Sampling Content from Online Social Networks: Comparing Random vs. Expert Sampling of the Twitter Stream

MUHAMMAD BILAL ZAFAR, Max Planck Institute for Software Systems, Germany
PARANTAPA BHATTACHARYA, Indian Institute of Technology Kharagpur, India; Max Planck Institute for Software Systems, Germany
NILOY GANGULY, Indian Institute of Technology Kharagpur, India
KRISHNA P. GUMMADI, Max Planck Institute for Software Systems, Germany
SAPTARSHI GHOSH, Max Planck Institute for Software Systems, Germany; Indian Institute of Engineering Science and Technology Shibpur, India

Analysis of content streams gathered from social networking sites such as Twitter has several applications ranging from content search and recommendation, news detection to business analytics. However, processing large amounts of data generated on these sites in real-time poses a difficult challenge. To cope with the data deluge, analytics companies and researchers are increasingly resorting to sampling. In this article, we investigate the crucial question of *how to sample content streams generated by users in online social networks*. The traditional method is to randomly sample all the data. For example, most studies using Twitter data today rely on the 1% and 10% randomly sampled streams of tweets that are provided by Twitter. In this paper, we analyze a different sampling methodology, one where content is gathered only from a relatively small sample (<1%) of the user population, namely, the *expert users*. Over the duration of a month, we gathered tweets from over 500,000 Twitter users who are identified as experts on a diverse set of topics, and compared the resulting expert sampled tweets with the 1% randomly sampled tweets provided publicly by Twitter. We compared the sampled datasets along several dimensions, including the popularity, topical diversity, trustworthiness, and timeliness of the information contained within them, and on the sentiment/opinion expressed on specific topics. Our analysis reveals several important differences in data obtained through the different sampling methodologies, which have serious implications for applications such as topical search, trustworthy content recommendations, breaking news detection, and opinion mining.

Categories and Subject Descriptors: H.3.5 [**On-line Information Services**]: Web-based Services; J.4 [**Computer Applications**]: Social and Behavioral Sciences; H.1.2 [**User/Machine Systems**]: Human Information Processing

General Terms: Experimentation, Human Factors, Measurement

Additional Key Words and Phrases: Sampling content streams, Twitter, random sampling, sampling from experts, Twitter Lists

---

This research was partially supported by grants from the Indo-German Max Planck Centre for Computer Science (IMPECS), and the Information Technology Research Academy (ITRA), DeITY, Government of India (Ref. No.: ITRA/15 (58) /Mobile/DISARM/05). Additionally, P. Bhattacharya was supported by a fellowship from Tata Consultancy Services, and S. Ghosh was supported by a postdoctoral fellowship from the Alexander von Humboldt Foundation.
This work is an extended version of the paper: Ghosh *et al.*, On Sampling the Wisdom of Crowds: Random vs. Expert Sampling of the Twitter Stream, Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM), pp. 1739–1744, 2013.
Authors' addresses: M. B. Zafar, K. P. Gummadi, and S. Ghosh, Max Planck Institute for Software Systems, Saarbruecken 66123, Germany; P. Bhattacharya and N. Ganguly, Department of Computer Science and Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India.

DOI: http://dx.doi.org/10.1145/2743023

ACM Reference Format:
Muhammad Bilal Zafar, Parantapa Bhattacharya, Niloy Ganguly, Krishna P. Gummadi, and Saptarshi Ghosh. 2015. Sampling content from online social networks: Comparing random vs. expert sampling of the twitter stream. ACM Trans. Web 9, 3, Article 12 (June 2015), 33 pages.
DOI: http://dx.doi.org/10.1145/2743023

1. INTRODUCTION

Online Social Networks (OSNs), such as Twitter and Facebook, are currently used by hundreds of millions of users, not only to communicate with friends, but also to find and share recent content on various topics of interest. The user-generated content in these social networking sites is increasingly being used for a variety of data analytics applications ranging from content search and recommendations [Teevan et al. 2011] to opinion mining [Hannak et al. 2012; Tumasjan et al. 2010]. Especially, the Twitter microblogging site (https://twitter.com/) has emerged as a popular platform for discovering real-time information on the Web, such as current events, news stories, and people's opinion about them [Kwak et al. 2010]. Recent studies on Twitter have shown that crowd-generated tweet-streams could be analyzed to gather public feedback and reaction to major events, be they movie releases [Asur and Huberman 2010], or political elections [Tumasjan et al. 2010], or natural emergencies [Sakaki et al. 2010].

However, Twitter has millions of users tweeting hundreds of millions of tweets every day [Twitter-stats 2014], and the sheer volume of the entire tweet stream (known as the *firehose*) presents an enormous logistic problem for data analysts. In addition to the huge cost required for accessing Twitter's *firehose*, it requires an enormous amount of resources (e.g., network bandwidth, storage space) to even gather and store the tweets gushing from Twitter's *firehose*, let alone processing them. So analytics companies and researchers often rely on sampled data rather than the entire dataset. Against this background, this paper investigates the following key question: *What is the most effective way to sample the data streams generated by users in OSNs?* Note that though the current work focuses on Twitter, this question is vital for sampling content from many other online media as well.

Today, most data analytics companies and researchers rely on *randomly sampled* tweet-streams. Twitter supplies 10% randomly sampled tweets (known as the *gardenhose*) from its *firehose* for a fee, and 1% randomly sampled tweets (originally known as *Spritzer*) for free. Note that the exact details of how these samples are selected are not made public by Twitter. However, the Twitter API documentation [Twitter-stream-api 2012] as well as several public blogs (e.g., spritzer-gnip-blog [2011]) state that these samples are random samples of all public tweets. Hence, we refer to these samples as *random samples*.

Random sampling is appealing for data analytics as the sampled tweets preserve a number of statistical properties of the global set of tweets, such as the fraction of tweets that are related to a given topic. However, random sampling also preserves the large amount of unwanted content in the original tweets, such as spam and *nontopical* conversational tweets. A growing number of content-centric applications like topical content search or breaking news detection, can benefit from a sampling methodology that filters out the unnecessary tweets and selectively captures tweets with the most important or interesting information, even if the sampled tweets were not representative of the global tweet population. In this article, we propose and analyze one such sampling methodology.

