Should We Use the Sample? Analyzing Datasets Sampled from Twitter's Stream API

YAZHE WANG, Singapore Management University
JAMIE CALLAN, Carnegie Mellon University
BAIHUA ZHENG, Singapore Management University

13

Researchers have begun studying content obtained from microblogging services such as Twitter to address a variety of technological, social, and commercial research questions. The large number of Twitter users and even larger volume of tweets often make it impractical to collect and maintain a complete record of activity; therefore, most research and some commercial software applications rely on samples, often relatively small samples, of Twitter data. For the most part, sample sizes have been based on availability and practical considerations. Relatively little attention has been paid to how well these samples represent the underlying stream of Twitter data. To fill this gap, this article performs a comparative analysis on samples obtained from two of Twitter's streaming APIs with a more complete Twitter dataset to gain an in-depth understanding of the nature of Twitter data samples and their potential for use in various data mining tasks.

Categories and Subject Descriptors: H.3.5 [**Online Information Services**]: Data Sharing

General Terms: Experimentation

Additional Key Words and Phrases: Twitter API, sample, data mining

ACM Reference Format:
Yazhe Wang, Jamie Callan, and Baihua Zheng. 2015. Should we use the sample? Analyzing datasets sampled from Twitter's stream API. ACM Trans. Web 9, 3, Article 13 (June 2015), 23 pages.
DOI: http://dx.doi.org/10.1145/2746366

1. INTRODUCTION

Microblogging is an increasingly popular form of lightweight communication on the Web. Twitter as a typical and quickly emerging Microblogging service has attracted much attention. Millions of Twitter users around the world form a massive online information network by initiating one-way "following" relationships to others. Twitter users post brief text updates, which are commonly known as tweets, with at most 140-characters. The tweets posted by a user are immediately available to his direct followers, and can be quickly disseminated through the network via retweeting. Different from traditional blog platforms, where users write long articles with low update frequency, Twitter generates short and real-time messages in large volume daily. Some

This research is supported by the Singapore National Research Foundation under its International Research Centre@Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA) and Baihua Zheng is supported by Singapore Management University. This research was also in part supported by National Science Foundation (NSF) grant NSF IIS-1160862.
Authors' addresses: Y. Wang and B. Zheng, School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore 178902; emails: yazhe.wang.2008@phdis.smu.edu.sg, bhzheng@smu.edu.sg; J. Callan, Carnegie Mellon University, 5000 Forbes Avenue, Gates Hillman Complex 5407, LTI, Pittsburgh, PA 15213; email: callan@cs.cmu.edu.

DOI: http://dx.doi.org/10.1145/2746366

studies of the Twitter network reveal a variegated usage including daily chatter, conversation, information sharing, news reporting [Java et al. 2007], and a diverse topic coverage such as arts, family and life, business, travel, sci-tech, health, education, style, world, and sports [Zhao et al. 2011]. Many researchers have analyzed Twitter content and made interesting observations with real business value. For example, Sakaki et al. [2010], utilize Twitter to detect earthquakes; Bakshy et al. [2011] study different methods of identifying influential Twitter users, which may be useful for online marketing and targeted advertising; and Bollen et al. [2011] analyze Twitter user sentiment to predict the stock market.

One obstacle to using Twitter data is its huge size, as measured by the size of the user base, the volume of tweets, and the velocity of updates. The number of registered user profiles on Twitter reached half a billion in 2012 [Semiocast 2012], and collectively, Twitter users now send over 400 million tweets every day [Bennett 2012]. These numbers keep growing rapidly. It is challenging for third-party researchers and developers to collect and manage such a huge amount of data.

Twitter provides API functions to facilitate third-party users to access the data (https://dev.twitter.com/docs/). There are two main types of Twitter APIs: the REST API and the stream API. The REST API supports queries to Twitter user accounts and tweets, and it usually has very strict limits on the query rate (e.g., 150 requests per hour). Although the REST API provides flexible access to Twitter data from almost every angle, the rate limits make it not suitable for collecting large amounts of Twitter data and monitoring updates. On the other hand, the stream API provides almost real-time access to Twitter's global stream of public tweets. Once the connection is built, tweet data are pushed into the client without any of the overheads incurred by pulling data from the REST API. The stream API produces near real-time samples of Twitter's public tweets in large amounts. Owing to the advantages of the Twitter stream API, it is used as the data source for many applications and mining tasks, for example, topic modeling [Hong et al. 2012; Pozdnoukhov and Kaiser 2011], disease outbreak surveillance [Sofean and Smith 2012], and popular trend detection [Mathioudakis and Koudas 2010]. The convenience and immediacy of the stream API makes it a common source of Twitter data for a variety of research tasks. However, prior research has not addressed the issue of how well the sample data provided by the stream API represent the original data, and if do not, toward which properties the sample data might be biased.

In this work, we focus on characterizing the sample data from the Twitter stream API, studying possible sampling bias, if any, and understanding the implications of the findings to related applications. The Twitter stream API has different access priorities. For example, according to Twitter, the default *Spritzer* access provides a 1% sample of the complete public tweets, whereas *Gardenhose* access provides a larger 10% sample. However, Twitter does not reveal how the samples are generated and does not even guarantee that the sampling ratios are stable. These deficiencies make it difficult to perform theoretical analysis of the sample data. Therefore, in this article, we conduct a study based on experimentally analyzing the properties of sampled data and comparing them with a baseline complete dataset.

Due to limited storage capacity and the API access rate restrictions, we could not afford to collect the complete set of tweets generated by all the Twitter users (over 400 million tweets per day). Instead, we gather the "complete dataset" based on a relatively small subset of Twitter users: the Singapore Twitter users. We use all the tweets generated by these Singapore Twitter users during May of 2012 collected via the Twitter REST API as the *complete dataset*. Meanwhile, we gather the tweets of these users returned by the Twitter stream API at the same time period with two different access priorities, respectively, as the *sample datasets*. We perform comparative analysis

of the sample datasets with the complete dataset in terms of the basic tweet statistics, the content representativeness, the user coverage, and the user interactions. We find that the actual sampling ratios of the Spritzer sample and the Gardenhose sample are around 0.95% and 9.6% respectively. The sampled Twitter data represent the general user activity patterns and the tweet content of the complete dataset well even with a sampling ratio as small as 0.95%. However, although the sample data provide good coverage of interactions among active users, their coverage of infrequent users is less complete due to their lower probability of appearing in a sample. Extending the sampling period and increasing the sampling rate both help to improve the coverage of the user base and the accuracy of the interaction based user popularity estimation.

The rest of the article is organized as follows. Section 2 reviews related work. Section 3 describes the datasets used and the collection methods. Section 4 presents the main analysis results. Section 5 concludes the article.

2. RELATED WORKS

The huge volume of user-generated content in modern online social networks presents challenges to researchers for collecting and analyzing these data. A common practice to deal with this problem is to generate and analyze a representative sample of the complete dataset. There are two main issues for generating the sample: What is a good sampling strategy, and what is a good sampling ratio. In the case of the Twitter streaming API, the sample data are generated by some unknown strategies designed by Twitter with approximately fixed sampling ratios. Therefore, our focus in this work is on the unresolved question of whether the sample data generated by the Twitter streaming API are good enough for various mining and analysis tasks.

Very recent work by Morstatter et al. [2013] studies the same problem; however, there are important differences between their work and ours. The main difference is that they use a sample dataset collected from the Twitter stream API that focuses on a particular event: The Syria conflict from December 2011 to January 2012. We analyze a dataset that is not event-specific to provide more general observations. Their work also does not address the issue of sampling ratio, whereas we study two different sampling ratios and discuss their effects on the quality of the data obtained. In terms of methodology, Morstatter et al. measure the daily sampling ratio, whereas we also study the retweet ratio and the user tweet frequency distribution to provide a more comprehensive analysis. When studying the tweet content, they analyze the correlation of the ranks of the top hashtags and compare the topic distribution of the sample data with that of the complete data. We do not compare the topic distributions because we consider topic alignment across unlabeled datasets to be difficult, subjective, and unreliable. In this work, we study a rich set of terms in the tweet content including text terms, hashtags, URLs, and URL domains, and discuss the similarity of the sample data using these content terms to the complete data based on vocabulary coverage and frequency correlations. We also perform a sentiment classification task to compare the results obtained from the sample datasets and the complete dataset. In order to study user relationships, their work focuses on the user retweet network, whereas we study not only the user retweet relationships but also the mention relationships. Finally, their work analyzes the geolocation distribution of the tweets. However, because our dataset is based on Singapore Twitter users, the tweets are mainly located in Singapore; thus, geolocation distribution adds no new information.

Several other works discuss different Twitter data sampling methods. For example, Ghosh et al. [2013] study an expert generated tweet set and compare it with a random Twitter sample. They find that each dataset has its own relative merits. The expert tweets are significantly richer in information, more trustworthy, and capture

breaking news marginally earlier. However, the random sample preserves certain important statistical properties of the entire dataset and captures more conversational tweets. Choudhury et al. [2011] propose a diversity-based sampling approach to generate topic-centric tweet set. Our work does not study new sampling approaches, rather it investigates the characteristics of the existing and widely used Twitter samples. We focus on understanding whether the quality of the samples is good enough for various mining and analysis tasks.

