Variation across Scales: Measurement Fidelity under Twitter Data Sampling

Siqi Wu[1,3] and **Marian-Andrei Rizoiu**[2,3] and **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 bias is the cornerstone of mitigating biases in social media research. This paper presents in-depth measurements of the effects of Twitter data sampling across different timescales and different subjects (entities, networks, and cascades). By constructing two complete tweet streams, we show that Twitter rate limit message is an accurate measure for the volume of missing tweets. Despite sampling rates having clear temporal variations, we find that the Bernoulli process with a uniform rate well approximates Twitter data sampling, and it allows to estimate the ground-truth entity frequency and ranking with the observed sample data. In terms of network analysis, we observe significant structure changes in both the user-hashtag bipartite graph and the retweet network. Finally, we measure the retweet cascades. We identify risks for information diffusion models that rely on tweet inter-arrival times and user influence. This work calls attention to the social data bias caused by data collection, and proposes methods to measure the systematic biases introduced by sampling.

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]

Fairness is a timely topic that receives broad attention. One of the main sources of unfairness is data bias (Tufekci 2014; Olteanu et al. 2019), which particularly affects data-driven studies. 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 bias 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[2], 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. Although Twitter allows the access to significant amounts of data, in high-volume cases that exceed its rate limit, Twitter only returns a portion of desired tweets. Therefore, once triggered, Twitter sampling will compromise the completeness of collected observational data. Twitter offers two streaming APIs for free, namely *sampled* stream and *filtered* stream. In this paper, we focus on empirically quantifying the data bias resulted from the sampling in the filtered streaming API.

This work addresses three open questions related to Twitter sampling. Firstly, **are tweets missing at random?** 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 streaming API 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, and retweet cascades. Lastly, **can we correct the sampling effects?** Within the scope of Twitter sampling, inferring ground-truth statistics from observed samples remains unexplored. The answers to these questions will help researchers shape appropriate questions, and 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 into multiple subcrawlers. We study the relevance of the Twitter rate

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

[2]Due to space limitation, we list the papers and their used Twitter APIs in Section 1 of (Appendix 2020).

limit messages. Contradicting observations made by Sampson et al. (2015), our results show that the returned messages closely approximate the volume of missing data.

Addressing the second question, we measure the effects of Twitter sampling over different subjects, e.g., the entity frequency, entity ranking, user-hashtag usage graph, retweet network, and retweet cascades. We find that (1) uniform random sampling is a reasonable approximation for the empirical entity distribution; (2) the ranks of top entities are disrupted; (3) the network structures change significantly with some components more likely to be preserved; (4) sampling introduces risks to information diffusion models as the distributions of inter-arrival times are substantially skewed. We remark that this work only studies the effects of Twitter sampling mechanism, but does not intend to reverse engineer it.

Concerning the third question, we find that the Bernoulli process can be used to infer the ground-truth statistics (entity frequency and ranking) based on sampled observations.

The main contributions of this work include:

- We show that the Twitter rate limit message is an accurate measure for the volume of missing tweets.
- A set of measurements on the Twitter sampling effects across different timescales and different subjects.
- We show how to uncover the entity frequency and ranking of the complete data using only the sample data.
- We release a software package to construct the complete data streams on Twitter[3].

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 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 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 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. More importantly, we take a step to correct the sampling effects.

Another important observation is that 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. The fairness problem has received growing attention in academic studies. 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 bias 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 uncovered.

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.

[3]The software package, collected data, and analysis code will be made publicly available soon.

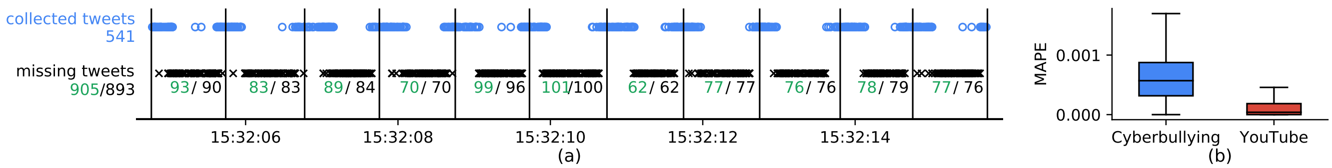

Figure 1: **(a)** Collected and missing tweets in an 11-second interval. blue circle: collected tweet; black cross: missing tweet; black vertical line: rate limit message. green number: estimated missing volume from rate limit messages; black number: count of missing tweets compared to the complete set. **(b)** MAPE of predicting 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 the Twitter API. The client will be rate limited if the number of matching tweets exceeds a preset threshold – 50 tweets per second as of 2020-01 (Twitter 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*.

To construct the complete data streams, we split the keyword set into multiple disjoint sets, and track 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 then 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). Hence, for the rest of this work, we treat the complete sets as if they contain no missing tweets.

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 2020e). Therefore, the difference between 2 consecutive rate limit messages should estimate the missing volume in between.

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 2 of (Appendix 2020).

	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%

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 measure for the number of missing tweets. Note that it only approximates the volume, but not the content of missing tweets.

Our observations contradict those from Sampson et 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

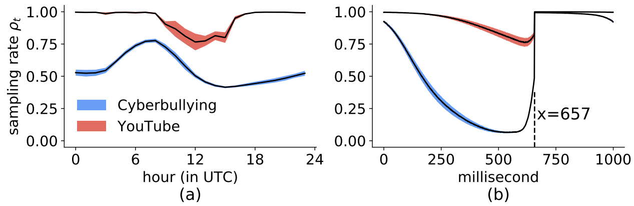

Figure 2: Sampling rates are uneven **(a)** in different hours or **(b)** in different milliseconds. black line: temporal mean sampling rates; color shades: 95% confidence interval.

0.13%, comparing N_c of the complete set to $N_c+\hat{N}_m$ of the sample set in Table 1). This discrepancy may result from a different design of rate limit message back in 2015. We provide a detailed analysis in Section 3 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 in relation to 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. We include the minutely and secondly sampling rates in Section 4 of (Appendix 2020).

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).