In contrast to random sampling, we gather content only from *topical experts*, that is, Twitter users whose followers consider them to be knowledgeable on some topic. The topical experts are often the primary drivers of interesting discussions on Twitter [Ardon et al. 2013]. Hence, our intuition is that by focusing on tweets from experts

on a wide variety of topics, we might be able to cut down the unwanted tweets in the sampled data while still gathering useful tweets related to a wide range of topics. The key challenge, however, lies in identifying an extensive set of experts covering the wide range of topics that Twitter users are interested in, including popular, niche, local and global topics. This is necessary to avoid being restricted to only few experts (or few topics) and thus falling prey to the long-standing expert versus crowd debate [Fama 1970; Poses et al. 1990]. We leveraged a technique that we recently proposed [Ghosh et al. 2012a; Sharma et al. 2012] to crowd-source expert detection, and identified over half a million experts on a diverse set of topics of interest to Twitter users (details in Section 3).

In this work, we present an extensive analysis of the merits and drawbacks of sampling data from these expert users, compared to random sampling of data from the entire user population. To this end, we gathered two samples of tweets—(i) the 1% randomly sampled tweet stream provided by Twitter and (ii) tweets from over half a million experts on different topics—over the course of a month and compared the resulting datasets. We compare the information content in the samples along several different aspects, including its quality, popularity, topical diversity, trustworthiness, timeliness, and diversity of opinions/sentiment on specific issues.

Our analysis reveals that the tweets sampled from experts have several properties that make this approach more suitable for data mining/information retrieval applications, as compared to random sampling of the tweet stream. Using a combination of topic modeling [Blei et al. 2003] and human feedback (obtained through the Amazon Mechanical Turk service) on the nature and topicality of tweets, we find that the tweet sample collected from experts is not only much richer in its information content but also contains useful information on a wide variety of topics, including politics, sports, science and technology, entertainment, and so on. In contrast, a very large majority of the tweets in the Twitter random sample either contain day-to-day conversation among users, or are related to only two broad topics—entertainment and lifestyle (e.g., music, television shows, fashion, celebrity news)—which are popular among the general population. We also find that the experts' tweets are more popular in the social network, as well as more trustworthy (i.e., their tweets have much fewer malicious URLs or spam content). We uncover the presence of *sampling spam* in the Twitter random samples, that is, some users deliberately tweeting more in order to have a greater presence in the randomly sampled streams; experts' tweets are not vulnerable to such sampling spam. Furthermore, tweets sampled from experts also tend to capture breaking news stories and important events in timely manner, often a little earlier than randomly sampled tweets. Finally, we find that opinions/sentiments mined from the tweets of a large and diverse set of experts match fairly well with the opinions mined from the random sample.

Our findings make a strong case for selectively sampling data from a large and diverse set of expert users, as it yields concise yet information-rich and trustworthy digests. We conclude that expert sampling is more efficient than random sampling for content-centric applications ranging from topical search and content recommendations to breaking news detection and opinion mining. However, random sampling also has its own merits, which can be useful in specific applications such as tracking the interests of the masses.

The rest of the article is organized as follows. Section 2 reviews related work. Section 3 describes the datasets (expert and random samples), and the methodology used to collect the datasets. Section 3 also states the aspects/dimensions along which the expert and random samples are compared in the subsequent sections—the characteristics of the sources, that is, the users who post the tweets (Section 4), and characteristics of the content, for example, popularity (Section 4), topical quality and

diversity (Section 5), trustworthiness (Section 6), timeliness (Section 7), and diversity of opinion in the samples (Section 8). Finally, Section 9 concludes the article by discussing some potential applications of the insights gained in this study.

2. RELATED WORK

There have been a number of studies on sampling of large data spaces, in various domains including signal processing, image processing, networked systems, and so on. This section briefly reviews some of these studies.

Sampling of large data spaces: Sampling of information signals and images has been extensively studied in signal processing and information theory literature [Candes and Wakin 2008; Kellogg 2006; Romberg 2008]. With the advent of large networked systems, there have been several studies on sampling large networks such as the Web-graph [Rusmevichientong et al. 2001] and social networks [Frank 1978; Gjoka et al. 2010; Granovetter 1976; Leskovec and Faloutsos 2006]. A large majority of these studies have focused on sampling of *nodes* from the networks. Different node sampling strategies (such as uniform sampling, random-walk–based sampling), have been studied to ascertain whether the subgraphs obtained by these strategies can be used to recover the properties of the complete network, including topological properties such as size and degree distributions [Gjoka et al. 2010; Katzir et al. 2011; Leskovec and Faloutsos 2006] as well as dynamic processes such as information diffusion [Choudhury et al. 2010] over the networks. Specifically for the Twitter OSN, [Krishnamurthy et al. 2008] made the first attempt to compare two datasets obtained by different sampling techniques—one through snowball crawling of the Twitter network starting from a small seed set of users, and the other through continuously accessing the Twitter public timeline (which provided the 20 most recent tweets) and extracting the users who posted the tweets. Again, Choudhury et al. [2010] examined several node (user) sampling strategies on Twitter users (e.g., random sampling, sampling users based on their location or their level of activity) and compared the information diffusion patterns observed from the samples. Whereas all these studies deal with sampling of users (nodes in the social network), we, in this work, focus on sampling data/content streams generated in OSNs, for which the methodologies discussed in the aforementioned studies are not directly applicable.

Sampling of data streams in Twitter: Data streams from Twitter are presently being used for a wide variety of purposes such as topical search [Choudhury et al. 2011a; Lin et al. 2011], identifying breaking news [Sakaki et al. 2010; Sankaranarayanan et al. 2009], opinion mining [Asur and Huberman 2010; Hannak et al. 2012; Tumasjan et al. 2010], and so on. Among these, a few studies [Choudhury et al. 2011a, 2011b; Lin et al. 2011] have used the complete *Firehose* for a limited duration of time; however, the *Firehose* is available to very few organizations in the world, and even for those to whom it is available, the amount of content in the *Firehose* is too large to analyze continuously in real time. Hence, most studies, especially the ones which focus on real-time analytics, use a random or uniform sample of all tweets. Random samples of various sizes (e.g., 1%, 10%) provided by Twitter have been used, for instance, to detect current trending topics [Mathioudakis and Koudas 2010], to summarize the tweet stream to provide user-specific and topic-specific temporal analytics [Yang et al. 2012a], to map the content to various topical dimensions [Ramage et al. 2010], and so on. Hybrid sampling strategies have also been used; for example, Sankaranarayanan et al. [2009] combined the *Gardenhose* with tweets from 2,000 hand-picked news media accounts to identify breaking news from Twitter. Some studies have also used Twitter search to gather tweets containing specific keywords, for example, tweets related to earthquakes in Japan [Sakaki et al. 2010] and the Singapore General Election [Li et al. 2012].

