bipartite graphs. We apply spectral clustering (Stella and Shi 2003) on the user-hashtag bipartite graph, with the number of clusters set at 6. The resulted clusters are summarized in Table 5, together with the most used 5 hashtags and a manually-assigned category. Apart from the cyberbullying keywords, there are significant amount of hashtags related to politics, live streaming, and Korean pop culture, which are considered as some of the most discussed topics on Twitter. We further quantify how the clusters traverse from the complete set to the sample set in Figure 8. Three of the complete clusters (CC1, CC2, and CC3) are maintained in the sample set (mapping to SC1, SC2, and SC3 respectively), since more than half of the entities preserve. The remaining three complete clusters disperse. Investigating the statistics for the complete clusters, the preserved ones have a larger average weighted degree, meaning

		CC1	CC2	CC3	CC4	CC5	CC6
complete set	size	**1,925,520**	**986,262**	**742,263**	1,289,086	1,389,829	1,562,503
	#users	**1,606,450**	**939,288**	**602,845**	1,080,359	1,227,127	1,390,276
	#hashtags	**319,070**	**46,974**	**139,418**	208,727	162,702	172,227
	avg. degree	**8.03**	**7.64**	**22.19**	3.46	4.74	4.07
	category	**politics**	**Korean pop**	**cyberbullying**	Southeast Asia pop	politics	streaming
	hashtags	**brexit**	**bts**	**gay**	peckpalitchoke(th)	kamleshtiwari	ps4live
		demdebate	**mamavote**	**pussy**	peckpalitchoke	standwithhongkong	bigolive
		afd	**blackpink**	**sex**	vixx	hongkong	10tv
		cdnpoli	**pcas**	**horny**	wemadeit	bigil	mixch.tv(ja)
		elxn43	**exo**	**porn**	mayward	lebanon	twitch
		SC1	**SC2**	**SC3**	SC4	SC5	SC6
sample set	size	**1,880,247**	**823,232**	**551,219**	822,436	549,589	805,852
	#users	**1,600,579**	**767,183**	**446,303**	686,609	465,339	688,922
	#hashtags	**279,668**	**56,049**	**104,916**	135,827	84,250	116,930
	avg. degree	**5.58**	**5.75**	**14.98**	3.06	3.51	3.28
	category	**politics**	**Korean pop**	**cyberbullying**	mixed	mixed	mixed
	hashtags	**ps4live**	**bts**	**gay**	mixch.tv(ja)	bigolive	Idolish7(ja)
		10tv	**mamavote**	**pussy**	bigil	kamleshtiwari	reunion
		brexit	**blackpink**	**sex**	peckpalitchoke(th)	bb13	Idolish7(ja)
		afd	**pcas**	**horny**	reality_about_islam(hi)	biggboss13	vixx
		demdebate	**bts(ko)**	**porn**	doki.live(ja)	execution_rajeh_mahmoud(ar)	vixx(ko)

Table 5: Statistics and the most used 5 hashtags in the 6 clusters of the user-hashtag bipartite graph. Three complete clusters maintain their structure in the sample set (**boldfaced**). The language code within brackets is the original language for the hashtag. ja: Japanese; ko: Korean; th: Thai; hi: Hinda; ar: Arabic.

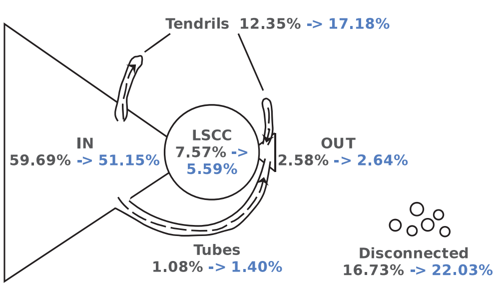

Figure 9: Visualization of bow-tie structure in complete set. The black number indicates the relative size of component in the complete set, blue number indicates the relative size in the sample set.

more tweets between the users and hashtags in these clusters. Another notable observation is that albeit the entities traverse to the sample clusters differently, all complete clusters have similar missing rates (28% to 34%). It suggests that Twitter data sampling impacts the community structure. Denser structures are more resilient to sampling.

6.2 User-user retweet network

Retweet network describes the information sharing between users. We build a user-user retweet network by following the "@RT" relation.. Each node is a user, and each edge is a directed link weighted by the number of retweets between two users. The user-user retweet network has been extensively investigated in literature (Sadikov et al. 2011; Morstatter et al. 2013; González-Bailón et al. 2014).