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

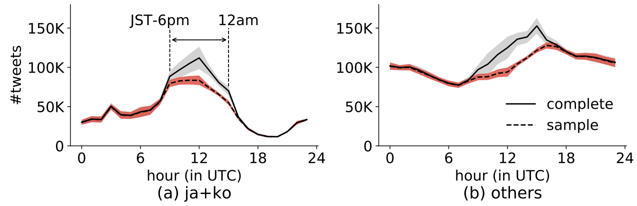

Figure 3: Hourly tweet volumes in YOUTUBE dataset. **(a)** Japanese+Korean; **(b)** other languages. black line: temporal mean tweet volumes; color shades: 95% confidence interval.

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

	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%

Tweet types. Currently there are 4 types of tweets. The users create a *root tweet* when they post new content from their home timeline. 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 the different types of tweets are different 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 sampling is not biased towards any tweet type.

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., user who posts the tweet, hashtags, or URLs. We first use a Bernoulli process to model the Twitter 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 rankings by sampling and how to correct them (Section 5.3). The statistics of entities are listed in Table 3. The analyses in this section, Section 6, and Section 7,

Table 3: Statistics of entities in CYBERBULLYING dataset.

	complete	sample	#missing	est. #missing
#users	19,802,506	14,649,558	5,152,948	5,171,746
#hashtags	1,166,483	880,096	286,387	283,558
#URLs	467,941	283,729	184,212	182,458

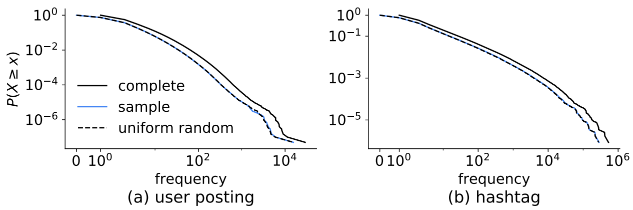

Figure 4: The frequency distributions of **(a)** user posting and **(b)** hashtag. The x-axis starts at 0 rather than 1, as the sample set and uniform random sample both have missing entities.

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%. First, 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, we measure the agreement between these distributions following the practices in (Leskovec and Faloutsos 2006), by using the Kolmogorov-Smirnov D-statistic, which is defined as

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

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, a Bernoulli process of constant rate closely models the observed entity frequency distributions[5].

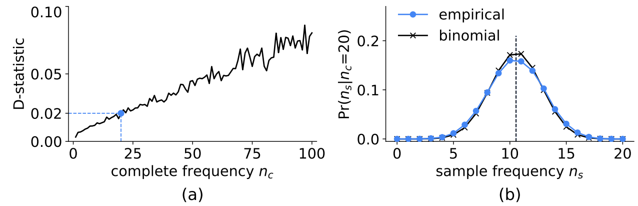

Figure 5: **(a)** D-statistic between empirical distribution and binomial distribution. **(b)** The probability distribution of observing n_s times in the sample set when n_c=20.

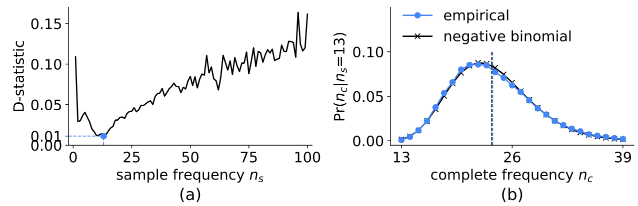

Figure 6: **(a)** D-statistic between empirical distribution and negative binomial distribution. **(b)** The probability distribution of posting n_c times in the complete set when n_s=13.

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 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 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

[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.

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 **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

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

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 $\sum_{n_c=1}^{\inf} (1-\bar{\rho})^{n_c} \hat{\mathbf{F}}[n_c]$ for the whole dataset. We show the estimated results in the rightmost column of Table 3. The relative errors (MAPE) are smaller than 1.0% 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).

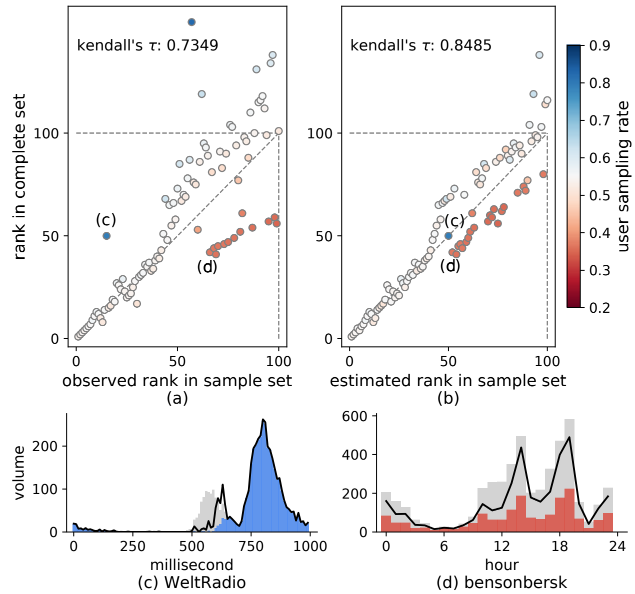

Figure 7: **(a)** Observed ranks in the sample set (x-axis) vs. ground-truth ranks in the complete set (y-axis). **(b)** Estimated ranks improve the agreement with the ground-truth ranks. **(c)** user WeltRadio, observed/ground-truth/estimated ranks: 15/50/50. **(d)** user bensonbersk, observed/ground-truth/estimated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

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 (González-Bailón et al. 2014). We measure how the Twitter 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.

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 the 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 (ρ_u=79.1%). On the contrary, Figure 7(d) shows a user bensonbersk with decreased rank in the sample set and low sampling rate (ρ_u=36.5%). Examining his posting pattern, this user mainly tweets in the low sampling rate hours (UTC-12 to 19).

Estimating complete ranks from the sample set. Apart from measuring the rank distortion between the complete and the sample sets, we investigate the possibility of estimating the complete ranks by using the observations from the sample set. From the rate limit messages, we extract 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 κ=(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 ground-truth 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 as it allows researchers to mitigate the rank distortion without constructing a complete data stream.