The previous discussion shows that various approaches have been used to sample tweet streams. However, there has been very little research on the suitability of different content sampling strategies for specific applications. To our knowledge, only a few prior studies have attempted to compare content streams collected through different sampling strategies in Twitter. Gonzalez-Bailon et al. [2014] sampled tweets from two sources—the Twitter "search API" and the 1% random sample—and compared the communication networks reconstructed from the two samples. Morstatter et al. [2013] compared the Twitter Firehose with data obtained through the Twitter "streaming API," which provides a sample of tweets matching some parameters preset by the API user. More recently, Morstatter et al. [2014] developed a methodology to estimate the bias in the samples collected through the Twitter streaming API during particular intervals of time.

There are important differences between the prior studies and the present work. Whereas the prior studies compared among the various sampling strategies that are already provided by Twitter, we compare a novel expert sampling strategy with the most commonly used random sampling. Additionally, the earlier studies focused only on tweets related to a specific event (political protests in Spain in mid-2012 [Gonzalez-Bailon et al. 2014], and the Syrian conflict during 2011–2012 [Morstatter et al. 2013]), whereas we study various aspects of the sampled tweet streams to derive far more general observations. Because our objective is to analyze the utility of the samples for data mining/information retrieval applications (e.g., topical search and recommendation, breaking news detection, opinion mining), we investigate issues such as the popularity, topical diversity, trustworthiness, timeliness and sentiment analysis, many of which have never been studied by the prior works on sampling OSNs.

Sampling experts versus crowds: As mentioned earlier, the present study compares tweet streams sampled from the general Twitter crowd (the Twitter random sample), and a set of chosen experts in Twitter. There has been a long-standing debate on the expert-versus-crowd question as to which source of information is better. Crowd wisdom has been found to be superior to expert wisdom in some specific applications. For instance, in the finance domain, the famous Efficient Market Hypothesis [Fama 1970] concludes that crowd wisdom is superior to expert wisdom in picking financial stocks. Poses et al. [1990] observed that multiple inexperienced physicians may collectively make better decisions than one experienced physician. On the other hand, expert opinion has been found to be extremely valuable in *web-based applications* such as search and recommendation. For instance, Amatriain et al. [2009] designed a recommendation system for movies on Netflix based on opinions from movie critics (experts) collected from the Web. Again, ranking algorithms used by Web search-engines, such as PageRank [Brin and Page 1998], give equal or higher weightage to a link from an important page (analogous to an expert) than links from multiple less important pages. However, to the best of our knowledge, there has not been any investigation as to which source (experts or crowds) is better for sampling content streams generated in OSNs.

Furthermore, the traditional experts-versus-crowds studies compare the wisdom of a *small number* of experts with that of the crowd which constitutes a much larger sample. On the other hand, the present work compares two tweet streams of comparable size, one sampled from the general Twitter crowd (the Twitter random sample) and the other from a *large* chosen set of experts. Hence, in effect, we address a related but different question: If one is able to sample a data space of a given size from Twitter (e.g., a certain number of tweets per day), then should one sample the data from the crowds, or from a set of chosen experts?

Finally, a preliminary version of the present work has been published as a short paper [Ghosh et al. 2013]. The present work thoroughly improves and extends the analyses in Ghosh et al. [2013]. For instance, we include a detailed characterization

of the users whose tweets are included in the random and expert samples; we also report the presence of *sampling spam* in the Twitter random sample—some users tweeting excessively to increase their chances of being represented in the random sample. Furthermore, the present work includes some completely new analyses, such as topic modeling of the random and experts' streams to identify the topics being discussed, and the representativeness of opinions/sentiment on specific events in the two streams.

3. DATA SAMPLES GATHERED

The objective of the present study is to collect and compare tweets obtained through two different sampling strategies: random sampling and sampling from the experts/most popular users. In this section, we describe how we gathered the samples and compare their high-level characteristics. We also list the aspects/dimensions considered for a more extensive comparison of the samples in the subsequent sections.

We start by discussing the dataset of users collected for the study. To identify the experts/most popular users in Twitter, we started a long-running crawl of the Twitter user accounts in the order in which the accounts were created in Twitter. Under the rate limits imposed by Twitter on such crawls [twitter-rate-limit 2013], we were able to crawl data (profile details, social links, and lists) for the first 50 million Twitter users. The experts/most popular users were identified from among these 50 million users, as described in the following text.

3.1. Sampling Methodologies

We considered the following sampling strategies for collecting the tweets.

Random sample: As mentioned earlier, the most commonly used methodology of sampling tweet streams is random sampling. To obtain a random sample of tweets, we used the publicly available Twitter streaming API to gather the 1% random sample of all tweets [Twitter-stream-api 2012] during December 2012.

Tweets posted by experts/popular users: The other sampling methodology considered in this study is to sample content from the most popular users/experts in the social network. There are several metrics to rank users in Twitter according to their popularity, and hence to identify most popular users [Cha et al. 2010; Ghosh et al. 2012a; Kwak et al. 2010]. These include the number of followers of a user (follower-rank), the PageRank of a user in the social network, the number of times a user is listed (List-rank), and so on. We ranked the 50 million users for whom we could gather data, using each of the aforementioned metrics.[1] We found that there is a significant overlap among the top-ranked users according to the various metrics. For instance, considering the top 500,000 users (i.e., top 1% of 50 million) according to the various metrics, we observed 68.2% overlap between the top 500,000 users according to follower-rank and PageRank, and 68.7% overlap between the top 500,000 according to follower-rank and List-rank. Hence, the choice of the specific ranking metric is not likely to cause significant differences in the tweet sample gathered from experts (the top-ranking users).

We decided to use the List-rank metric where users are ranked according to the number of Lists in which they are included. Lists are an organizational feature in Twitter, by which a user can group together experts on specific topics [lists-howtouse 2013]. For instance, a user who is interested in music may create a List named "Music and musicians" and add popular musicians like "BritneySpears" and "LadyGaga" as

[1]The PageRank scores were computed over the network of follow-links among the 50 million users whose data we could gather.