The topic of network sampling and the effect of the imperfect data on the common network measurements have been widely studied. The earlier work of Granovetter proposes a network sampling algorithm that allows estimation of the basic network properties [Granovetter 1976]. Later, many common network sampling techniques are studied such as snowball sampling, random-walk–based sampling, node sampling, and link sampling [Lee et al. 2006; Yoon et al. 2006; Leskovec and Faloutsos 2006; Maiya and Berger-Wolf 2011]. These works compare the structural properties of the sample networks obtained by different methods with those of the original network and address the sampling bias of different methods. There are also works that discuss the effect of data errors and missing data on common network measures (e.g., centrality) [Kossinets 2003; Borgatti et al. 2006; Costenbader 2003]. Recently, William and Samuel [2010] study the forest fire network sampling method with different seed user selection strategies, and discuss their impact on the discovery of information diffusion on Twitter. However, this prior research does not apply to our problem because we study data that are sampled from the Twitter public tweet stream, not the Twitter user network. The (unknown) sampling mechanisms used by Twitter to generate data are presumably different from the network sampling methods discussed in this prior research.

The rising popularity of Twitter has inspired research into its characteristics. Kwak et al. [2010] conduct an exploratory analysis of the entire Twittersphere to study the topological characteristics of the Twitter network and information diffusion on it. Their results show a remarkable deviation from known characteristics of human social networks. They find that the Twitter network has a non–power-law degree distribution, short effective diameter, and low reciprocity, which establish Twitter's role as a new medium of information sharing. This study collects the entire Twitter network snapshot in its early stage (i.e., 2009). With the rapid growth of Twitter population, it becomes more and more difficult to handle the whole Twitter network, not to mention tracking its frequent information update. Therefore, much research has been performed on incomplete Twitter data. Java et al. [2007] analyze a Twitter subset with 76,000 users and 1 million tweets and categorize the users based on their intentions on Twitter. Their dataset is collected by periodically retrieving the most recent public tweet updates using an old version of the Twitter stream API that is no longer supported by Twitter. Naaman et al., study Twitter users' activity based on a small set of sampled non-organizational users, and classify them as "Meformers" and "Informers" according to whether they like to post tweets that are self-related or general informational [Naaman et al. 2010]. Zhao et al. [2011] characterize Twitter with topic modeling based on tweets collected from the Twitter stream API. They classify tweets into different topic categories and study the size distribution of these categories. Huberman et al. [2009] study the user activities and interactions in Twitter and reveal that the usage of Twitter is driven by a hidden network of connections underlaying the "declared" friend and follower relationships. The dataset they use consists of over 300,000 Twitter users and their tweets; however, the method of collection is not described. These works study Twitter datasets collected in several different ways, but none of them provides a discussion of the strengths and limitations of the data collection methods used and the representativeness of their datasets.

Krishnamurthy et al. [2008] perform descriptive analysis of the Twitter user base and make the first attempt to compare results of two datasets crawled by different techniques. The first dataset is collected by the snowball crawling of the Twitter network using the Twitter REST API. It starts with a small set of seed users and expands the user set by adding partial lists of the users being followed by the current users. The second dataset is obtained by the Twitter public timeline API, which provides continuously the 20 most recent tweet updates. The users associated with these tweets are extracted. They find that the analysis results on the two datasets are similar in terms of the user class, daily activity pattern, source interface usage, and geographic distribution. The work presented in our paper also analyzes Twitter datasets collected in different ways. However, it differs from the above work in three aspects. Firstly, the datasets analyzed have different properties. Our work analyzes three datasets based on the same set of Twitter users: (1) a complete Singapore user tweet dataset collected by crawling the Twitter REST API and (2) two sample datasets obtained via the Twitter stream API with different access priorities; our sample datasets are proper subsets of the complete dataset. In contrast, Krishnamurthy et al. [2008] study two datasets that may cover different sets of Twitter users. Second, the two studies have different purposes. In this article, our focus is not on characterizing the Twitter user base but on characterizing the Twitter stream API and understanding how well the data collected from the stream API represents the complete Twitter data space. Finally, due to the different study objectives, we perform analysis on different aspects of the datasets, including not only the users but also the tweet statistics, contents, and user interactions.

In addition to Twitter, several other popular social networks have been studied. YouTube is studied to understand the characteristics of user generated contents [Cha et al. 2007; Gill et al. 2007]. Kumar et al. [2006] analyze the structural properties of the Flickr and Yahoo!360 networks including the path lengths, density, change over time, and component structure. Mislove et al. [2007] verify the power-law, small-world, and scale-free properties of many popular online social networks including Flickr, YouTube, LiveJournal, and Orkut. Benevenuto et al. [2009] characterize the behaviors of a set of 37,000 collected users on online social networks such as Orkut, MySpace, Hi5, and LinkedIn. None of these works address the relationship between their analysis results and the data collection methods. Ahn et al. [2007] compare the topological characteristics and growth pattern of three large-scale online social networks: Cyworld, MySpace, and Orkut. They evaluate the validity of the snowball sampling method, which they use to crawl the networks. Their results reveal that with a sampling ratio above a certain threshold, snowball sampling captures the scaling behavior of the node degree distribution correctly, but it cannot estimate other metrics such as the clustering coefficient distribution and the degree correlation. Our article is different from that work because the sample datasets that we study are not obtained by the snowball sampling of the Twitter social network but by an unknown sampling method developed by Twitter on the public tweet stream.

3. DATA COLLECTION

In order to study sampling bias, we need the complete Twitter dataset to serve as the baseline, with which the sample datasets can be compared. However, collecting the complete Twitter stream is not practical for our study because of its cost. Instead of considering the full set of more than 500 million Twitter users, we focus on the complete set of Singapore Twitter users, which is a smaller group. We used all the tweets posted by these Singapore Twitter users within a 1-month period as the complete dataset. We also gathered all tweets by these Singapore users that appeared in the Spritzer and Gardenhose Twitter streams during the same timespan to create two sample datasets.

Table I. Description of the Datasets

Datasets	*Complete*	$Sample_{Gardenhose}$	$Sample_{Spritzer}$
API used for collection	REST	Stream (Gardenhose)	Stream (Spritzer)
Time period	May 2012		
Num. of users	151,041		
Num. of Tweets	13,468,661	1,297,304	128,647

The complete dataset was collected with the help of the social network mining research group of Singapore Management University.[1] To locate the Singapore Twitter users, a set of 58 popular Singapore Twitter users were manually selected as seeds. Initially, the user set only contained these seed users. The user set was then expanded by exploring the follower and friend lists of users in the set. A follower or a friend of a current user was added to the user set if either he specified his location to be "Singapore" or he followed at least three of the known Singapore users. In this way, a set of 151,041 Singapore Twitter users in 2012 was identified, which we believe covered the majority of the Singapore Twitter users.

After the set of users was constructed, the Twitter REST API was invoked to crawl the tweets generated by these users for a 1-month period beginning on May 1, 2012 and ending on May 31, 2012. The collected tweets formed the complete dataset, referred as *Complete*.

We collected two sample datasets at the same time period via the Twitter stream API using Spritzer and Gardenhose access priorities respectively. The Spritzer and Gardenhose streams output samples of the entire public tweet stream with different sampling ratios. According to Twitter, Spritzer provides an approximately 1% sample of the complete public tweets, whereas Gardenhose generates a larger sample with the sampling ratio around 10%. Twitter does not provide any description of the algorithms that generate the samples nor does it guarantee the sampling ratios to be stable. From the sampled tweets, we extracted the subsets that were posted by the identified Singapore users. In this way, two samples of the complete dataset were obtained, referred as $Sample_{Spritzer}$ and $Sample_{Gardenhose}$, respectively. Table I provides some basic information about the datasets.

4. ANALYSIS OF RESULTS

In this section, we perform detailed comparative analysis of the collected sample and complete datasets. Specifically, we compare them in terms of the tweet statistics, content representativeness, user coverage, and user interactions. Through the comparison, we try to understand the nature of the sample datasets, for which properties the sample datasets are representative of the complete dataset, and for which properties the sample datasets are not representative, and discuss the implications of our findings for certain mining tasks.

4.1. Tweet Statistics

We first study the sampling ratio and the basic tweet statistics in this section. We perform the analysis on the datasets collected over the 1-month time period and also present results on daily bases.

We begin the analysis by examining the actual sampling ratios of the two sample datasets from the Twitter stream API and present the average daily sampling ratios and standard deviations in Table II. As shown in Table II(a), the Singapore users generate around a half million tweets a day, on average. The Spritzer and Gardenhose

[1]https://sites.google.com/site/socnetmine/.

Table II. Average Daily Sampling Ratios

(a) Daily sampling ratios for tweets.