6 Impacts on networks of entities

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 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 from such bipartite graphs. We apply spectral clustering 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 (Dodds et al. 2019). We further quantify how the clusters transform from the complete set to the sample set in Figure 8. Three of the complete clusters (CC1, CC5, and CC6) are maintained in the sample set (mapping to SC1, SC2, and SC5 respectively). Investigating the statistics for the complete clusters, the preserved ones have a larger average weighted degree

Table 4: Statistics of user-hashtag bipartite graphs in CYBERBULLYING dataset. Ratio (rightmost column) compares the value of the sample set against that of the complete set.

	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%

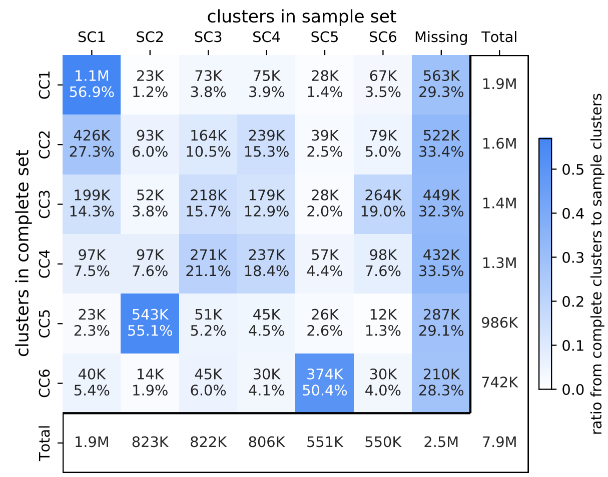

Figure 8: The change of clusters from complete set to sample set. Each cell denotes the volume (top number) and the ratio (bottom percentage) of entities (users and hashtags) transforms from a complete set to a sample cluster.

(see Table 5), meaning more tweets between the users and hashtags in these clusters. Another notable observation is that albeit the entities transfer to the sample clusters differently, all complete clusters have similar missing rates (28% to 34%). It suggests that the Twitter sampling impacts the structure of the communities and of the discussions. Denser structures are also 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 reach-

Table 5: Statistics and the most used 5 hashtags in the 6 clusters of the user-hashtag bipartite graph. Clusters are ordered by size (#users+#hashtags). 3 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.

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

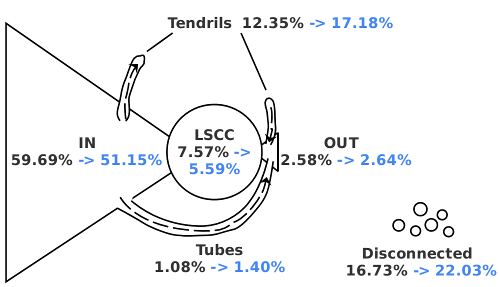

Figure 9: Visualization of bow-tie structure in complete set. The black number indicates the relative size of component in the complete set, blue number indicates the relative size in the sample set.

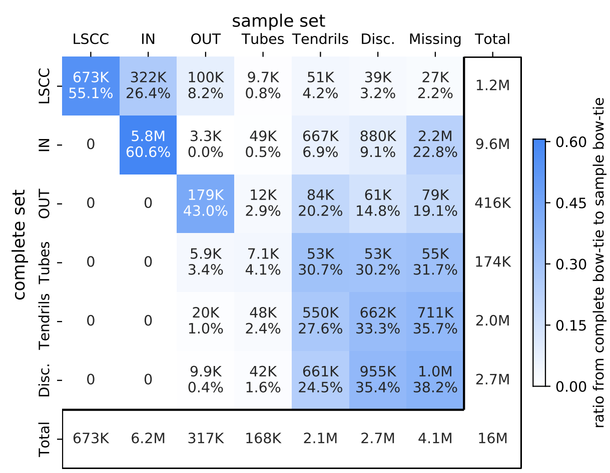

Figure 10: The change of bow-tie components from complete set to sample set. Each cell denotes the volume (top) and the ratio (bottom) of users transforms from a component in complete set to a component in sample set.

able 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 31% and 39% after sampling.

Figure 10 shows the node flow of each components from the complete 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 an 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).

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 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%) small cascades are complete (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 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 in the complete set (black dashed line), meaning 50% retweets happen within 23 seconds from last retweet. After sampling, the median increases 5-fold to 105.7 seconds (blue dashed line). For research that uses the sample inter-arrival time, this poses the risk of miss-calibrating models and of underestimating the virality of the cascades.

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 – comparing to the observation over 14 days (5.1% and 11.3%, respectively).

Table 6: Statistics of cascades in CYBERBULLYING dataset.

	complete	sample	ratio
#cascades	3,008,572	1,168,896	38.85%
avg. retweets per cascade	15.63	10.97	70.18%
#cascades (≥50 retweets)	99,952	29,577	29.59%

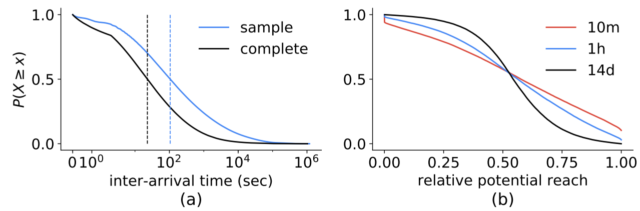

Figure 11: CCDFs of **(a)** inter-arrival time and **(b)** relative potential reach.

8 Conclusion

This work presents a set of in-depth measurements on the effects of Twitter data sampling. We validate that the 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. Across different subjects (entities, networks, cascades), we measure a set of features for each subject, and identify risks for measurement and modeling studies. For some collective statistics, we show that the Bernoulli process is a reasonable approximation for the Twitter sampling. We also show how to uncover ground-truth statistics in the complete data using only the sample data.

Limitations. These observations in this paper apply to current Twitter APIs (as of 2020-01) 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 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 uncovering ground-truth statistics, which allows researchers to mitigate the risks in their datasets without collecting the complete data. Our research also provides methods and tool kits for collecting sampled and complete data on Twitter. This can fuel data sources to many other research topics, e.g., building sampling-robust models for community detection and information diffusion. Future works include quantifying the robustness of existing models against Twitter data sampling, e.g., SEISMIC (Zhao et al. 2015).