Table I. Examples of Twitter Lists: Names, Descriptions, and Some Sample Members

List Name	Description (extracts)	Example Members
News	News media accounts	nytimes, BBCNews, WSJ, cnnbrk, CBSNews
Music	Musicians	Eminem, britneyspears, ladygaga, rihanna, BonJovi
Tennis	Tennis players and Tennis news	andyroddick, usopen, Bryanbros, ATPWorldTour
Politics	Politicians and people who talk about them	BarackObama, nprpolitics, whitehouse, billmaher, Al Gore, Rachel Maddow
Food	Love food? Chef's, cookers and other experts in food	ChefChiarello, Rick_Bayless, Paula_Deen, epicurious, LATimesfood

Table II. Examples of Topical Experts Identified by the List-Based Methodology [Ghosh et al. 2012a; Sharma et al. 2012] for Specific Topics

Topic	Experts identified by List-based method
Music	Katy Perry, Lady Gaga, Justin Timberlake, coldplay, P!nk, Marshall Mathers
Politics	Barack Obama, Al Gore, Bill Maher, NPR Politics, Sarah Palin, John McCain
Environment	GreenPeace USA, NYTimes Environment, TreeHugger.com, National Wildlife
Medicine	NEJM, Nature Medicine, American Medical News, FDA Drug Information
Physics	Institute of Physics, Physics World, Fermilab Today, CERN, astroparticle
Neurology	Neurology Today, AAN Public, Neurology Journal, Oliver Sacks, ArchNeurology
Anthropology	Leakey Foundation, AmericanAnthro, anthropologyworks, Popular Anthropology

The experts are indicated by the real names or the screen names as given in the profile of their Twitter accounts.

members of the List. Table I shows some examples of Lists, giving the List names, descriptions and some example members included in the Lists.

We preferred the List-rank metric to the more commonly used follower-rank metric because our prior studies observed that follow-links in Twitter can easily be farmed by spammers/link farmers [Ghosh et al. 2012b]; hence, having a high follower-rank does not necessarily imply that the user is a genuine expert. On the other hand, because Lists have a specific name and description, List-memberships are much more difficult to farm as compared to follow-links. Furthermore, Lists have an additional advantage: List names and other metadata can be used to infer the topical expertise of the members of the Lists, as shown in our prior studies [Ghosh et al. 2012a; Sharma et al. 2012] as well as in other recent studies [Wagner et al. 2012; Wu et al. 2011].

Similar to Ghosh et al. [2012a], we consider a Twitter user as a "topical expert" if and only if the user has been listed at least 10 times on some particular topic. Out of the 50 million users whose data we could collect, 584,759 users were listed at least 10 times on some specific topic; hence, we consider these 584,759 users as our sample set of experts. Table II shows some example topics and Twitter users identified as experts on the topic, using the List-based methodology. We collected all tweets posted by the 584,759 experts over the month of December 2012.

Note that a vital strength of the List-based methodology (which relies on crowd-sourced social annotations) is that it not only identifies hundreds of thousands of experts but also the set of experts cover a large and diverse set of topics, ranging from very popular topics such as music and politics to niche, specialized topics such as physics, neurology, and anthropology (see Table II).[2] This is essential to ensure that information on a large number of topics/events is available in the experts' tweets.

[2]Interested readers are referred to our prior work [Bhattacharya et al. 2014] for a detailed discussion on the diversity of the topics covered in the set of experts.

Table III. Number of Tweets and Distinct Users Who Tweeted in the Three Digests—the Random 1% Digest, the Expert Digest, and a Subsampled Random 1% Digest Containing the Same Number of Tweets as the Expert Digest–across Three Different Durations—a Day, a Week, and a Month

Sample of tweets	Day (Dec 3, 2012)		Week (Dec 3–9, 2012)		Month (Dec 2012)	
	# Tweets	# Users	# Tweets	# Users	# Tweets	# Users
Random 1% digest	4,051,763	3,145,879	27,410,736	13,050,061	124,253,878	30,046,582
Expert digest	2,264,904	260,339	15,517,042	378,180	63,497,081	427,674
Subsampled random digest	2,264,904	1,930,045	15,517,042	9,105,185	63,497,081	21,941,041

To further investigate the effects of the choice of the metric to identify experts, we also collected the tweets posted by the top 100,000 users according to the follower-rank metric, during the month of December 2012. We compared the tweet-sample comprising of the tweets posted by these 100,000 top-followed users, with a similarly sized tweet sample consisting of tweets posted by the most listed 584,759 experts.[3] We found that the two samples have very similar properties along all dimensions that are discussed later in the article. This further shows that the choice of the specific metric to identify experts does not result in significant differences in the tweet samples collected from the experts.

Note that like any other method for identifying experts, the aforementioned methodology also has a few limitations. The primary one is that the set of experts identified by us is limited to users who joined Twitter early. Hence, our set of experts is biased heavily toward certain countries (e.g., the United States and some English-speaking countries) where Twitter became popular earlier than in other countries. However, our objective here is *not* to identify all experts or obtain an unbiased sample of experts in Twitter. Rather we wish to study the differences between randomly sampled tweets and those obtained from a (any) large set of experts. We believe that the set of experts we identified is sufficient for the purpose of our study.

3.2. High-Level Sample Characteristics

We gathered both 1% randomly sampled tweets and the tweets posted by the experts during the entire month of December 2012. We refer to the resulting tweet samples as the *random digest* and the *expert digest,* respectively.

We compared the two tweet digests sampled over three time durations: a day (December 3, 2012), a week (December 3–9, 2012), and the entire month of December 2012. Table III gives the number of tweets and the number of distinct users who posted the tweets in both the digests during the three durations. In each of the three durations, the random digest has about 1.8 to 2 times as many tweets as the expert digest. This implies that our expert digest contains about 0.55% of all tweets posted on Twitter. Table III also shows that out of the 584,759 experts whom we identified from among the first 50 million Twitter user accounts (as detailed earlier), only 427,674 posted a tweet during the entire month of December 2012. Hence, in essence, this study compares the Twitter random sample with the content sampled from these 427,674 experts.

It can be noted that the goal of the present study is to compare the two sampling strategies (random sampling versus expert sampling), rather than to compare the exact set of tweets posted by our chosen set of experts with the Twitter 1% random sample. To enable a fair comparison between the two sampling strategies, it is necessary to consider *equal-sized* sets of tweets obtained using the strategies. Hence, we (randomly) subsampled the Twitter 1% random digest to contain the same number of tweets as

[3]Note that 84.4% of the 100,000 top followed users are already included among the 584,759 experts identified using Lists.

the expert digest over each of the three durations; the statistics for the subsampled random digest are also shown in Table III. In the rest of the article, we always compare the expert digest with the similarly sized random digest.