Daily statistic	$Complete$	$Sample_{Gardenhose}$		$Sample_{Spritzer}$	
	tweet#	tweet#	sampling ratio	tweet#	sampling ratio
Daily avg.	481,024	46,332	9.62%	4,634	0.96%
Std. dev.	67,446	6,637	0.15%	664	0.014%

(b) Daily sampling ratios for users.

Daily statistic	$Complete$	$Sample_{Gardenhose}$		$Sample_{Spritzer}$	
	user#	user#	sampling ratio	user#	sampling ratio
Daily avg.	35,316	15,769	44.55%	3,625	10.22%
Std. dev.	2,407	1,601	1.88%	484	0.85%

Table III. Daily Tweets and Retweets Ratios

Daily statistic	$Complete$		$Sample_{Gardenhose}$		$Sample_{Spritzer}$	
	tweet%	retweet%	tweet%	retweet%	tweet%	retweet%
Daily avg.	84.41%	15.59%	84.24%	15.76%	84.21%	15.79%
Std. dev.	0.56%	0.56%	0.54%	0.54%	0.76%	0.76%

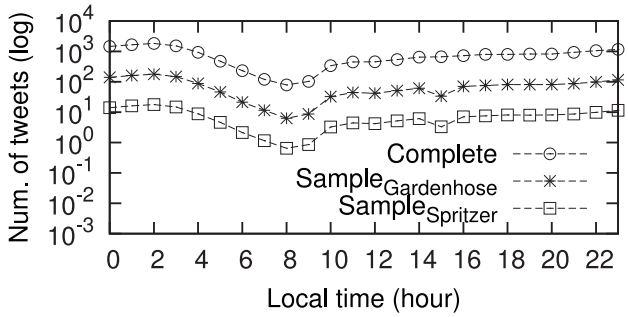
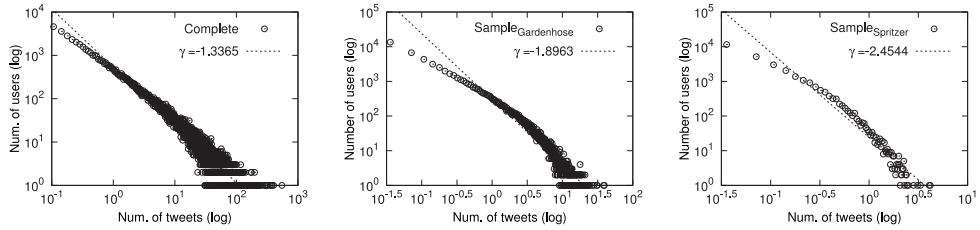


Fig. 1. Average hourly tweet count of the Singapore Twitter users of a 1-month period.

samples return around 0.96% and 9.6% of them, respectively. The actual tweet sampling ratios are both slightly lower than what Twitter announced (i.e., 1% and 10%). Table II(b) shows the sampling ratios on users each day. We find that, on average, there are around 35,000 Singapore users who generate tweets each day, and the Spritzer and Gardenhose samples capture around 10% and 45% of them, respectively. The sampling ratio for users is much higher than it is for tweets, which is not surprising; each tweet appears just once in the complete dataset, whereas a user may appear many times, thus increasing the likelihood that he will also appear in a sample.

Next, we study whether the sample datasets preserve the general tweeting patterns of the Twitter users. Table III lists the average proportions of the original tweets and retweets generated by Twitter users each day. As observed from the table, among all the tweets published daily, about 85% are original tweets and 15% are retweets. The same ratio between original tweets and retweets is captured by both sample datasets. Figure 1 further illustrates the average hourly tweet counts of the three datasets for all the 31 studied days. We observe that the Singapore users tend to be more active at the nighttime. The tweeting frequency increases rapidly after 17:00, and peaks at 22:00. Then it drops quickly through midnight and hits the bottom at 4:00. Thereafter, as the new day starts, the users gradually regain activity, and the tweeting frequency rises slowly through the day. The sampled datasets both reflect the same hourly tweeting frequency pattern of the users. The results indicate that both the small Spritzer sample



(a) $Complete$
$SE_\gamma = 0.008$, $R^2 = 0.91$

(b) $Sample_{Gardenhose}$
$SE_\gamma = 0.024$, $R^2 = 0.93$

(c) $Sample_{Spritzer}$
$SE_\gamma = 0.072$, $R^2 = 0.94$

Fig. 2. Average daily tweeting frequency distributions of the Singapore users of a 1 month period. γ is the estimated power-law exponent. SE_γ is the standard error of γ. R^2 is the square error of the power-law fitting.

and the larger Gardenhose sample obtained via the Twitter stream API reflect the general user tweeting patterns of the complete dataset accurately.

In addition, we analyze the tweet patterns based on individual users. We plot the distributions of the user average daily tweeting frequency in Figure 2. The average daily tweeting frequency distribution of the Singapore users approximates a power-law distribution with the exponent of –1.3365. However, the user average daily tweeting frequency distributions captured by the Spritzer and Gardenhose samples fit the power-law distributions with the different exponents of –2.4544 and –1.8963, respectively. Therefore, the sample data preserve the scaling pattern of the user tweeting frequency distribution (i.e., power-law) but tend to overestimate the proportion of the users with low tweeting frequency, and the overestimation is more serious in the sample with a smaller sampling ratio (e.g., Spritzer).

4.2. Content Representativeness

Twitter data are also widely used for performing mining tasks such as event detection, sentiment analysis, content summarization, and topic modeling. As many of these tasks are built upon analyzing the tweet contents, it is important to understand if the tweet contents in the sample datasets from the Twitter stream API accurately represent those in the complete dataset.

For each dataset, we extracted the vocabularies of four common types of text representation: text terms, hashtags, URLs, and URL domains. We performed lightweight processing of the text terms by eliminating stopwords,[2] punctuation, and non-English terms. For each dataset and each method of representation (e.g., Spritzer URLs), we record the frequency of the vocabulary item and its rank each day and for the entire 1-month timespan, as described in Table IV. We analyze the correspondence of the vocabularies of the complete dataset and the sample datasets using four metrics as described in Table V and display the results in Table VI.

We measure how well the vocabulary of a sample dataset covers the vocabulary of the complete dataset using two metrics: *size ratio* metric and the *collection term frequency (CTF)* metric. Each metric provides a different perspective on how well the sample vocabulary covers the complete vocabulary.

The *size ratio* metric calculates the proportion of the unique terms in a vocabulary of the complete dataset that are captured by a sample dataset. As observed from Table VI(a), the Spritzer sample only covers around 6% of the text vocabulary, 2.5% of the hashtag vocabulary, 1% of the URL vocabulary, and 3.7% of the URL domain vocabulary in each day. The size ratios of the vocabularies of the Gardenhose sample

[2]We use a stopword dictionary with 429 distinct words (http://www.lextek.com/manuals/onix/stopwords1.html).

Table IV. Symbols for a Vocabulary

Symbol	Description
V	The vocabulary for all the text words/hashtags/URLs/URL domains appearing in the set of tweets of the studied time period (e.g., 1 day or 1 month).
$\|V\|$	The size of the vocabulary V; It is the number of terms that exist in the vocabulary.
t	A term in a vocabulary representing a text word/hashtag/URL/URL domain.
$t.f^V$, $\overline{f^V}$	$t.f^V$ is the frequency of t in vocabulary V; It is the number of times that t appears in the tweet set on which the vocabulary is built. $\overline{f^V}$ is the average frequency of all the terms in V.
$t.r^V$, $\overline{r^V}$	$t.r^V$ is the rank of t in vocabulary V; It is the rank of the t in the vocabulary based on its frequency. $\overline{r^V}$ is the average value of all the term ranks in V.

Table V. Comparison Metrics for Vocabularies

Metric	Description
S_{size}	The size ratio of a vocabulary of the sample tweet set (V_S) and the corresponding vocabulary of the complete tweet set (V_C). $S_{size} = \frac{\|V_S\|}{\|V_C\|}$
S_{ctf}	The collection term frequency (CTF) ratio of a vocabulary of the sample tweet set (V_S) and the corresponding vocabulary of the complete tweet set (V_C).$S_{ctf} = \frac{\sum_{t \in V_S} t.f^{V_C}}{\sum_{t \in V_C} t.f^{V_C}}$
S_{pcc}	The Pearson product-moment correlation coefficient of the term frequency values of a vocabulary of the sample tweet set (V_S) and the corresponding vocabulary of the complete tweet set (V_C).$S_{pcc} = \frac{\sum_{t \in V_S \cap V_C} (t.f^{V_S} - \overline{f^{V_S}})(t.f^{V_C} - \overline{f^{V_C}})}{\sqrt{\sum_{t \in V_S \cap V_C} (t.f^{V_S} - \overline{f^{V_S}})^2} \sqrt{\sum_{t \in V_S \cap V_C} (t.f^{V_C} - \overline{f^{V_C}})^2}}$
S_{scc}	The Spearman's rank correlation coefficient of the term rank values of a vocabulary of the sample tweet set (V_S) and the corresponding vocabulary of the complete tweet set (V_C).$S_{scc} = \frac{\sum_{t \in V_S \cap V_C} (t.r^{V_S} - \overline{r^{V_S}})(t.r^{V_C} - \overline{r^{V_C}})}{\sqrt{\sum_{t \in V_S \cap V_C} (t.r^{V_S} - \overline{r^{V_S}})^2} \sqrt{\sum_{t \in V_S \cap V_C} (t.r^{V_C} - \overline{r^{V_C}})^2}}$

are much higher due to the higher sampling ratio (i.e., the Gardenhose sample covers around 26% of the text vocabulary, 16% of the hashtag vocabulary, 9% of the URL vocabulary, and 18% of the URL domain vocabulary in each day). In addition, we find that the size ratio of the URL vocabulary almost equals the tweet sampling ratio, while the size ratios of text terms, hashtags, and URL domains are much larger than the tweet sampling ratio. This result is easily explained. Many of the URL terms occur only once in the complete dataset; thus, the odds of seeing them in a sample depend strongly on the sample size. In contrast, many individual text terms, hashtags, and URL domains have higher occurrence frequencies, thus samples tend to cover more of these vocabularies.