Acknowledgments. This work is supported by Asian Office of Aerospace Research and Development (AOARD) Grant 19IOA078, Australian Research Council Discovery 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. http://tiny.cc/w1priz. [Anonymous link during submission].

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*.

Dodds, P. S.; Minot, J. R.; Arnold, M. V.; Alshaabi, T.; Adams, J. L.; Dewhurst, D. R.; Reagan, A. J.; and Danforth, C. M. 2019. Fame and ultrafame: Measuring and comparing daily levels of "being talked about" for United States' presidents, their rivals, God, countries, and K-pop. *arXiv preprint arXiv:1910.00149*.

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 dont make a whole: Comparing simultaneous samples from Twitters 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 Twitters 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*.

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

Twitter. 2020a. Filter realtime tweets. https://developer.twitter.com/en/docs/tweets/filter-realtime/overview/statuses-filter. [Online; accessed Jan-15-2020].

Twitter. 2020b. Rate limit in labs streaming endpoint. https://developer.twitter.com/en/docs/labs/filtered-stream/troubleshooting. [Online; accessed Jan-15-2020].

Twitter. 2020c. Sample realtime tweets. https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/GET_statuse_sample. [Online; accessed Jan-15-2020].

Twitter. 2020d. Search tweets. https://developer.twitter.com/en/docs/tweets/search/overview. [Online; accessed Jan-15-2020].

Twitter. 2020e. Twitter rate limit notices. https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/streaming-message-types. [Online; accessed Jan-15-2020].

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 Twitters 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*.

Supplementary Information
Variation across Scales: Measurement Fidelity under Twitter Data Sampling

Siqi Wu, Marian-Andrei Rizoiu and Lexing Xie

1 Twitter data in ICWSM papers (2015-2019)

82 (31%) out of 265 ICWSM full papers used Twitter data from 2015 to 2019. Twitter search API has been used 25 times, sampled stream 12 times, filtered stream 18 times, firehose 8 times. 12 papers used multiple Twitter APIs for data collection. 7 papers did not clearly specify their Twitter API choices.

Table 1: Year 2015. 20 out of 64 papers used Twitter data. search API: 4; sampled stream: 5; filtered stream: 3; firehose: 1; unspecified: 4; multiple APIs: 3.

Id	Paper	APIs	Notes
1	Audience analysis for competing memes in social media	search	searched keywords “Russia”, “meteor”, “Fox”, and “Obama”
2	Making use of derived personality: The case of social media ad targeting	filtered	mention at least one term related to NYC and one term related to traveling
3	The many shades of anonymity: Characterizing anonymous social media content	unspecified, possibly sampled	500 random publicly available tweets
4	On analyzing hashtags in Twitter	sampled; search	10M messages crawled in December 2013; 200 tweets for each hashtag in our original dataset
5	WhichStreams: A dynamic approach for focused data capture from large social media	sampled; filtered	5000 users first to use one of the keywords “Obama”, “Romney” or “#USElections”
6	Characterizing silent users in social media communities	filtered	all tweets of 140,851 Singapore-based users and 126,047 Indonesia-based users
7	Predicting user engagement on Twitter with real-world events	firehose	nearly 2.7 billion English tweets during August of 2014
8	Geolocation prediction in Twitter using social networks: A critical analysis and review of current practice	sampled	10% sampled stream
9	Characterizing information diets of social media users	sampled	500 randomly selected tweets from Twitter’s 1% random sample
10	Degeneracy-based real-time sub-event detection in Twitter stream	unspecified, possibly filtered	several football matches that took place during the 2014 FIFA World Cup in Brazil, between the 12th of June and the 13th of July 2014
11	CREDBANK: A large-scale social media corpus with associated credibility annotations	sampled	1% random sample
12	Understanding musical diversity via online social media	search	collected U.S. Twitter users who share their Last.fm accounts, then we collected all publicly available tweets
13	Smelly maps: The digital life of urban smellscapes	unspecified	collected 5.3M tweets during year 2010 and from October 2013 to February 2014
14	Project recommendation using heterogeneous traits in crowdfunding	search	retrieving tweets containing URLs that begin with `http://kck.st`
15	Don’t let me be #misunderstood: Linguistically motivated algorithm for predicting the popularity of textual memes	sampled	approximately 15% of the Twitter stream in six month period
16	SEEFT: Planned social event discovery and attribute extraction by fusing Twitter and web content	unspecified, possibly search	querying Twitter API with 3 event types, namely concerts, conferences, and festivals
17	A bayesian graphical model to discover latent events from Twitter	sampled	1% sampled stream and 10% sampled stream
18	Patterns in interactive tagging networks	sample; search	randomly sampled 1 million seed users from sample streams on December 2014; following network starting from the same 1 million seed users
19	Hierarchical estimation framework of multi-label classifying: A case of tweets classifying into real life aspects	search	collected 2,390,553 tweets posted from April 15, 2012 to August 14, 2012, each of which has “Kyoto” as the Japanese location information
20	The lifecyle of a Youtube video: Phases, content and popularity	filtered	tweets containing keyword “youtube” OR (“youtu” AND “be”)

Table 2: Year 2016. 23 out of 52 papers used Twitter data. search API: 10; sampled stream: 2; filtered stream: 5; firehose: 3; unspecified: 1; multiple APIs: 2.