We choose to characterize the retweet network using the bow-tie structure. Initially proposed to measure the World Wide Web (Broder et al. 2000), the bow-tie structure was also used to measure the QA community (Zhang, Ackerman, and Adamic 2007) or YouTube video networks (Wu, Rizoiu, and Xie 2019). The bow-tie structure characterizes a network into 6 components: (a) the largest strongly connected component (LSCC) as the central part; (b) the IN component contains nodes pointing to LSCC but not reachable from LSCC; (c) the OUT component contains nodes that can be reached by LSCC but not pointing back to LSCC; (d) the Tubes component connects the IN and OUT components; (e) the Tendrils component contains nodes pointing from In component or pointing to OUT component; (f) the Disconnected component includes nodes not in the above 5 components. Figure 9 visualizes the bow-tie structure of the user-user retweet network, alongside with the relative size for each component in the complete and sample sets. The LSCC and IN components, which make up the majority part of the bow-tie, reduce the most in both absolute size and relative ratio due to sampling. OUT and Tubes are relatively small in both complete and sample sets. Tendrils and disconnected components enlarge 39% and 32% after sampling.

Figure 10 shows the node flow of each components from the complete set to the sample set. About a quarter of LSCC component shift to the IN component. For the OUT, Tubes, Tendrils, and Disconnected components, 20% to 31% nodes move into the Tendrils component, resulting in a slight increase of absolute size for Tendrils. Most notably, nodes in the LSCC has a much smaller chance of missing (2.2%, other components are with 19% to 38% missing rates).

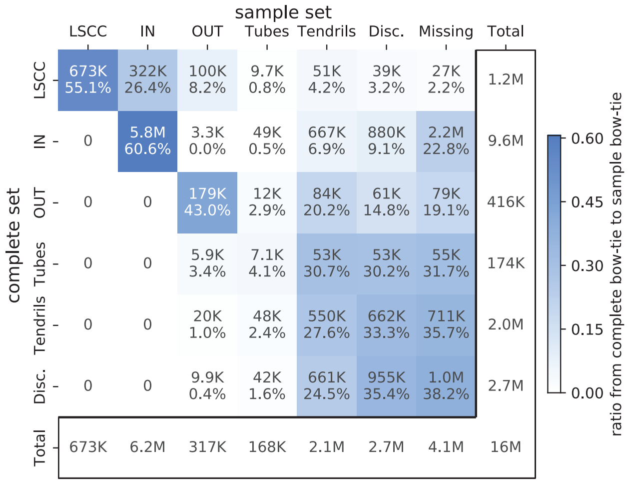

Figure 10: The change of bow-tie components from complete set to sample set. Each cell denotes the volume (top) and the ratio (bottom) of users that traverse from a component in complete set to a component in sample set.

7 Impacts on retweet cascades

Information diffusion is perhaps the most studied social phenomenon on Twitter. A retweet cascade consists of two parts: a root tweet and its subsequent retweets. A number of models have been proposed for modeling and predicting retweet cascades (Zhao et al. 2015; Mishra, Rizoiu, and Xie 2016; Martin et al. 2016). However, these usually make the assumption of observing all the retweets in cascades. In this section, we analyze the impacts of Twitter sampling on retweet cascades and identify risks for existing models. We first construct cascades without missing tweets from the complete set. Next, we measure the sampling effects for some commonly used features in modeling retweet cascades, e.g., inter-arrival time and potential reach.

Constructing complete cascades. When using the filtered streaming API, if a root tweet is observed, the API should return all its retweets. This is because the API also tracks the keywords in the `retweeted_status` field of a tweet (i.e., the root tweet), which allows us to construct a set of complete cascades from the complete set. In the sample set, both the root tweet and any of its retweets could be missing. If the root tweet is missing, we miss the entire cascade. If some retweets are missing, we observe a partial cascade. Table 6 lists the obtained cascades in the complete and the sample sets. Notably, there are 3M cascades in the complete set, but only 1.17M in the sample set (38.85%), out of which only 508k (16.88%) cascades are complete and their sizes are relatively small (i.e., they don't miss any retweet, max cascade size: 23, mean size: 1.37). Prior literature (Zhao et al. 2015) often concentrates on retweet cascades with more than 50 retweets. There are 99,952 such cascades in the complete set, but only 29,577 in the sample set, out of which none is complete.

Inter-arrival time. One line of work models the information diffusion as point processes (Zhao et al. 2015; Mishra, Rizoiu, and Xie 2016). These models use a memory kernel as a function of the time gap Δt between two consecutive events, which is also known as inter-arrival time. Figure 11(a) plots the CCDFs of inter-arrival times in the complete and the sample sets. The distribution shifts right, towards larger values. This is expected as the missing tweets increase the time gap between two observed tweets. The median inter-arrival time is 22.9 seconds in the complete set (black dashed line), meaning 50% retweets happen within 23 seconds from last retweet. After sampling, the median increases almost 5-fold to 105.7 seconds (blue dashed line). For research that uses tweet inter-arrival time, this presents the risk of miss-calibrating models and of underestimating the virality of the cascades.

	complete	sample	ratio
#cascades	3,008,572	1,168,896	38.9%
#cascades ($\geq$50 retweets)	99,952	29,577	29.6%
avg. retweets per cascade	15.6	11.0	70.2%
med. inter-arrival time (s)	22.9	105.7	461.6%

Table 6: Statistics of cascades in CYBERBULLYING dataset.