The size ratio metric indicates that the sample datasets cover only small proportions of the vocabularies for different representations of the complete dataset. However, the size ratio metric treats every term equally, and it is skewed by the many terms that appear just a few times in the dataset. Based on our observation, the infrequent terms are more likely to be typographical errors and/or user-created words, which may be less important for studying Twitter contents.

In order to distinguish the frequent terms from the infrequent ones, we adopt another vocabulary coverage metric, namely *collection term frequency (CTF)* ratio [Callan and Connell 2001]. This metric also calculates the proportion of the terms in the complete vocabulary that are covered by a sample vocabulary, but it weights each term with its

Table VI. Content Representativeness Based on Four Types of Vocabularies Daily and for all the Studied Days (i.e., 1 Month)

(a) Size ratio (S_{size})

Daily statistic	$Sample_{Gardenhose}$				$Sample_{Spritzer}$			
	text	hashtag	URL	URL domain	text	hashtag	URL	URL domain
Daily avg.	0.257	0.160	0.090	0.182	0.064	0.025	0.010	0.037
Std. dev.	0.021	0.019	0.013	0.023	0.002	0.002	0.001	0.003
All days	0.237	0.185	0.092	0.184	0.062	0.032	0.010	0.034

(b) CTF ratio (S_{ctf})

Daily statistic	$Sample_{Gardenhose}$				$Sample_{Spritzer}$			
	text	hashtag	URL	URL domain	text	hashtag	URL	URL domain
Daily avg.	0.915	0.622	0.121	0.915	0.750	0.371	0.034	0.837
Std. dev.	0.013	0.044	0.018	0.013	0.022	0.038	0.006	0.020
All days	0.977	0.791	0.144	0.960	0.939	0.603	0.054	0.926

(c) Pearson product-moment correlation coefficient (S_{pcc})

Daily statistic	$Sample_{Gardenhose}$				$Sample_{Spritzer}$			
	text	hashtag	URL	URL domain	text	hashtag	URL	URL domain
Daily avg.	0.997	0.9715	0.911	0.987	0.975	0.856	0.655	0.974
Std. dev.	0.002	0.021	0.045	0.005	0.005	0.040	0.195	0.013
All days	0.100	0.993	0.990	0.988	0.999	0.979	0.973	0.985

(d) Spearman's rank correlation coefficient (S_{scc})

Daily statistic	$Sample_{Gardenhose}$				$Sample_{Spritzer}$			
	text	hashtag	URL	URL domain	text	hashtag	URL	URL domain
Daily avg.	0.812	0.641	0.433	0.736	0.705	0.552	0.268	0.754
Std. dev.	0.009	0.016	0.022	0.018	0.013	0.044	0.050	0.036
All days	0.817	0.691	0.442	0.706	0.811	0.623	0.266	0.692

frequency of occurrence to give more credits to the frequent terms, which are believed to be more important in the dataset. The closer the CTF ratio is to 1, the more a sample contains the terms that are frequent, and thus presumably important, in the complete dataset.[3] The results of the CTF ratio are displayed in Table VI(b). Generally, we find that the CTF ratio for every vocabulary is much higher than the corresponding size ratio, and the Gardenhose sample has higher CTF ratios than the Spritzer sample due to the higher sampling ratio.

Closer inspection reveals that different types of text representations behave differently.

—For the daily text vocabularies, the CTF ratios were significantly high (e.g., they were around 0.75 for the Spritzer sample, and they exceeded 0.9 for the Gardenhose sample). Therefore, even small sample datasets capture the important text terms very well.
—The CTF ratios for Spritzer samples were about 0.37 for the daily hashtag vocabularies, whereas the Gardenhose CTF ratios were about 0.62. Sampling over a 1-month timespan improves Spritzer coverage to 0.60 and Gardenhose coverage to 0.80. These results suggest that one might want to be cautious about drawing conclusions from daily variations in hashtag occurrences in a Spritzer stream, and even conclusions based on a Gardenhose stream (10× larger) will miss significant amounts of hashtag activity. Observations based on a 1-month timespan are more reliable but will necessarily miss a significant amount of hashtag activity.

[3]Note that stopwords are not included in this comparison. If stopwords were included, they would dominate the weighting, and all methods would have a CTF ratio close to 1.

—The CTF ratios for URLs are very low for both sample datasets, due primarily to the fact that many of the URLs only occur once in the dataset. One may interpret this result as indicating that many tweeted URLs are unimportant and thus safely ignored; or, it may mean that tweeting frequency is a less reliable method of determining the importance of a tweeted URL. In any case, Spritzer and Gardenhose streams provide only very approximate information about the distribution of URLs in the underlying complete stream.
—Instead of using individual URLs, some researchers may be interested in only the domains that produce URLs, for example, to identify information sources that are popular with Twitter users. We find that the Spritzer sample has very high CTF ratios (i.e., around 0.83) for URL domains, and the CTF ratios are even higher for Gardenhose samples (i.e., about 0.92). Therefore, even though the sample datasets do not provide enough information for studying the popularity of individual URLs, they preserve the important URL domains very well.

In addition to analyzing the daily vocabularies, we also analyze the cumulative vocabularies of the 1-month period. We find that the cumulative data obtained by extending the sampling time period does not improve the raw vocabulary coverage, as the size ratios are not significantly improved with the cumulative vocabularies. This observation is consistent with Heaps' Law, which predicts the continued growth of the vocabulary as more texts are observed [Heaps 1978]. However, the long sampling period helps to improve the coverage of the frequent terms, as indicated by the increase in CTF ratios for the cumulative vocabularies.

The size ratio and CTF ratio metrics only evaluate the proportions of the (frequent) terms that are captured by the sample datasets. They do not evaluate whether the frequency information of the captured terms is well preserved by the sample datasets. In other words, they do not evaluate whether the term frequency information obtained from the sample datasets is correlated with the actual term frequencies in the complete dataset. For mining tasks such as event detection, tweet content summarization, and sentiment analysis, the term frequency information is crucial. Therefore, in the following analysis, we use another two metrics to study the quality of the term frequency information in the sample datasets: the Pearson product-moment correlation coefficient (PCC) and the Spearman's rank correlation coefficient (SCC).

The PCC metric measures the linear dependency of the *term frequencies* in a sample vocabulary and the complete vocabulary. Its value is in the range of $[-1, 1]$. The value is close to 1 if the term frequencies in the sample dataset and the complete dataset are strongly correlated, the value is 0 when the term frequencies are uncorrelated, and the value is –1 when the term frequencies are inversely correlated. SCC measures the linear dependency of the frequency-based *term rankings* of a sample vocabulary and the complete vocabulary. To calculate the SCC score, the terms in a vocabulary are ranked by the decreasing order of their frequencies in the dataset. Rank ties are handled by assigning a rank that equals to the average of their positions in the ranked list. For example, if the top two terms both have the highest frequency in the ranked list, that is, there is a tie between position 1 and position 2, the ranks assigned to these two terms are both $1.5 = \frac{1+2}{2}$. The SCC score is calculated based on the term rankings with a similar function as the PCC metric (see Table V), and it has the same value range.

The results of these two metrics are shown in Tables VI(c) and VI(d). We find that in most of the cases, the PCC values are above 0.8 and 0.9 for the Spritzer and Gardenhose vocabularies, respectively. Thus, the term frequencies of the sample datasets are linearly correlated with those of the complete dataset. In other words, the sample datasets

Table VII. Features Derived for Sentiment Classification

Feature category	Features
Overall scores (6)	Sum of positive and negative scores for Adjectives. Sum of positive and negative scores for Adverbs. Sum of positive and negative scores for Verbs.
Score ratios to number of terms (6)	Ratios of positive and negative scores to total number of terms for each part of speech.
Positive to negative score ratios (3)	Positive to negative scores ratio for each part of speech.
Negation (1)	Percentage of negated terms in a tweet

accurately estimate the relative frequencies of the terms in every vocabulary. The results of the SCC metric are not as good as those of the PCC metric, but still are reasonably high, except for the URL vocabularies. Therefore, the sample datasets predict the term rankings of text terms, hashtags, and URL domains to certain extent. The degradation of the SCC performance is mainly caused by the ties in the term ranking. The URL vocabularies have the most rank ties because many of the URLs only appear once in the datasets and thus have the worst SCC performance. The improvement of the PCC and SCC scores by extending the sampling time period is not very obvious.