Id	Paper	APIs	Notes
1	Are you charlie or ahmed? Cultural pluralism in charlie hebdo response on Twitter	search	#JeSuisCharlie, #JeSuisAhmed, and #CharlieHebdo – from 2015-01-07 to 2015-01-28
2	When a movement becomes a party: Computational assessment of new forms of political organization in social media	filtered	extracted 373,818 retweets of tweets that (1) were created by, (2) were retweeted by, or (3) mentioned a user from the list
3	Journalists and Twitter: A multidimensional quantitative description of usage patterns	search	contained 5,358 accounts of journalists and news organizations, crawled all their 13,140,449 public tweets
4	Social media participation in an activist movement for racial equality	filtered	#ferguson, #BlueLivesMatter, #BlackLivesMatter, #AllLivesMatter, #Baltimore, #BaltimoreRiots, #BaltimoreUprising, and #FreddieGray
5	Understanding communities via hashtag engagement: A clustering based approach	firehose	tweets from all English language Twitter users in the United States that used a hashtag at least once during the 30 day study period starting January 15, 2015
6	Investigating the observability of complex contagion in empirical social networks	filtered; search	45 Nigerian cities with populations of 100,000 or more using a radius varying from 25 to 40 miles; collected tweets from the timelines of selected users
7	Dynamic data capture from social media streams: A contextual bandit approach	sampled; filtered	leverage sampled stream to discover unknown users; filtered stream for realtime data of the subset users
8	On unravelling opinions of issue specific-silent users in social media	search	asked the Twitter users to provide their Twitter screen names so as to crawl their Twitter data
9	Distinguishing between topical and non-topical information diffusion mechanisms in social media	search	a dataset that is nearly complete and contains all public tweets produced by users until September 2009 and a snapshot of the social graph crawled in September 2009
10	TweetGrep: Weakly supervised joint retrieval and sentiment analysis of topical tweets	search	the queries are issued to the Twitter Search Web Interface via a proxy that we developed
11	What the language you tweet says about your occupation	search	we download these users' 3,000 most recent tweets
12	TiDeH: Time-dependent hawkes process for predicting retweet dynamics	firehose	SEISMIC dataset by Zhao et al. 2015
13	Emotions, demographics and sociability in Twitter interactions	search	collect tweets from an area that included Los Angeles, then collect all (timeline) tweets from subset users
14	Analyzing personality through social media profile picture choice	search	we have collected up to 3,200 most recent tweets for each user
15	Cross social media recommendation	unspecified, possibly sampled	corpora were sampled between 2012/09/17 and 2012/09/23
16	Understanding anti-vaccination attitudes in social media	firehose	first did a manual examination of 1000 Twitter posts, then snowball sampling from a sample of firehose
17	Twitter's glass ceiling: The effect of perceived gender on online visibility	sampled	10% sampled stream
18	Mining pro-ISIS radicalisation signals from social media users	search	Twitter user timeline of 154K users
19	Predictability of popularity: Gaps between prediction and understanding	sampled	URLs tweeted by 737k users for three weeks of 2010
20	Theme-relevant truth discovery on Twitter: An estimation theoretic approach	search	collected through Twitter search API using query terms and specified geographic regions related to the events
21	#PrayForDad: Learning the semantics behind why social media users disclose health information	filtered	collect tweets in English and published in the contiguous United States during a four-month window in 2014
22	Your age is no secret: Inferring microbloggers' ages via content and interaction analysis	filtered	record all the tweets which contain one of the keywords "happy *y*th birthday" with y ranging from 14 to 70
23	EigenTransitions with hypothesis testing: The anatomy of urban mobility	filtered	collected geo-tagged Tweets generated within the area covering New York City and Pittsburgh from Jul 15, 2013 to Nov 09, 2014.

Table 3: Year 2017. 9 out of 50 papers used Twitter data. search API: 3; filtered stream: 3; multiple APIs: 3.

Id	Paper	APIs	Notes
1	Who makes trends? Understanding demographic biases in crowdsourced recommendations	sampled; search	1% random sample; queried search API every 5 minutes and collected all topics which became trending in US
2	#NotOkay: Understanding gender-based violence in social media	sampled; filtered	1% random sample; collect tweets from October 26th to November 26th, 2016 that contain the indicated hashtags
3	Online popularity under promotion: Viral potential, forecasting, and the economics of time	filtered	tweets containing keyword "youtube" OR ("youtu" AND "be")
4	Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on Twitter	filtered	tracked "shooter, shooting, gunman, gunmen, gunshot, gunshots, shooters, gun shot, gun shots, shootings" between January 1 and October 5, 2016
5	State of the geotags: Motivations and recent changes	filtered	selected all coordinate-geotagged tweets within 0.2 degrees latitude and longitude from Pittsburgh
6	Online human-bot interactions: Detection, estimation, and characterization	search	collected the most recent tweets produced by those accounts
7	Identifying effective signals to predict deleted and suspended accounts on Twitter across languages	sampled; search	1% random sample; batches of 100 unique users were queried against the public Twitter API
8	Adaptive spammer detection with sparse group modeling	search	crawled a Twitter dataset from July 2012 to September 2012 via the Twitter Search API
9	Wearing many (social) hats: How different are your different social network personae?	search	76% of About.me users in our dataset have linked their profiles to their alternate account in Twitter

Table 4: Year 2018. 13 out of 48 papers used Twitter data. search API: 2; sampled stream: 3; filtered stream: 4; firehose: 2; multiple APIs: 2.

Id	Paper	APIs	Notes
1	Peer to peer hate: Hate speech instigators and their targets	sampled; search	1% random sample; we use search API to fetch tweet traces of users
2	Characterizing audience engagement and assessing its impact on social media disclosures of mental illnesses	search	obtain the list of individuals who have retweeted each tweet from the disclosers during this period of analysis
3	Facebook versus Twitter: Cross-platform differences in self-disclosure and trait prediction	search	we collected participants' social media posts
4	Can you verifi this? Studying uncertainty and decision-making about misinformation using visual analytics	filtered	collected 103,248 tweets posted by these 178 accounts along with account metadata from May 23, 2017 to June 6, 2017
5	Using longitudinal social media analysis to understand the effects of early college alcohol use	firehose	extract 639k tweets that match these keywords in August-December 2010 in our organization's archive of the Twitter firehose
6	Modeling popularity in asynchronous social media streams with recurrent neural networks	filtered; firehose	tweets containing keyword "youtube" OR ("youtu" AND "be"); SEISMIC dataset by Zhao et al. 2015
7	The effect of extremist violence on hateful speech online	sampled	10% random sample
8	You are your metadata: Identification and obfuscation of social media users using metadata information	sampled	random sample of the tweets posted between October 2015 and January 2016
9	#DEBATENIGHT: The role and influence of socialbots on Twitter during the first 2016 U.S. presidential debate	firehose	Twitter discussions that occurred during the 1st 2016 U.S presidential debate between Hillary Clinton and Donald Trump
10	Ecosystem or echo-system? Exploring content sharing across alternative media domains	filtered	tracked various keyword terms related to the Syrian conflict including geographic terms of affected areas
11	COUPLENET: Paying attention to couples with coupled attention for relationship recommendation	filtered	collected tweets with emojis contains the keyword "heart" in its description
12	Beyond views: Measuring and predicting engagement in online videos	filtered	tweets containing keyword "youtube" OR ("youtu" AND "be")
13	Understanding web archiving services and their (mis)use on social media	sampled	1% random sample