3.3. Dimensions Along Which to Compare Data Samples

The goal of this study is to compare the tweet digests obtained through the two different sampling strategies, to ascertain their utility for applications such as topical search and recommendation, breaking news detection, sentiment/opinion mining, and so on. Hence, we consider the following dimensions for comparing the tweet samples:

(1) **Sources:** Do the samples contain information posted by very different types of users? Also, does the expert sample contain the posts of only the elite users, or also the voices of the crowd?
(2) **Popularity:** How popular are the tweets contained in the samples, in the global Twitter network?
(3) **Quality:** Do the samples contain mostly conversational babble or useful information?
(4) **Topical diversity:** What are the various topics which are covered in the samples?
(5) **Trustworthiness:** What is the amount of spam (e.g., blacklisted URLs) included in the samples?[4]
(6) **Timeliness:** How quickly is the information of a recent, unforeseen event (e.g., an accident or natural calamity) available in the samples?
(7) **Sentiment/opinion:** Do the samples express similar sentiment on specific issues, such as recently released movies and sociopolitical issues?

4. SOURCES AND POPULARITY OF CONTENT

We start by comparing the sources of information in the expert and random digests (i.e., the users whose tweets are included in the two digests). Specifically, we ask two things: (i) How similar are the users whose tweets are included in the two digests?; and (ii) Does the expert digest reflect the views of only the experts, or does it also succeed in capturing interesting information tweeted by the Twitter crowds? We also study the popularity of the tweets included in the samples in the global Twitter network.

4.1. Comparing Sources of Tweets

Comparing the number of users in the two digests over the three different durations shown in Table III, it is clear that the random digest includes tweets from a substantially larger (by one to two orders of magnitude) population of Twitter users than the expert digest which is limited to tweets from a relatively small set of expert users. However, although the expert digest includes all tweets posted by the selected experts during a certain time period (a day/week/month), the random sample is likely to include only a small fraction of the tweets posted by a given user. This is demonstrated in Figure 1, which plots the distribution of the number of times a given user is sampled in the two digests during the month of December 2012. It is seen that for more than 90% of the users included in the random digest, less than 10 tweets are included per user.

Computing the overlap between the set of users and the set of tweets included in the two digests, we found that about 35% of the users in the expert digest are included in the corresponding random digest over the duration of one month (i.e., at least one tweet from these experts has been included in the random digest), whereas only 0.15%

[4]Note that there are multiple aspects to "trustworthiness" of content, such as presence or absence of rumors or false information. However, in this study, we limit ourselves to investigating the presence of spam.

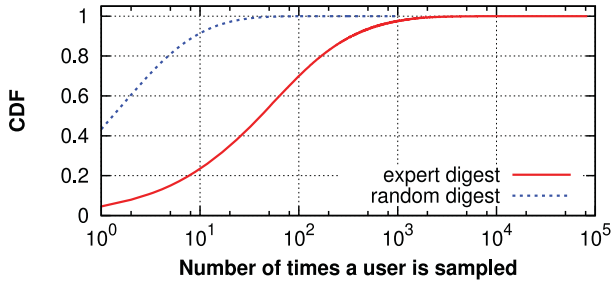
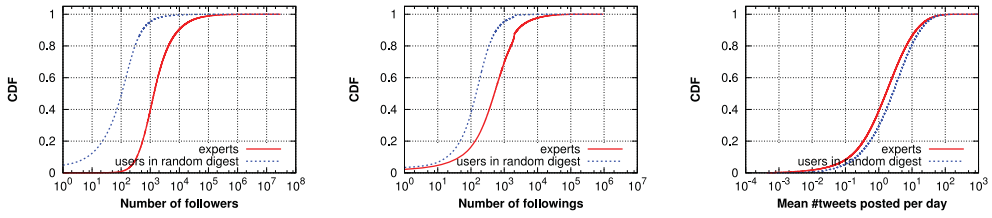


Fig. 1. Distribution of the number of times a particular user's tweets are included in the two digests. Most users are sampled only a few times in the random digest, whereas the expert digest contains all tweets posted by the experts.



(a) Number of followers (b) Number of followings (c) Mean number of tweets posted per day

Fig. 2. Comparing the users whose tweets are included in the random and expert digests—(a) number of followers, (b) number of followings, and (c) mean number of tweets posted per day.

of the tweets are common between the two digests. In other words, even though a sizable fraction of the experts gets included in the random sample over a relatively long duration of time, such instances occur only intermittently.

Next, we compare the characteristics of the users encountered in the Twitter random and expert digests. Figures 2(a) through 2(c), respectively, show the distributions of the number of followers, number of followings, and the mean (average) number of tweets posted per day, for the two sets of users.[5] The plots clearly show that the expert and random digests draw their tweets from very different Twitter user populations; compared to users in the random digest, experts have many times more followers and followings and are considerably more popular within the Twitter social network. However, both sets of users post similar number of tweets per day on average, as seen from the closely matching plots in Figure 2(c). Note that though the users in the two digests generally show similar activity in terms of posting tweets, the random digest includes thousands of tweets from a few users who tweet excessively in order to increase their visibility in the Twitter random sample; these users are described in Section 6.

4.2. Comparing Original Sources of Retweets

The differences between the user populations in the expert and random digests (as reported earlier) are not surprising. However, these differences raise a potential concern that by sampling tweets only from experts, one might miss useful and interesting information that is tweeted by the masses of nonexpert users in Twitter. To check if this concern holds, we analyzed the retweets in the expert digest to see whether experts

[5]The mean number of tweets posted per day by a user can be computed from the profile information of a user, which contains the total number of tweets posted by her to date, as well as the date on which the account was created.

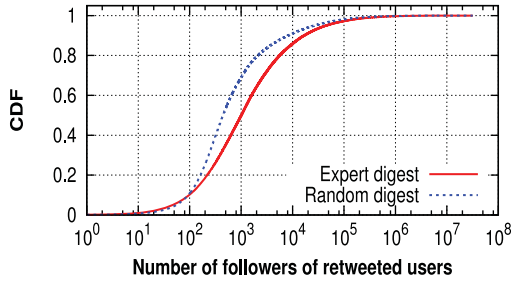


Fig. 3. Distribution of the number of followers of users who are retweeted in each digest. The experts frequently retweet tweets that are originally posted by nonexperts having low number of followers.

retweet (repost or forward) only tweets from other experts, or whether they retweet interesting information posted by ordinary users as well.