4.2.1. Sentiment Analysis. To demonstrate the usefulness of the sample data for analyzing Twitter content, we perform a sentiment classification task on the sample datasets and the complete dataset and then compare the results. Sentiment classification is an opinion mining activity concerned with determining what is the overall sentiment orientation of the opinions contained within a given document (e.g., tweet). The sentiment orientation can be classified as positive or negative. We implement the binary classifier described by Ohana and Tierney to analyze the sentiment orientation of tweets [Ohana and Tierney 2009]. We extract sentiment features of tweets using SentiWordNet and then train a SVM classifier to assign sentiment labels to the tweets in each dataset. SentiWordNet is a lexical database for opinion mining. Given a term and its part-of-speech tag, SentiWordNet returns three sentiment scores ranging from 0 to 1—positivity, negativity, and objectivity—each indicating the term's sentiment bias. The sum of the three scores equals to 1. In our experiment, the GATE Twitter part-of-speech tagger (https://gate.ac.uk/wiki/twitter-postagger.html) was used to perform part-of-speech analysis on tweets. SentimentWordNet scores were then calculated for terms found. Sentiment features were derived from the scores as described in Table VII. A total of 16 features were generated.

We train a linear SVM classifier with a set of 1,224 manually classified tweets from a separate dataset. The training data consist of 570 positive tweets and 654 negative tweets. We apply the trained classifier to predict the sentiment orientation of the tweets in our datasets.

First, we analyze Twitter's daily overall sentiment polarity by counting the percentages of tweets with positive and negative sentiment orientation respectively in each day. In order to compare the difference of the sample datasets with the complete dataset, we calculate the absolute difference of the positive tweet percentages of a sample dataset and the complete dataset in each day.[4] Figure 3 presents the percentage difference distributions of the Spritzer and Gardenhose samples over the studied 30 days. We observe that for all the 30 days, the percentage differences of both datasets are fairly small (i.e.,

[4]The absolute difference of the negative tweet percentages is the same as that of the positive tweets percentages.

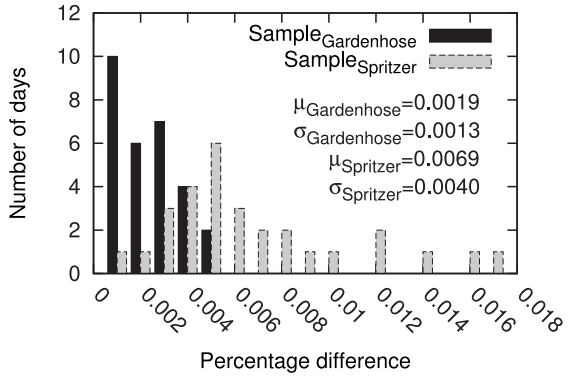
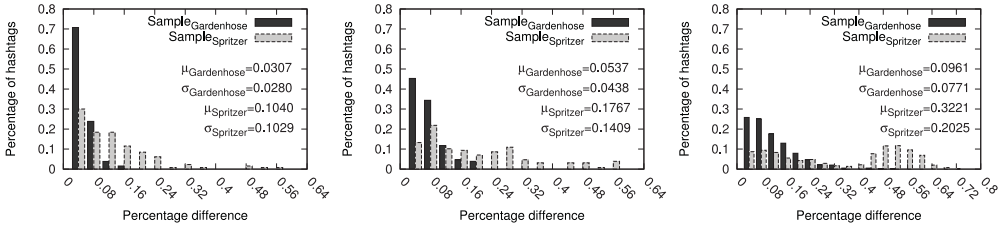


Fig. 3. The percentage difference of the positive (or negative) tweets daily. The x-axis is the binned difference. The y-axis is the count of days in each bin. μ is the average difference of 30 days, and σ is the standard deviation.



(a) Very popular hashtags (b) Moderately popular hashtags (c) Less popular hashtags

Fig. 4. The percentage difference of the positive (or negative) tweets in different hashtag groups with different popularity. The x-axis is the binned difference. The y-axis is the percentage of hashtags in each popularity group. μ is the average difference of all the hashtags in each popularity group, and σ is the standard deviation

less than 1.8%), which indicates that both sample datasets can reflect Twitter's daily sentiment orientation very accurately. It is not surprising that the larger Gardenhose sample shows better accuracy at predicting Twitter's daily sentiment polarity. Its daily percentage differences are all smaller than 0.6%.

Besides the overall sentiment polarity orientation, we also study Twitter's sentiment orientation toward certain hashtags. We group the hashtags that are captured by both sample datasets based on their popularity in the complete dataset. We categorize the hashtags that were used by more than 1,000 tweets as very popular, the hashtags that were used by less than 1,000 but more than 500 tweets as moderately popular, and the hashtags that were used by less than 500 tweets as less popular. We ignore the hashtags that were used by less than 100 tweets because of their lack of popularity. We infer Twitter's sentiment orientation to a hashtag also by the percentages of the positive and negative tweets containing this hashtag. We use the percentage difference to evaluate the error of the sample datasets for predicting Twitter's sentiment polarity to each hashtag. Figure 4 shows the percentage difference distributions of the three hashtag groups. We find that the Gardenhose sample relatively accurately reflects sentiment orientation to the hashtags in the very popular and moderately popular groups. In these two groups, the Gardenhose sample captures the sentiment orientation for most of the hashtags (i.e., more than 90% and 80% of the hashtags respectively) with

Table VIII. Average Number of Tweets Daily of the Captured and Missed Users

Daily statistics	$Sample_{Gardenhose}$		$Sample_{Spritzer}$	
	captured users	missed users	captured users	missed users
Daily avg.	23.99	4.20	48.45	10.13
Std. dev.	2.56	0.08	2.49	0.55

the percentage difference less than 8%. The performance of the Gardenhose sample decreases greatly for the less popular hashtags. It can only guarantee small percentage difference (e.g., less than 8%) for around 50% of the hashtags. Not surprisingly, the performance of the smaller Spritzer sample is not as good as the Gardenhose sample. It can only achieve less than 10% percentage difference for around 70% of the very popular hashtags. For those not so popular hashtags, percentage differences for most of them are not small.

To sum up, in this subsection, we find that both the Spritzer and Gardenhose samples can be used to estimate Twitter's overall sentiment orientation. However, for individual hashtags, the Gardenhose sample can be used to estimate Twitter's sentiment orientation for very popular and moderately popular hashtags, while the Spritzer sample may be only suitable for estimating Twitter's sentiment for very popular hashtags.

4.3. User Coverage

According to the analysis in Section 4.1, the daily user sampling ratios of the Spritzer dataset and the Gardenhose dataset are about 10% and 45%, respectively. Thus, more than half of the users in the complete dataset who tweet each day are not captured by the sample datasets. In the following analysis, we compare the properties of the users captured by the sample datasets with those of the users that are missed. Here, the missed users refer to the users who generate tweets and are included in the complete dataset, but their tweeting behaviors are not captured by the sample datasets.

First, we calculate the average numbers of tweets generated daily by the captured and missed users, respectively, and report the results in Table VIII. The results show that the captured users tend to publish more tweets than the missed users. To be more specific, users captured by the Spritzer sample publish more than 48 tweets daily, on average, while the missed users generate only around 10 tweets daily. The average daily tweeting frequencies of the captured users and the missed users in the Gardenhose sample are around 24 and 4, respectively. These results imply that the samples generated by Twitter Stream API cover more information from the active users who tweet frequently every day, and they may lose the voice of those less active users. The bias is more significant with the sample dataset having a smaller sampling ratio (e.g., $Sample_{Spritzer}$).

Although the user samples tend to cover active users, we expect that the cover of low-frequency users can be improved by extending the sampling period. In other words, with the extension of the sampling period, the chance of discovering inactive users will increase. To verify how fast the sample user set grows as the sampling period increases, we first identified the 37,124 Singapore users who published tweets on the first day of the complete dataset as the baseline user set. Then, we monitored the Twitter stream API to see when these users were observed. Once a user's tweets were spotted, we added that user to the sample user set. We performed this monitoring using both the Spritzer stream and the Gardenhose stream for 60 days and maintained two sample user sets, respectively. Figure 5 plots the sizes of these sample user sets over the 60 days.