Table 5: Year 2019. 17 out of 51 papers used Twitter data. search API: 6; sampled stream: 2; filtered stream: 3; firehose: 2; unspecified: 2; multiple APIs: 2.

Id	Paper	APIs	Notes
1	Linguistic cues to deception: Identifying political trolls on social media	firehose	a list of 2,752 Russian troll accounts, then collected all of the trolls' discussions
2	Tweeting MPs: Digital engagement between citizens and members of parliament in the UK	search	we fetched all the users (?4.28 Million) who follow MPs and also the users that MPs followed (869K)
3	View, like, comment, post: Analyzing user engagement by topic at 4 levels across 5 social media platforms for 53 news organizations	filtered	collecting all posts from a news organization
4	A large-scale study of ISIS social media strategy: Community size, collective influence, and behavioral impact	firehose	a large dataset of 9.3 billion tweets representing all tweets generated in the Arabic language in 2015 through full private access to the Twitter firehose
5	Who should be the captain this week? Leveraging inferred diversity-enhanced crowd wisdom for a fantasy premier league captain prediction	unspecified	collected their soccer related tweets by scraping Twitter user timelines (for a total 4,299,738 tweets)
6	Multimodal social media analysis for gang violence prevention	search	we scraped all obtainable tweets from this list of 200 users in February 2017
7	Hot streaks on social media	search	we obtained all tweets, followers, and retweeters of all tweets using the Twitter REST API
8	Understanding and measuring psychological stress using social media	search	601 active users who completed the survey
9	Studying cultural differences in emoji usage across the east and the west	sampled	10% random sample
10	What Twitter profile and postedImages reveal about depression and anxiety	search	downloaded the 3200 most recent user tweets for each user, leading to a data set of 5,547,510 tweets, out of which 700,630 posts contained images and 1 profile image each across 3498 users
11	Polarized, together: Comparing partisan support for Trump's tweets using survey and platform-based measures	sampled; search	collecting a large sample of Twitter users (approximately 406M) who sent one or more tweets that appeared in the Twitter Decahose from January 2014 to August 2016; select from this set the approximately 322M accounts that were still active in March 2017
12	Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 U.S. cities	sampled	1% sample of Twitter's public stream from January 1st, 2011 to December 31st, 2016
13	A social media study on the effects of psychiatric medication use	sampled; filtered	public English posts mentioning these drugs between January 01, 2015 and December 31, 2016
14	SENPAI: Supporting exploratory text analysis through semantic&syntactic pattern inspection	filtered	gathered a dataset of Twitter messages from 103 professional journalists and bloggers who work in the field of American Politics
15	Empirical analysis of the relation between community structure and cascading retweet diffusion	search	we used the Search API and collected Japanese tweets using the query q=RT, lang=ja
16	Measuring the importance of user-generated content to search engines	unspecified	a row of three cards with one tweet each. Google obtains the tweets either from Twitter's search (a SearchTweetCarousel) or a single user (a UserTweetCarousel)
17	Detecting journalism in the age of social media:Three experiments in classifying journalists on Twitter	filtered	tracking a set of event-related keywords and hashtags

2 Constructing the complete data streams

To collect the complete streams, we split the same set of keywords into multiple streaming clients. The CYBERBULLYING and YOUTUBE datasets are respectively crawled by 8 and 12 clients based on different combinations of keywords and languages. We carefully choose the split criteria to ensure that each client would not be rate limited by much ($\bar{\rho}$>99% for all subcrawlers). Table 6 and Table 7 list the statistics for all streaming clients of the two datasets.

- 25 keywords for the CYBERBULLYING dataset: `nerd`, `gay`, `loser`, `freak`, `emo`, `whale`, `pig`, `fat`, `wannabe`, `poser`, `whore`, `should`, `die`, `slept`, `caught`, `suck`, `slut`, `live`, `afraid`, `fight`, `pussy`, `cunt`, `kill`, `dick`, `bitch`
- rules for the YOUTUBE dataset: "`youtube`" `OR` ("`youtu`" `AND` "`be`")
- 66 language codes on Twitter, denoted as set S: en, es, ja, ko, und, ar, pt, de, tl, fr, cs, it, vi, in, tr, pl, ru, sr, th, el, nl, hi, zh, da, ro, is, no, hu, fi, lv, et, bg, ht, uk, lt, cy, ka, ur, sv, ta, sl, iw, ne, fa, am, te, km, ckb, hy, eu, bn, si, my, pa, ml, gu, kn, ps, mr, sd, lo, or, bo, ug, dv, ca

Table 6: Subcrawler configurations for CYBERBULLYING dataset. $S\backslash$en is all languages excluding "en".

Id	Keywords	Languages	#collected tweets	#rate limit	#est. missing tweets	sampling rate
1	`should`	en	29,647,814	1,357	7,324	99.98%
2	`should`	$S\backslash$en	801,904	0	0	100.00%
3	`live`	en	16,526,226	1,273	25,976	99.84%
4	`live`	$S\backslash$en	7,926,325	233	7,306	99.91%
5	`kill`, `fight`, `poser`, `nerd`, `freak`, `pig`	all	15,449,973	16	108	100.00%
6	`dick`, `suck`, `gay`, `loser`, `whore`, `cunt`	all	13,164,053	15	125	100.00%
7	`pussy`, `fat`, `die`, `afraid`, `emo`, `slut`	all	21,333,866	89	1,118	99.99%
8	`bitch`, `wannabe`, `whale`, `slept`, `caught`	all	14,178,366	64	666	100.00%
complete	subcrawlers 1-8	all	114,488,537	3,047	42,623	99.96%
sample	all 25 keywords	all	60,400,257	1,201,315	54,175,503	52.72%

Table 7: Subcrawler configurations for YOUTUBE dataset.