We observe that both digests contain nearly the same fraction of retweets—18% of the tweets in the expert digest and 21.6% of the tweets in the random digest are retweets. This suggests that experts are also forwarding tweets from other users at the same rate as users in the random digest. Next, we investigate whether experts only retweet tweets posted by other experts. Figure 3 shows the distribution of popularity (measured as number of followers) of those users whose tweets were retweeted by experts. For comparison, we also show the distribution of popularity of those users whose tweets were retweeted by users included in the random digest. We find that the popularity distributions for the retweeted users in the two digests (Figure 3) are much more similar than the popularity distributions of all users whose tweets are included in the two digests (Figure 2(a)). Put differently, though the experts are far more popular than the average user in the random sample, the experts are not limiting their retweets to their fellow experts. They retweet interesting information from both expert and nonexpert users, just as a random Twitter user would retweet. In fact, as much as 73% of the retweets in the expert digest were originally posted by nonexperts.

The aforementioned observation about retweeting behavior of experts has important implications: It suggests that concerns about expert digest not including interesting tweets from the Twitter crowds might be unfounded. In fact, our data suggests that experts themselves are interacting with nonexpert crowds within Twitter, and they are filtering/forwarding information from the crowds that they deem interesting and useful.

We give an interesting example of this phenomenon. When the US-Mexican singer Jenni Rivera died in an air crash on December 9, 2012, the news was first posted in Twitter by the user @IvanSamuelLuz, a *nonexpert* having only 180 followers. This tweet was retweeted by the user @Rodpac, an *expert* having 23,000 followers. The post by Rodpac got retweeted several times by his large number of followers, and one of these retweets subsequently appeared in the Twitter random sample (see Section 7 for details and other examples). Thus, important information posted by nonexperts are picked up by experts and captured in the expert digest.

4.3. Global Popularity of the Digests

Now we compare the *global popularity* of the tweets contained in the expert digest and the random digest. We estimate the global popularity of a given tweet by the number of times it has been retweeted in the global Twitter network, which is also known as the retweet count of the tweet. We obtained the retweet counts for the original tweets (i.e., tweets which are not retweets themselves) included in the two digests, using the

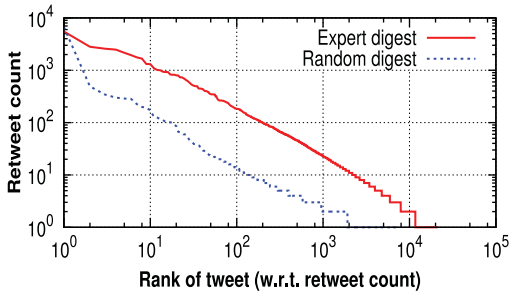


Fig. 4. Distribution of the number of times an original tweet (i.e., not a retweet) in the two digests is retweeted globally in Twitter. The original tweets in the expert digest are retweeted many more times than those in the random digest.

public Twitter API.[6] The popularity (retweet count) distributions of the original tweets in the expert and random digests are shown in Figure 4. Here, the y-axis gives the retweet count, and the x-axis gives the rank of the tweets when ordered in decreasing order of the retweet count. The plot clearly shows that the tweets authored by the experts have an order of magnitude higher retweet counts than the original tweets in the random digest. An original tweet in the expert digest is retweeted 2.18 times, on average, whereas an original tweet in the random digest is retweeted only 0.33 times. Thus, compared to random sampling, expert sampling is biased toward more popular and influential tweets, which can be expected given the much higher popularity of the experts compared to random users (as reported earlier in this section).

Summary: We find that the expert digest and the random digest include information posted by very different users. As expected, the experts are much more popular than the users whose tweets are included in the random sample. As a result of the higher popularity of the experts, the tweets in the expert digest are also more popular (retweeted) in the social network. However, we also see that experts selectively forward (retweet) information that was originally posted by common users; hence, the expert digest also includes the views of the masses.

5. QUALITY AND TOPICAL DIVERSITY OF CONTENT

In this section, we compare the quality and topical diversity of the information contained in the expert digest and the random digest. Our analysis is driven by the following questions: (i) How much useful information is contained in the two digests?, and (ii) how diverse are the topics covered by the tweets in two digests?

We compared the information content in expert and random digests gathered over three different durations: a day, a week, and a month (as stated in Section 3). We observed very similar trends in results obtained over each of the three durations. Hence, in this section, we only report results obtained from comparing digests gathered over the duration of a day (specifically, December 3, 2012).

5.1. Comparing Content Quality in the Digests

We start by studying what fraction of tweets in the two digests contain useful topical information. Since judging whether a tweet contains useful information on some topic is inherently subjective, we judged the quality of information in the tweets through human feedback. To gather feedback from a large and varied set of people, we used the

[6]Because of Twitter API rate limits, we could query the retweet counts of only 100,000 randomly selected original tweets in both digests.

Table IV. Categories of Tweets, as Judged by Evaluators in a Survey Conducted through Amazon Mechanical Turk

Category	Random digest	Expert digest
Conversational/Babble	82.11% (11.5%)	40.64% (22.4%)
Opinion/Sentiment (without any topical information)	8.42% (37.5%)	10.16% (31.6%)
Advertisement/Sales promotion	1.05% (100%)	5.88% (90.9%)
Spam	1.58% (0.0%)	0% (0.0%)
Information on a specific topic	6.84% (53.9%)	43.32% (90.12%)

The first number in each entry shows the percentage of tweets (in the two digests) that were judged to be of the corresponding category. The second number (within parentheses) shows what percentage of the tweets in a particular category contained URLs or hashtags.

Amazon Mechanical Turk (AMT) service where human evaluators judged the nature/topic of the tweets using a web-based feedback service.[7] Note that using AMT for assessing the topical relevance of documents is an established procedure [Grady and Lease 2010].

For the AMT survey, we selected at random 200 English-language tweets each from the expert digest and the random digest.[8] During the survey, each AMT worker was shown a tweet and five categories—(i) conversational tweet/babble, (ii) tweet containing sentiment/opinion but no topical information, (iii) advertisement/sales promotion, (iv) spam, and (v) tweet containing useful topical information—and was asked to judge under which category the tweet falls. In the AMT survey, we opted for 'master workers' who are known to perform tasks more accurately. In total, 12 AMT workers were used for the evaluation, and each tweet was judged by three distinct evaluators. There was a high degree of agreement between the judgements of various evaluators on each individual tweet; there was unanimous agreement for over 47% of the tweets and a majority agreement for over 93% of the tweets. Table IV summarizes the majority decisions for the categories of the tweets.