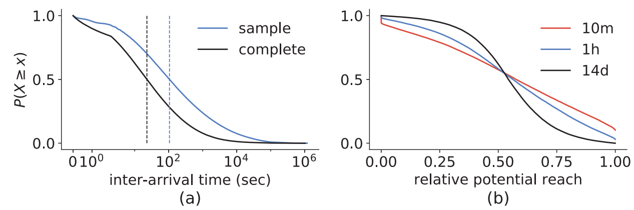

Figure 11: CCDFs of **(a)** inter-arrival time and **(b)** relative potential reach.

Potential reach. Online influence is another well-studied phenomenon on Twitter, and one of its proxies is the number of followers of a user. We define potential reach as the total number of all observed retweeters' followers. This approximates the size of the potential audience for the root tweet. We compute the relative potential reach as the ratio of potential reach in the sample cascade against that in the complete cascade, and we plot the CCDFs in Figure 11(b). When observing cascades for as much as 14 days, 50% of the cascades have the relative potential reach below 0.544. This indicates that when using the sampled Twitter data, researchers can severely underestimate the size of the potential audience. Another common setting is to perform early prediction, i.e., after observing 10 minutes or 1 hour of each retweet cascade. Figure 11(b) shows that the relative potential reach is more evenly distribution for shorter time windows – 21.0% cascades have relative potential reach below 0.25 and 33.7% cascades above 0.75 within 10 minutes span – comparing to the observation over 14 days (5.1% and 11.3%, respectively).

8 Conclusion

This work presents a set of in-depth measurements on the effects of Twitter data sampling. We validate that Twitter rate limit messages closely approximate the volume of missing tweets. Across different timescales (hour, minute, second, millisecond), we find that the sampling rates have distinct temporal variations at each scale. We show the effects of sampling across different subjects (entities, networks, cascades), which may in turn distort the results and interpretations of measurement and modeling studies. For counting statistics such as number of tweets per user and per hashtag, we find that the Bernoulli process with a uniform rate is a reasonable approximation for Twitter data sampling. We also show how to estimate ground-truth statistics in the complete data by using only the sample data.

Limitations. These observations in this paper apply to current Twitter APIs (as of 2020-03) and are subject to the changes of Twitter's proprietary sampling mechanisms. We are aware of that Twitter plans to release a new set of APIs in near future. Consistent with the current streaming APIs, the rate limit threshold for the new APIs is also set to 50 tweets per second (Twitter.com 2020b). Therefore, we believe the observations of this paper will hold.

Practical implications and future work. This work calls attention to the hidden biases in social media data. We have shown effective methods for estimating ground-truth statistics, which allows researchers to mitigate the risks in their datasets without collecting the complete data. Our research also provides methods and toolkits for collecting sampled and complete data streams on Twitter. Our findings provide foundations to many other research topics using sampled data, such as community detection and information diffusion algorithms that are robust to data subsampling. Future works include measuring a larger set of activity and network measurements under data sampling, generalizing the results of this work to other social media platforms and data formats, and quantifying the robustness of existing network and diffusion models against data sampling.

Acknowledgments

This work is supported in part by AOARD Grant 19IOA078, Australian Research Council Project DP180101985, and the National eResearch Collaboration Tools and Resources (Nectar). Marian-Andrei Rizoiu's time is partially supported by a Facebook content policy award.

References

Almuhimedi, H.; Wilson, S.; Liu, B.; Sadeh, N.; and Acquisti, A. 2013. Tweets are forever: A large-scale quantitative analysis of deleted tweets. In *CSCW*.

Appendix. 2020. Online appendix: Variation across scales: Measurement fidelity under Twitter data sampling. https://arxiv.org/pdf/2003.09557.pdf#page=12.

Arakelyan, S.; Morstatter, F.; Martin, M.; Ferrara, E.; and Galstyan, A. 2018. Mining and forecasting career trajectories of music artists. In *Hypertext*.

Bovet, A., and Makse, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Commun.*

Boyd, D., and Crawford, K. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*.

Broder, A.; Kumar, R.; Maghoul, F.; Raghavan, P.; Rajagopalan, S.; Stata, R.; Tomkins, A.; and Wiener, J. 2000. Graph structure in the web. *Computer networks*.

Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *FAT**.

De Choudhury, M.; Lin, Y.-R.; Sundaram, H.; Candan, K. S.; Xie, L.; and Kelliher, A. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media? In *ICWSM*.

De Choudhury, M.; Jhaver, S.; Sugar, B.; and Weber, I. 2016. Social media participation in an activist movement for racial equality. In *ICWSM*.

Gaffney, D., and Matias, J. N. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PloS one*.

González-Bailón, S.; Wang, N.; Rivero, A.; Borge-Holthoefer, J.; and Moreno, Y. 2014. Assessing the bias in samples of large online networks. *Social Networks*.

Joseph, K.; Landwehr, P. M.; and Carley, K. M. 2014. Two 1%s don't make a whole: Comparing simultaneous samples from Twitter's streaming API. In *SBP*.

Kergl, D.; Roedler, R.; and Seeber, S. 2014. On the endogenesis of Twitter's Spritzer and Gardenhose sample streams. In *ASONAM*.

Kossinets, G. 2006. Effects of missing data in social networks. *Social networks*.

Leskovec, J., and Faloutsos, C. 2006. Sampling from large graphs. In *KDD*.

Martin, T.; Hofman, J. M.; Sharma, A.; Anderson, A.; and Watts, D. J. 2016. Exploring limits to prediction in complex social systems. In *WWW*.

Mishra, S.; Rizoiu, M.-A.; and Xie, L. 2016. Feature driven and point process approaches for popularity prediction. In *CIKM*.

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. M. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *ICWSM*.

Morstatter, F.; Pfeffer, J.; and Liu, H. 2014. When is it biased?: Assessing the representativeness of Twitter's streaming API. In *WWW*.

Mosteller, F. 1949. The pre-election polls of 1948. *Social Science Research Council*.

Nand, P.; Perera, R.; and Kasture, A. 2016. How bullying is this message?: A psychometric thermometer for bullying. In *COLING*.

Newman, M. E. 2001. The structure of scientific collaboration networks. *PNAS*.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kiciman, E. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*.

Pfeffer, J.; Mayer, K.; and Morstatter, F. 2018. Tampering with Twitter's sample API. *EPJ Data Science*.

Rizoiu, M.-A.; Xie, L.; Sanner, S.; Cebrian, M.; Yu, H.; and Van Hentenryck, P. 2017. Expecting to be HIP: Hawkes intensity processes for social media popularity. In *WWW*.

Ruths, D., and Pfeffer, J. 2014. Social media for large studies of behavior. *Science*.

Sadikov, E.; Medina, M.; Leskovec, J.; and Garcia-Molina, H. 2011. Correcting for missing data in information cascades. In *WSDM*.

Sampson, J.; Morstatter, F.; Maciejewski, R.; and Liu, H. 2015. Surpassing the limit: Keyword clustering to improve Twitter sample coverage. In *Hypertext*.

Stella, X. Y., and Shi, J. 2003. Multiclass spectral clustering. In *ICCV*.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM*.

Twitter.com. 2020a. Filter realtime tweets. https://developer.twitter.com/en/docs/tweets/filter-realtime/overview/statuses-filter. [Online; accessed Mar-31-2020].

Twitter.com. 2020b. Rate limit in labs streaming endpoint. https://developer.twitter.com/en/docs/labs/filtered-stream/troubleshooting. [Online; accessed Mar-31-2020].

Twitter.com. 2020c. Sample realtime tweets. https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/GET_statuse_sample. [Online; accessed Mar-31-2020].

Twitter.com. 2020d. Search tweets. https://developer.twitter.com/en/docs/tweets/search/overview. [Online; accessed Mar-31-2020].

Twitter.com. 2020e. Twitter rate limit notices. https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/streaming-message-types. [Online; accessed Mar-31-2020].

Wagner, C.; Singer, P.; Karimi, F.; Pfeffer, J.; and Strohmaier, M. 2017. Sampling from social networks with attributes. In *WWW*.

Wang, Y.; Callan, J.; and Zheng, B. 2015. Should we use the sample? Analyzing datasets sampled from Twitter's stream API. *TWEB*.

Wu, S.; Rizoiu, M.-A.; and Xie, L. 2018. Beyond views: Measuring and predicting engagement in online videos. In *ICWSM*.

Wu, S.; Rizoiu, M.-A.; and Xie, L. 2019. Estimating attention flow in online video networks. In *CSCW*.

Zhang, J.; Ackerman, M. S.; and Adamic, L. 2007. Expertise networks in online communities: structure and algorithms. In *WWW*.

Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman, A.; and Leskovec, J. 2015. SEISMIC: A self-exciting point process model for predicting tweet popularity. In *KDD*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*.