Id	Keywords	Languages	#collected tweets	#rate limit	#est. missing tweets	sampling rate
1	"`youtube`"	en	10,312,498	323	3,582	99.97%
2	"`youtube`"	ja	6,620,927	118	3,211	99.95%
3	"`youtube`"	ko	714,992	36	1,339	99.81%
4	"`youtube`"	es	2,106,474	0	0	100.00%
5	"`youtube`"	und	1,418,710	0	0	100.00%
6	"`youtube`"	$S\backslash\{$en,ja,ko,es,und$\}$	5,264,150	20	169	100.00%
7	"`youtu`" `AND` "`be`"	en	11,188,872	530	10,328	99.91%
8	"`youtu`" `AND` "`be`"	ja	8,389,060	619	9,657	99.89%
9	"`youtu`" `AND` "`be`"	ko	4,560,793	1,193	43,584	99.05%
10	"`youtu`" `AND` "`be`"	es	2,271,712	27	829	99.96%
11	"`youtu`" `AND` "`be`"	und	2,856,415	37	1,556	99.95%
12	"`youtu`" `AND` "`be`"	$S\backslash\{$en,ja,ko,es,und$\}$	7,351,671	158	2,800	99.96%
complete	subcrawlers 1-12	all	53,557,950	3,061	77,055	99.86%
sample	"`youtube`" `OR` ("`youtu`" `AND` "`be`")	all	49,087,406	320,751	4,542,397	91.53%

3 A detailed comparison with Sampson et al. (2015) on the rate limit messages

In this section, we investigate a sampled dataset that contains tweets mentioning YouTube video URLs in Jan, 2017. The creation time of this dataset is closer to 2015. We believe that the design of rate limit messages is the same with the study in [1]. Differing from the observations we make in the Section 3 of the submitted paper, we find that the rate limit messages are designed differently back in 2017.

Figure 1 (left) plots the arrival of Twitter rate limit messages. We plot the arrival orders on the x-axis and the associated timestamps on the y-axis. Within the 1-minute span, every second has several rate limit messages arriving. Figure 1 (right) shows the bar chart of the secondly rate limit message number in the whole dataset. We obtain up to 4 rate limit messages per second. This observation is different with what we observe in the datasets curated in 2019, in which we obtain at most 1 rate limit message for each second.

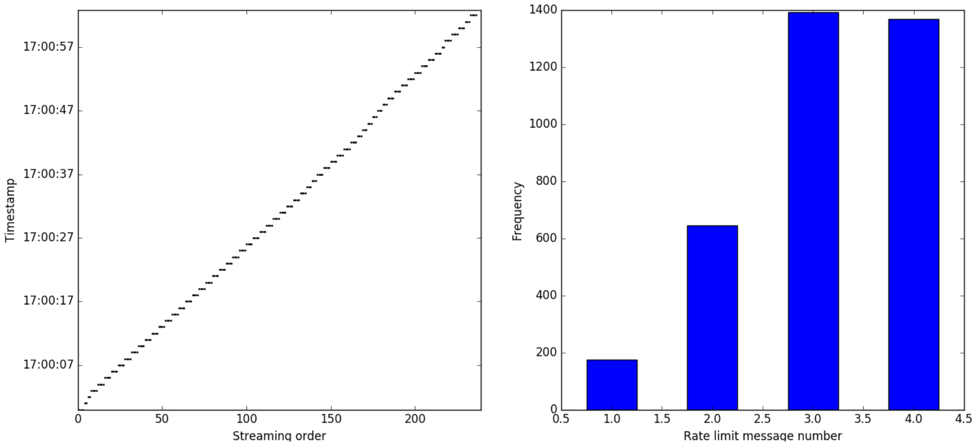

Figure 1: (left) The arrival of rate limit messages. x-axis: arrival order; y-axis: the associated timestamp. (right) The number of rate limit messages per second, shown as a bar chart.

Figure 2 (left) shows the scatter plot of rate limit messages with the timestamp on the x-axis and the associated integer on the y-axis. We notice that the integers of the arriving rate limit messages are not increasing monotonically, which confronts Twitter's official statement that "Limit notices contain a total count of the number of undelivered Tweets since the connection was opened [2]". However, in our 2019 datasets, the associated integers are increasing monotonically. The above observations prompts us to believe the rate limit messages (and the streaming client) are splitted into 4 parallel threads rather than 1. An explanation for the number less than 4 could be the streaming client has delivered all tweets within that particular thread, thus no rate limit message is returned for that thread. We propose Algorithm 1, which fits rate limit messages to 4 monotonic incremental lines for measuring the number of undelivered tweets. The simulation result is shown in Figure 2 (right).