The sample user set from the Gardenhose stream grew quickly for the initial 10 days, eventually reaching about 87% of the baseline user set. After that, it converged slowly

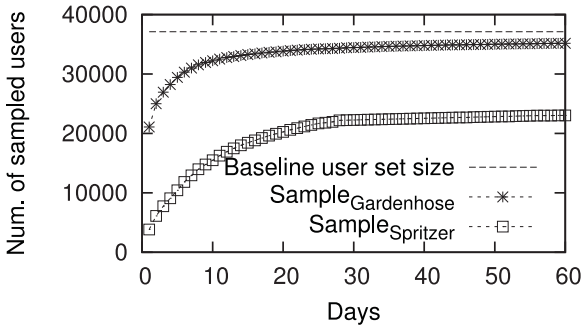


Fig. 5. Total number of users observed over time.

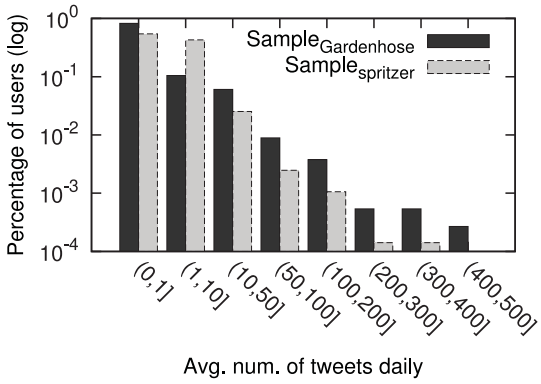


Fig. 6. Average daily tweeting frequency distribution of the missed users.

to the baseline user set. After 60 days of monitoring, 35,174 Singapore users were discovered, which covered about 95% of the baseline user set. However, the set of users found in the Spritzer stream converged much more slowly due to the low sampling ratio. After 60 days, it only had captured 23,036 users, which covered 62% of the baseline user set. We conclude that extending the sampling period helps to improve the user coverage, and a period of 10 days is enough for discovering most of the baseline users (i.e., more than 85% of them) using the Gardenhose stream.

After 60 days of monitoring, there were still some users that had not been seen in the Spritzer and Gardenhose samples. We examined the daily tweeting frequency distributions of these missed users. The results are displayed in Figure 6. More than 90% of the missed users in both datasets tweeted infrequently (i.e., generated no more than 10 tweets a day). Therefore, it is not surprising that these users did not appear in any sample.

The daily tweeting frequency distribution of users not found in the Gardenhose sample is significantly skewed toward the extremely low frequency users (i.e., the users who tweet less than once a day, on average) compared with those of the Spritzer sample. This result confirms that the higher sampling rate increases the chance of discovering low-activity users and that only the extremely inactive users are missed.

However, there is a group of users who are very active (i.e., post more than 100 tweets daily) but do not appear in any sample. Although this group is small (less than 1% of the missed users), it is perhaps surprising that these users do not appear in either of our samples. We manually checked the profiles and the tweets of these users. Most of these users are organizational users or marketers who periodically tweet

Table IX. The Proportions of the Reciprocal and Directed User Mention Interactions Extracted from the Complete and Sample Datasets Daily and for all the Studied Days (i.e., 1 month)

Daily statistics	*Complete*		$Sample_{Gardenhose}$		$Sample_{Spritzer}$	
	reciprocal	directed	reciprocal	directed	reciprocal	directed
Daily avg.	11.44%	88.56%	7.18%	92.82%	4.10%	95.90%
Std. dev.	0.34%	0.34%	0.51%	0.51%	0.47%	0.47%
All days	10.67%	89.33%	9.25%	90.75%	5.73%	94.27%

URLs linking to external websites or product promotions. We believe that these users' tweets are intentionally excluded from the sample stream, perhaps because Twitter has identified them as robots, spammers, or other undesirable information producers. This group of missing users may not be a problem for most researchers, because it is a tiny group, and because the information that they provide may not represent "real" user content. However, they might be important to researchers who study robot and spammer behavior.

4.4. User Interactions

Another type of valuable information embedded in the Twitter data is the interactions between users. There are two main types of interactions between Twitter users: mention interactions and retweet interactions. A Twitter user *mentions* another user by inserting "@username" into the body of his tweet. Mentions are usually used to signify quotes from other users' posts or to send direct messages to the users that are mentioned. A Twitter user *retweets* another user's tweet by clicking the "Retweet" button under that tweet. A retweet is a reposting of someone else's tweet.

Mention and retweet interactions can be directed or reciprocal. Usually, a directed interaction indicates an "informational relationship" between users because the information only flows one way from a user to another, while a reciprocal interaction indicates a "friendship relationship" because there is communication between users. In the rest of the article, we use the term "interaction" and "relationship" interchangeably.

Mention and retweet information can be obtained from tweet metadata. They are commonly used for the tasks such as understanding users' roles in Twitter, identifying key players, and modeling information diffusion. In this section, we extract the mention and retweet relationships among the Singapore Twitter users from the complete and sample datasets, respectively, and study the representativeness of the sample datasets on these relationships. We first analyze the mention relationships. The same analysis is performed on the retweet relationships as well.

We first examine whether the sample datasets represent the proportions of the reciprocal and directed mention relationships in the complete dataset. Table IX provides the results. Among all the Singapore Twitter users, about 11.4% of the mention relationships captured daily in the complete dataset are reciprocal, and 88.5% of them are directed. However, in the Spritzer sample, 4% of the mention relationships captured are reciprocal, and 96% of them are directed; while in the Gardenhose sample, the proportions of the reciprocal and directed relationships are around 7% and 93%, respectively. These observations tell us that the sample datasets tend to underestimate the amount of the reciprocal relationships, and the underestimation of the Spritzer sample is more serious than that of the Gardenhose sample. Again, we find that extending the sampling period improves the estimation of the proportions of the reciprocal and directed relationships. The Gardenhose sample for a 1-month period has similar proportions of these relationships with the complete dataset.

Next, we study how many of the relationships in the complete dataset are captured by the sample datasets. We calculate the recall of the reciprocal, directed, and all

Table X. The Recall of the User Mention Interactions Daily and for all the Studied Days (i.e., 1 month)

(a) Recall on all users

Daily statistic	$Sample_{Gardenhose}$			$Sample_{Spritzer}$		
	$Recall_{Dir.}$	$Recall_{Rec.}$	$Recall_{All}$	$Recall_{Dir.}$	$Recall_{Rec.}$	$Recall_{All}$
Daily avg.	0.181	0.109	0.173	0.021	0.007	0.020
Std. dev.	0.013	0.021	0.014	0.001	0.001	0.001
All days	0.244	0.208	0.240	0.040	0.020	0.038

(b) Recall on captured users

Daily statistic	$Sample_{Gardenhose}$			$Sample_{Spritzer}$		
	$Recall_{Dir.}$	$Recall_{Rec.}$	$Recall_{All}$	$Recall_{Dir.}$	$Recall_{Rec.}$	$Recall_{All}$
Daily avg.	0.213	0.143	0.206	0.064	0.053	0.063
Std. dev.	0.014	0.026	0.016	0.004	0.008	0.004
All days	0.283	0.263	0.281	0.095	0.086	0.095

relationships, and list the results in Table X(a). As observed from the table, the Spritzer dataset and the Gardenhose dataset recover around 2% and 17% of all the daily interactions among all the Singapore Twitter users, respectively. Given the tweet sampling ratios of the two datasets (i.e., less than 1% and 10%), these Recall values are reasonable. However, many applications that utilize the user interaction information need a complete view of the user relationships, for example, to analyze the properties of the user network, and study network based information diffusion. For these applications, the sample datasets do not provide sufficient information.

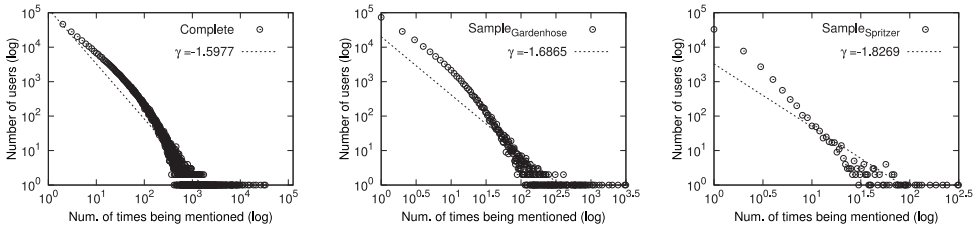
Even though the recall of the mention relationships is increased by extending the sampling period to 1 month, it is still far from complete; that is, in the best case, the Gardenhose sample captures only 24% of all the interactions of the Singapore Twitter users based on the 1-month period of sampling. To get the nearly complete user relationships, much longer sampling time may be needed. However, the relationships among the Twitter users are relatively dynamic. The information extracted from the historical data may lose effectiveness. We also notice that the Recall of the reciprocal relationships is generally smaller than the Recall of the directed relationships, which indicates that reciprocal relationships are harder to capture from sample data.