Over 90% of the tweets in the random digest were judged to be merely conversational or expressing some personal sentiment/opinion (i.e., without having any topical information). In sharp contrast, 43% of the tweets from the expert digest were judged to contain useful information on some specific topic. Also note that 1.6% of the tweets in the random digest were judged to be spam (e.g., adult content, or promoting mechanisms to acquire more followers in Twitter), whereas none of the tweets in the expert digest were judged to be spam.

Table IV also shows the percentage of tweets in each category that contained hashtags or URLs (the second set of values within parentheses). We see that a large majority of the tweets that were judged to contain useful information on some topic also contained URLs or hashtags (especially in the expert digest). Thus, whether a tweet contains useful topical information is highly correlated with the presence of hashtags or URLs in the tweet (as has also been observed in prior studies [Teevan et al. 2011]). Hence, we compare the number of hashtags and URLs contained in the expert digest and a similarly sized random digest (as described in Section 3) in Table V. Compared to the random digest, the expert digest has *twice as many tweets with hashtags* and *four times as many tweets with URLs*. The abundance of hashtags and URLs in expert digest suggests that it is a much richer source of information than the random digest. A detailed comparison of the hashtags and URLs contained in the two digests is given later in Section 5.3.

[7]Amazon Mechanical Turk: https://www.mturk.com/mturk/welcome.

[8]Following the approach of Hannak et al. [2012], we consider a tweet to be in English if at least 70% of the words contained in it appear in the `wamerican-small` English dictionary.

Table V. Comparison of Information Content in the Expert Digest and a Similarly Sized Random Digest (Subsampled Version of the Twitter 1% Digest)

Property	Random digest	Expert digest
# Tweets	2,264,904	2,264,904
# Retweets	490,057 (21.6%)	409,920 (18.1%)
# Tweets with Hashtags	290,602 (12.8%)	571,662 (25.2%)
# Tweets with URLs	281,484 (12.4%)	1,183,070 (52.2%)
Distinct Retweets	407,749	342,633
Distinct Hashtags	135,471	165,986
Distinct URLs	246,057	994,967

5.2. Comparing Topical Diversity in the Digests

We now analyze the variety of topics covered by the tweets in the two digests. Since the digests contain millions of tweets, it is infeasible to manually identify the topic of each tweet. Hence, we use *topic models*, which are statistical models for discovering topics in a text corpus. Note that topic models have been used by several prior studies to discover topics from large volumes of tweets [Morstatter et al. 2013; Ramage et al. 2010; Yin et al. 2011].

We use the widely used topic modeling algorithm Latent Dirichlet Allocation (LDA) [Blei et al. 2003]. LDA identifies topics as mixtures of terms, and treat documents as mixtures of topics. Each topic is represented by a probability distribution over the vocabulary of all terms, where the probability values convey the affinity for a given term to a particular topic.

For our experiments, we considered only English tweets in the two digests (i.e., tweets for which at least 70% of the words appear in the `wamerican-small` English dictionary [Hannak et al. 2012]). We preprocessed the tweets by case-folding (converted all English characters to lowercase), and removed URLs, @user mentions, punctuation, emoticons, and so on, and also removed a set of well-known English stop-words. Hashtags were treated similar to other words in this analysis, for example, the hashtag "#christmas" and the word "christmas" were considered as the same term. We then used the GibbsLDA++ implementation of LDA [Phan and Nguyen 2007] to identify topics.

The LDA algorithm takes three input parameters: the number of topics N, a hyperparameter α for the Dirichlet prior topic distribution, and a hyperparameter β for the Dirichlet prior word distribution. Choosing optimal values for these parameters is a challenging problem and is not the focus of this work. Hence, we use $N = 100$ topics (as also used by Morstatter et al. [2013]), and default values of the GibbsLDA++ tool for the other parameters: $\alpha = 50/N$, and $\beta = 0.1$. Note that it is possible that the optimal values of the aforementioned LDA parameters are different for the two samples. However, because the objective of this study is to perform a fair comparison between the two sampling strategies, we use the same parameter values for both the expert and random digests and compare the topics discovered by LDA from the two digests.

Mapping the topics discovered by LDA to broad topics: A topic identified by LDA (henceforth denoted as a LDA-topic) is essentially a distribution over the terms in the vocabulary of the dataset. We conducted another AMT survey to map the LDA-topics to broad topics understandable by human beings, such as politics, sports, education, entertainment, and so on. The complete list of broad topics considered is given in Table VI).[9] Each AMT evaluator was shown the top 10 terms in a LDA-topic, (i.e., the 10 terms that were assigned the highest probability values for this topic by LDA) and

[9]This set of topics is a reduced version of the Yahoo category directory (http://dir.yahoo.com/).

Table VI. Mapping the Topics Discovered by LDA (LDA topics) to Broad Topics through AMT Survey

Broad topic	Random digest	Expert digest
Lifestyle & Culture (including religion, fashion, celebrity news, pets, food, . . .)	33	21
Entertainment (including movies, music, TV shows, theatre, comedy, . . .)	17	12
Government & Politics (including law, crime, military affairs, . . .)	8	11
Science & Technology (including basic sciences, computers and Internet, online social media, consumer electronics, . . .)	7	13
Sports	6	15
Education	4	1
Business & Economy (including job search)	2	11
Health & Fitness	2	4
Arts & Humanities (including history, geography, literature, books, . . .)	0	3
Environment & Weather	0	3
Could not be mapped to any of the above broad topics	21	6

The columns show the number of LDA topics (among the 100 LDA topics discovered for each digest), which were judged to be related to the corresponding broad topic by majority judgement.

was asked to indicate the most relevant broad topic. The evaluators were also given the option to indicate that a particular LDA-topic could not be related to any of the given broad topics. Similar to the AMT survey described in Section 5.1, each LDA-topic was judged independently by three evaluators; there was unanimous agreement in case of about 48% of the LDA-topics, and a majority agreement (i.e., two of the three evaluators indicated the same broad topic) in case of about 90% LDA-topics.[10]

The results of the user survey are summarized in Table VI. It is seen that a substantial fraction (21 out of 100) of the LDA-topics identified from the random digest could not be mapped to any of the broad topics. These topics are mostly related to daily conversation, greetings, colloquial/slang words, and so on (some examples are shown later in Table VIII). Furthermore, 50% of the LDA-topics identified from the random digest are related to only 2 of the 10 broad topics—lifestyle and culture (e.g., celebrity news, Christmas celebrations, fashion), and entertainment (e.g., music and television shows)—while the other broad topics are much less represented. In fact, two of the broad topics—arts & humanities and environment & weather—are not at all represented among the LDA-topics identified from the random digest. In contrast, not only could LDA identify a much larger set of meaningful topics from the expert digest (94 out of the 100 LDA-topics could be mapped to one of the broad topics), but also these LDA-topics cover all the specified broad topics.[11] This shows that the tweets in the expert digest cover a much wider diversity of topics, compared to the random digest.