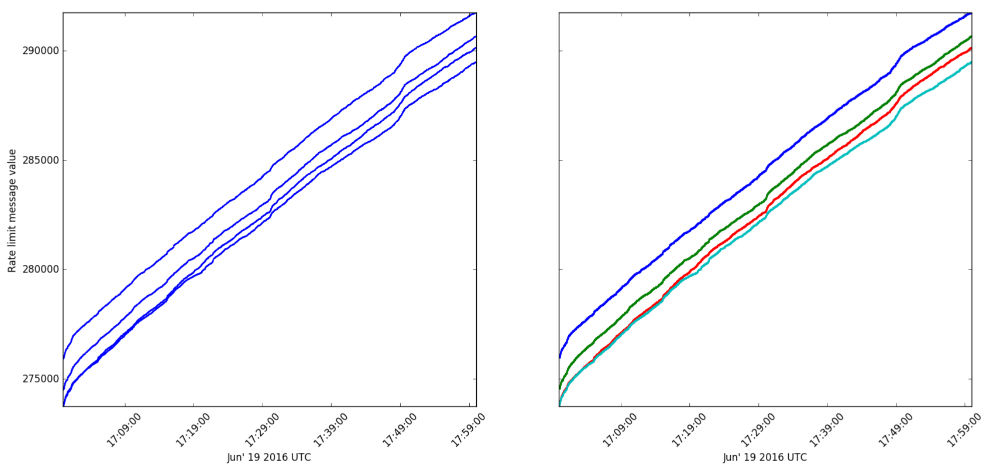

Figure 2: (left) Scatter plot of rate limit messages. x-axis: the timestamp; y-axis: the associated integer. (right) Coloring each rate limit message by using Algorithm 1.

```
input  : streaming rate limit messages
output: multiple monotonic incremental lines that constitute of rate limit messages
initialize lines by elements in first set of rate limit messages ;
while not at end of streaming do
    read rate limit messages then sort them reversely into CurRates;
    read last element from each line into PrevRates;
    m ⟵ len(PrevRates);
    n ⟵ len(CurRates);
    for i ⟵ 0 to n do
        rate ⟵ CurRates[i];
        if rate <= min(PrevRates) then
            add list(rate) to lines
        else
            for j ⟵ i to m do
                if rate > PrevRates[j] then
                    add rate to lines[j];
                    update PrevRates[j] by rate;
                    break;
                else if lines[j] not updated yet then
                    add PrevRates[j] to lines[j];
            end
    end
end
```

Algorithm 1: Fit rate limit messages to monotonic incremental lines.

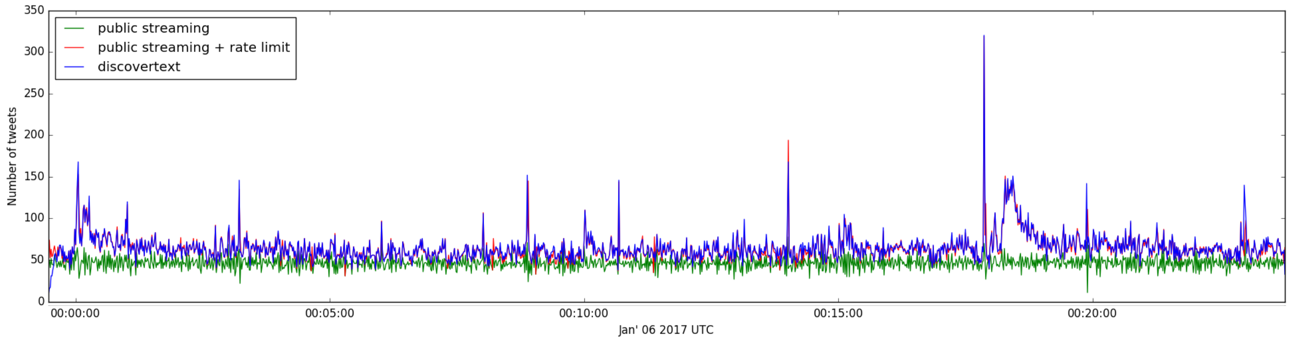

Figure 3: (left) Scatter plot of rate limit messages. x-axis: the timestamp; y-axis: the associated integer. (right) Coloring each rate limit message by using Algorithm 1.

To validate the 4-thread counters for rate limit messages, we obtain the complete data stream from a 3rd party data provider "**discovertext.com**[1]". This company provides access to the Twitter Firehose service at a cost. We started two streaming clients simultaneously on 2017-01-06, one with the public filtered streaming API, the other with the discovertext service. We track the temporal tweet volumes at three level: (1) the filtered stream (green); (2) the filtered stream plus the estimated missing volume from rate limit messages (red); (3) the complete volume from discovertext (blue). The missing volume is estimated by using Algorithm 1, which account the integers with rate limit messages into 4 parallel counters. The red line and blue line almost overlap each other. Our experiments show that the 4-thread counters are accurate measures for the missing volume.

If one uses the 1-thread counter to compute the missing volume, then the total volume is underestimated as the estimated missing volume now reduces to about 25%. This in turn makes the complete tweets appears to be much larger than the estimated total volume, which was the researchers found in [1]. We believe the discrepancy between our work and [1] are a result of different designs of the rate limit messages in 2015 and 2019.

[1] `https://discovertext.com/`

4 Minutely and secondly sampling rates

The temporal variations are much less prominent at the minutely and secondly levels, as shown in Figure 4.

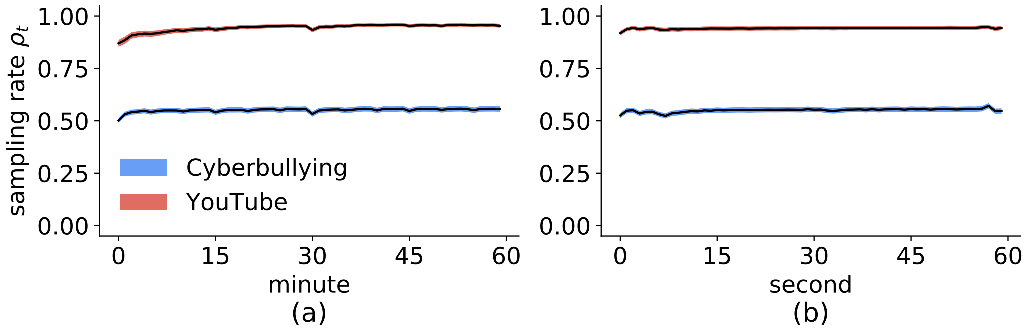

Figure 4: Minutely and secondly sampling rates.

References

[1] Sampson J, Morstatter F, Maciejewski R, Liu H. Surpassing the limit: Keyword clustering to improve Twitter sample coverage. In Proceedings of the 26th ACM conference on Hypertext and Social Media 2015 Aug 24 (pp. 237-245). ACM.

[2] Twitter Developer. Twitter rate limit notices. https: //developer.twitter.com/en/docs/tweets/filter-realtime/guides/ streaming-message-types 2020.