The analysis described in Section 4.3 found that a sample dataset only covers a proportion of the active users every day. To make our analysis more fair, we also calculate the Recall of the interactions between these captured users; those results are displayed in Table X(b). As observed, if we only focus on the captured users, the Recall values of all the daily interactions of the Spritzer dataset and the Gardenhose dataset increase to around 6% and 20%, respectively. By extending the sampling period to 1 month, the two datasets capture around 9.5% and 28% of all the mention interactions between the captured users respectively. Since much of the interaction information between users is missing from the sample data, researchers cannot construct a user mention network from the sample data that has properties similar to the user network constructed from the complete dataset.

In the final set of analysis, we study the intensity of the users being mentioned. This piece of information is important for studying users' role and locating key players in Twitter. Usually, the popular users that are mentioned many times by many users are more important than the less frequently mentioned users in the Twitter space. To perform the analysis, we first extracted the frequencies of the users being mentioned in the complete and the sample datasets respectively, and produced rankings of the users based on these frequencies. Then the CTF ratio, PCC score, and SCC score were calculated based on the extracted information to evaluate the effectiveness of the sample datasets at preserving the user popularity information. The results are presented in Table XI.

Table XI. Estimation of User Popularity Based on the Frequency of Being Mentioned Using the Data Daily and for All the Studied Days (i.e., 1 month)

Daily statistic	$Sample_{Gardenhose}$			$Sample_{Spritzer}$		
	S_{ctf}	S_{PCC}	S_{SCC}	S_{ctf}	S_{PCC}	S_{SCC}
Daily avg.	0.544	0.938	0.547	0.168	0.808	0.321
Std. dev.	0.048	0.025	0.021	0.014	0.040	0.018
All days	0.906	0.994	0.825	0.576	0.973	0.553



(a) $Complete$ $SE_\gamma = 0.022$, $R^2 = 0.82$

(b) $Sample_{Gardenhose}$ $SE_\gamma = 0.050$, $R^2 = 0.81$

(c) $Sample_{Spritzer}$ $SE_\gamma = 0.104$, $R^2 = 0.81$

Fig. 7. Distribution of the frequency of users being mentioned based on the tweets of the 1-month period. γ is the estimated power-law exponent. SE_γ is the standard error of γ. R^2 is the square error of the power-law fitting.

One day of the 1% Spritzer sample contains tweets that mention the users who are responsible for about 16% of the mentions in a day of the complete dataset (Table XI). One day of the 10% Gardenhose sample contains tweets that mention the users who are responsible for about 55% of the mentions in the complete dataset. The PCC scores based on the daily samples are very high, that is, 0.8 and 0.93 for the Spritzer and Gardenhose samples, respectively, which indicates that the mention frequency of users in the sample datasets is strongly correlated with the mention frequency in the complete dataset. If a user is mentioned in the sample data, the frequency information is relatively reliable. However, many of the users mentioned frequently in the complete dataset are not observed in a 1-day sample.

Extending the sampling period to 1 month greatly improves the results. The CTF ratios are greatly improved, especially for the Gardenhose sample (i.e., over 0.9). We also find that the PCC score and the SCC score based on the extended sampling period increase to 0.99 and 0.82, respectively, for the Gardenhose sample. Therefore, by extending the sampling period, the Gardenhose sample successfully captures most of the popular users and accurately predicts the users' relative popularities in terms of the frequency of being mentioned. Even though the CTF ratio of the Spritzer sample is also improved to 0.57 by extending the sampling period, it is still at the risk of missing many important users. Therefore, it is preferable to use the Gardenhose sample with extended sampling period for studying the users' role of popularity.

We also analyzed the user popularity distribution based on the frequency of being mentioned using the data of 1-month period. The results are displayed in Figure 7. We find that the user popularity distribution in the complete dataset approximates a power-law distribution with the exponent of –1.5977. The user popularity distributions of the sample datasets preserve the power-law property but with smaller exponents; thus, they overestimate the proportions of the less popular users. Again, we find that the user popularity distribution in the Gardenhose sample is more similar to the original distribution comparing with that in the Spritzer sample.

We performed exactly the same analysis on the retweet interactions. Despite variations in the numbers and details, the results show similar trends as the results based

on the mention interactions. All the observations made with the mention interactions in this section apply to the retweet interactions. The detailed results are provided in the Appendix.

5. CONCLUSIONS

This article provides a descriptive study of Twitter data samples obtained from the Twitter stream API with two different access priorities (i.e., Spritzer and Gardenhose). These two data streams are data sources for a variety of research and commercial activities. By comparing the sample data with the corresponding complete dataset from different perspectives, we explore the nature of the sample data, its biases, and how well it represents the complete data stream. Our results provide insights about the sample data obtained from the Twitter stream API and provide incentives for people to use or not to use them for their research.

We find that the Twitter stream API with the Spritzer and Gardenhose access priorities provides samples of the entire public tweets with actual sampling ratios around 0.95% and 9.6%, respectively. The sample datasets truthfully reflect the daily and hourly activity patterns of the Twitter users in the complete dataset. Moreover, the sample datasets capture the approximate power-law property of the user tweeting frequency distribution in the complete dataset but with smaller exponents. In other words, the sample datasets preserve the same scaling behavior of the user tweeting frequency distribution with the complete dataset, but tend to overestimate the proportions of low-frequency users. The overestimation is more serious when the sampling ratio is small. These observations indicate that the sample datasets, even with very small sampling ratios such as the Spritzer stream (i.e., 0.95%), are good for studying Twitter user activity patterns in general. However, researchers should be careful about the overestimation of the low-frequency users when trying to analyze users based on their activity levels (i.e., tweeting frequencies), and if possible, use a larger sample (e.g., Gardenhose) to reduce the estimation error.

Even with a very small sampling ratio (i.e., 0.95%), the sample datasets are able to capture certain important tweet contents (e.g., text terms and URL domains) and preserve the relative importance (i.e., frequency of appearance) of the content terms. Our work supports the viability of using sample datasets for research that analyzes tweet contents for tasks such as event detection, sentiment analysis, and tweet summarization. For some other types of content (e.g., hashtags), the small Spritzer sample is not adequate to preserve accurate information, and a larger Gardenhose sample is needed. However, for the content entities like URLs, of which the appearances in the tweets are temporal (e.g., only appears once or a few times), the importance of the terms is not reinforced by the recurrence. In this case, the sample datasets may only capture small portions of the data and may miss lots of crucial information.

In terms of the coverage of users, the sample datasets provide good coverage of active users but lower coverage of low-frequency users, as one might expect. We find that extending the sampling period or increasing the sampling ratio both help to improve the user coverage. By carefully examining the users that are difficult to sample, we find that the majority of them are extremely inactive with very low average daily tweeting frequency (e.g., post less than 1 tweet a day). A small proportion of unsampled users are highly active, but probably spammers that Twitter deliberatly excludes. For the tasks of studying the general Twitter user base, these two types of users are the least interesting because the extremely low-activity users hardly contribute anything to Twitter, and the spammers most likely only generate noise information. Therefore, the Twitter stream API can be used for collecting representative Twitter users.

Finally, we find that due to the low sampling ratios on tweets, the sample datasets cover only small proportions of the user interactions (i.e., mentions and retweets)

embedded in the tweets. For example, in the best case, the Gardenhose sample captures around 28% and 34% of the mention and retweet relationships between the captured users in a 1-month sample of data. The sample datasets do not provide a complete view of the user interaction network, thus are unsuitable for tasks that study the user network properties and information diffusion. However, for the tasks which study the users' popularities based on their frequencies of being mentioned or retweeted, the Gardenhose sample for a 1-month period provides relatively accurate information.

In general, the Twitter data samples obtained via the Twitter stream API preserve enough information for the research or applications conducted based on the general tweet or content statistics, such as user activity pattern characterization, event detection, sentiment analysis, and tweet summarization. They may also be useful for analyzing the Twitter user base and users' popularity. However, they do not provide the complete view of the user interaction network for tasks such as user network analysis and information diffusion modeling.

Although our results provide new information about the quality of Twitter data streams, they are limited by the scope of the datasets, which were collected based on a set of Singapore Twitter users. Even though our work focuses on general patterns and metrics that are not population specific, analysis of a different user population might lead to different conclusions. We believe that our observations about the Spritzer and Gardenhose samples will apply to other populations; however, this remains an open question.

We notice that the Spritze and Gardenhose samples have many characteristics that are similar to what people could expect from random samples (e.g., user activity pattern, retweet ratio, and tweeting frequency distribution). However, we could not conclude that these samples are truly random samples because we have observed that the public tweets from certain active users are excluded from the samples presumably due to their suspected spam behavior. We think it is an interesting problem for the future works to compare the Twitter Spritzer and Gardenhose samples with some truly random samples.

We also note that the sample datasets obtained from the Twitter Stream API are sampled sets of tweets. Although we can extract the user IDs and interaction information from tweet data, tweets do not contain the "follower" and "friend" relationship information declared by users, which is important metadata. It is an interesting open question whether these relationships can be inferred from tweets and the user interaction information that is available in sample data streams.

APPENDIX

This appendix provides the detailed results from analyzing retweet interactions in terms of the proportions of reciprocal and directed interactions, the Recall of interactions, and user popularity. Observations can be made from these data that are similar to observations made in Section 4.4 when studying user mention interactions.

Table XII. The Proportions of the Reciprocal and Directed User Retweet Interactions Extracted from the Complete and Sample Datasets Daily and for All the Studied Days (i.e., 1 month)

Daily statistic	*Complete*		*Sample*$_{Gardenhose}$		*Sample*$_{Spritzer}$	
	reciprocal	directed	reciprocal	directed	reciprocal	directed
Daily avg.	16.71%	83.29%	6.57%	93.42%	1.42%	98.58%
Std. dev.	0.46%	0.46%	0.99%	0.99%	0.41%	0.41%
All days	18.28%	81.72%	8.17%	88.75%	4.32%	95.68%

Table XIII. The Recall of User Retweet Interactions Daily and for All the Studied Days (i.e., 1 month)

(a) Recall on all users

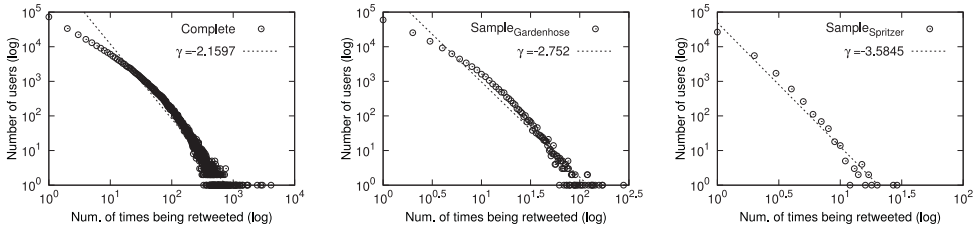
Daily statistic	$Sample_{Gardenhose}$			$Sample_{Spritzer}$		
	$Recall_{Dir.}$	$Recall_{Rec.}$	$Recall_{All}$	$Recall_{Dir.}$	$Recall_{Rec.}$	$Recall_{All}$
Daily avg.	0.229	0.081	0.204	0.028	0.002	0.023
Std. dev.	0.021	0.021	0.021	0.001	0.001	0.001
All days	0.332	0.188	0.306	0.059	0.012	0.050

(b) Recall on captured users

Daily statistic	$Sample_{Gardenhose}$			$Sample_{Spritzer}$		
	$Recall_{Dir.}$	$Recall_{Rec.}$	$Recall_{All}$	$Recall_{Dir.}$	$Recall_{Rec.}$	$Recall_{All}$
Daily avg.	0.271	0.110	0.247	0.088	0.025	0.084
Std. dev.	0.023	0.026	0.023	0.005	0.009	0.006
All days	0.381	0.204	0.342	0.138	0.061	0.131

Table XIV. Estimation of User Popularity Based on the Frequency of Being Retweeted Using the Data Daily and for All the Studied Days (i.e., 1 month)

Daily statistic	$Sample_{Gardenhose}$			$Sample_{Spritzer}$		
	S_{ctf}	S_{PCC}	S_{SCC}	S_{ctf}	S_{PCC}	S_{SCC}
Daily avg.	0.458	0.679	0.489	0.086	0.248	0.193
Std. dev.	0.049	0.038	0.022	0.005	0.072	0.033
All days	0.879	0.942	0.799	0.453	0.703	0.484



(a) $Complete$
$SE_\gamma = 0.024$, $R^2 = 0.93$

(b) $Sample_{Gardenhose}$
$SE_\gamma = 0.065$, $R^2 = 0.94$

(c) $Sample_{Spritzer}$
$SE_\gamma = 0.191$, $R^2 = 0.95$

Fig. 8. Distribution of the frequency of users being retweeted based on the tweets of the 1-month period. γ is the estimated power-law exponent. SE_γ is the standard error of γ. R^2 is the square error of the power-law fitting.

REFERENCES

Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. 2007. Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th International Conference on World Wide Web (WWW'07)*. ACM, New York, NY, 835–844.

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an influencer: Quantifying influence on Twitter. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*. ACM, New York, NY, 65–74.

Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. 2009. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference (IMC'09)*. ACM, New York, NY, 49–62.

Shea Bennett. 2012. Twitter Now Seeing 400 Million Tweets per Day, Increased Mobile Ad Revenue, Says CEO@ONLINE. Retrieved from http://www.mediabistro.com/alltwitter/twitter-400-million-tweets_b23744.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computer Science* 2, 1, 1–8.

S. Borgatti, K. Carley, and D. Krackhardt. 2006. On the robustness of centrality measures under conditions of imperfect data. *Social Networks* 28, 2 (May 2006), 124–136.

Jamie Callan and Margaret Connell. 2001. Query-based sampling of text databases. *ACM Transactions on Information Systems* 19, 2 (April 2001), 97–130.

Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. 2007. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC'07)*. ACM, New York, NY, 1–14.

Munmun De Choudhury, Scott Counts, and Mary Czerwinski. 2011. Find me the right content! Diversity-based sampling of social media spaces for topic-centric search. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM'11)*. The AAAI Press.

E. Costenbader. 2003. The stability of centrality measures when networks are sampled. *Social Networks* 25, 4 (Oct. 2003), 283–307.

Saptarshi Ghosh, Muhammad Bilal Zafar, Parantapa Bhattacharya, Naveen Sharma, Niloy Ganguly, and Krishna Gummadi. 2013. On sampling the wisdom of crowds: Random vs. expert sampling of the Twitter stream. In *Proceedings of the 22Nd ACM International Conference on Conference on Information & Knowledge Management (CIKM'13)*. ACM, New York, NY, 1739–1744.

Phillipa Gill, Martin Arlitt, Zongpeng Li, and Anirban Mahanti. 2007. YouTube traffic characterization: A view from the edge. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC'07)*. ACM, New York, NY, 15–28.

Mark Granovetter. 1976. Network sampling: Some first steps. *The American Journal of Sociology* 81, 6, 1287–1303.

H. S. Heaps. 1978. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., Orlando, FL.

Liangjie Hong, Amr Ahmed, Siva Gurumurthy, Alexander J. Smola, and Kostas Tsioutsiouliklis. 2012. Discovering geographical topics in the Twitter stream. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*. ACM, New York, NY, 769–778.

Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. 2009. Social Networks that matter: Twitter under the microscope. *First Monday* 14, 1.

Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we Twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD'07)*. ACM, New York, NY, 56–65.

Gueorgi Kossinets. 2003. Effects of missing data in social networks. *Social Networks* 28, 247–268.

Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. 2008. A few chirps about Twitter. In *Proceedings of the 1st Workshop on Online Social Networks (WOSN'08)*. ACM, New York, NY, 19–24.

Ravi Kumar, Jasmine Novak, and Andrew Tomkins. 2006. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. ACM, New York, NY, 611–617.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, New York, NY, 591–600.

SangHoon Lee, Pan-Jun Kim, Hawoong Jeong, and Fang Wu. 2006. Statistical properties of sampled networks. *Physical Review E* 73, 1.

Jure Leskovec and Christos Faloutsos. 2006. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*. ACM, New York, NY, 631–636.

Arun S. Maiya and Tanya Y. Berger-Wolf. 2011. Benefits of bias: Towards better characterization of network sampling. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11)*. ACM, New York, NY, 105–113.

Michael Mathioudakis and Nick Koudas. 2010. TwitterMonitor: Trend detection over the Twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data (SIGMOD'10)*. ACM, New York, NY, 1155–1158.

Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC'07)*. ACM, New York, NY, 29–42.

Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter Firehose. In *Proceedings of the 7th International Conference on Weblog fs and Social Media (ICWSM'13)*. The AAAI Press.

Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me?: Message content in social awareness streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW'10)*. ACM, New York, NY, 189–192.

B. Ohana and B. Tierney. 2009. Sentiment classification of reviews using SentiWordNet. In *Proceedings of the Ninth IT&T Conference*. 13.

Alexei Pozdnoukhov and Christian Kaiser. 2011. Space-time dynamics of topics in streaming text. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks (LBSN'11)*. ACM, New York, NY, 1–8.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10)*. ACM, New York, NY, 851–860.

Semiocast. 2012. Twitter Reaches Half a Billion Accounts More Than 140 Millions in the U.S. @ONLINE. Retrieved from http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US.

Mustafa Sofean and Matthew Smith. 2012. A real-time architecture for detection of diseases using social networks: design, implementation and evaluation. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT'12)*. ACM, New York, NY, 309–310.

W. Cohen William and Gosling Samuel. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media. In *Proceedings of the 4th International Conference on Weblogs and Social Media (ICWSM'10)*. The AAAI Press.

Sooyeon Yoon, Sungmin Lee, Soon-Hyung Yook, and Yup Kin. 2006. Statistical properties of sampled networks by random walk. *Physical Review E* 73, 1.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and traditional media using topic models. In *Proceedings of the 33rd European Conference on Advances in Information Retrieval (ECIR'11)*. Springer-Verlag, Berlin, 338–349.

Received July 2013; revised December 2014; accepted March 2015