Comparing the topics discovered from the two digests: We now compare the topics discovered by LDA from the two digests. Because LDA-topics have no implicit ordering, we need to match them based on the similarity of the terms in the topics. As

[10]There were a few LDA-topics for which there was no majority decision (i.e., which were judged differently by each of the three evaluators). These LDA-topics were found to contain a mixture of words most of which are unrelated to any one broad topic. Hence, we consider that these topics could not be mapped to any of the broad topics, and are counted in the last category in Table VI, in addition to those LDA-topics that were explicitly judged to be not related to any broad topic by a majority of the evaluators.

[11]Education is the only broad topic for which there are more LDA-topics in the random digest than in the expert digest. This is because the random digest contains a number of tweets posted by school/college students about their examinations, homework, classes, and so on.

Table VII. Examples of LDA-Topics Identified from Expert Digest That Were Not Found in the Random Digest

Sl.	Top 10 terms in topic identified by LDA	Broad topic
1	data, cloud, network, technology, security, services, management, software, access, solutions	Science & Tech
2	money, david, chicago, age, reasons, theatre, author, wilson, tea, startups	
3	sign, tax, support, starbucks, edition, amazon, companies, newsletter, browns, raiders	Business & Eco
4	internet, future, conference, rules, talks, freedom, valley, startup, leanstartup, dubai	Science & Tech
5	america, country, eye, storm, land, cities, river, mary, residents, philippines	Env & Weather
6	mark, side, west, island, jets, tim, ave, jersey, greg, sanchez	Sports
7	support, government, rights, congress, members, youth, disabilities, senate, leader, society	Politics & Govt
8	house, bill, hair, move, paper, floor, rep, session, order, smith	Politics & Govt
9	bank, cash, market, estate, rate, credit, debt, crisis, economy, investment	Business & Eco
10	fire, series, test, track, career, africa, bike, mountain, progress, australia	Sports
11	winter, trade, meetings, mike, source, boston, scott, rick, mlb, agent	Sports
12	head, coach, board, paul, romney, jeff, dave, kevin, florida, moore	Sports
13	health, nurse, cost, insurance, safety, healthcare, system, benefits, costs, doctor	Health & Fitness
14	email, group, tumblr, info, number, contact, mail, campaign, spread, details	Science & Tech
15	syria, attack, weapons, staff, security, war, turkey, iran, government, army	Politics & Govt
16	car, brand, challenge, japan, lincoln, company, ford, cars, auto, japanese	Business & Eco
17	media, content, ways, pinterest, socialmedia, strategy, brands, engagement, audience, infographic	Science & Tech
18	science, space, george, earth, research, system, surprise, sky, nasa, mission	Science & Tech
19	record, study, tech, brain, case, carbon, hits, records, degrees, louisiana	Science & Tech
20	check, john, article, issue, interview, magazine, location, writer, features, articles	Arts & Humanities
21	airport, winds, pressure, weather, snow, mph, rain, fog, field, county	Env & Weather
22	art, price, artist, piece, miami, pop, arts, museum, history, culture	Arts & Humanities

The top 10 terms (case-folded to lowercase) in each LDA-topic are shown. The last column identifies the relevant broad topic, as judged by the majority in an AMT survey (blank if no broad topic could be identified).

stated earlier, a topic identified by LDA is actually a probability distribution over all terms in the vocabulary; however, for simplicity, we consider a LDA-topic as a set (bag of words) of the top 10 terms according to the probability values for that topic assigned by LDA. We then measure the similarity between two LDA-topics by the Jaccard score of the corresponding sets of terms (similar to Morstatter et al. [2013]). In other words, the similarity between a LDA-topic T_i^R identified from the random digest (R) and a topic T_j^E identified from the expert digest (E) is computed as:

$$\text{sim}\left(T_i^R, T_j^E\right) = \frac{|T_i^R \cap T_j^E|}{|T_i^R \cup T_j^E|}. \quad (1)$$

For each LDA-topic in one digest, we attempt to find similar topic(s) in the other digest. We find that there are several pairs of very similar LDA-topics identified from the two digests. For instance, both digests contain topics related to technology (e.g., iPhone), sports (e.g., baseball and basketball in the United States), celebrity news (e.g., on the issue of the baby of the England royal family) and society (e.g., Christmas celebrations) which were popular topics of discussion in December 2012.

Because our objective is to understand the differences among the two digests, we focus more on the LDA-topics identified from one of the digests, which were not found

Table VIII. Examples of Topics Identified by LDA from Random Digest That Were Not Found in the Expert Digest

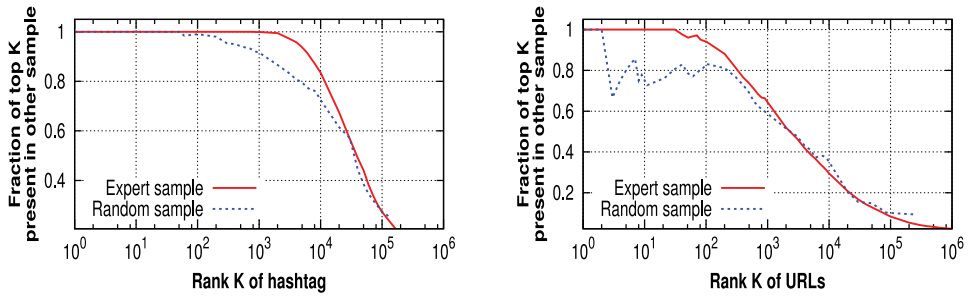
Sl.	Top 10 terms in topic identified by LDA	Broad topic
1	hoy, dormir, sin, ahora, estoy, dice, comer, bueno, puedo, eso	
2	hell, ball, dude, nope, jingle, balls, ocean, vampire, shot, screw	Entertainment
3	birthday, hope, bday, 21st, xoxo, crocs, xxx, march, hun, yrs	Lifestyle & Culture
4	practice, fact, pisces, security, haters, cheer, babies, williams, spread, dope	Lifestyle & Culture
5	life, story, candy, button, toy, decisions, moments, sea, thug, american	Lifestyle & Culture
6	feelings, mood, thoughts, chick, relationships, information, taurus, gemini, virgo, emotions	Lifestyle & Culture
7	hair, towie, towielive, color, sam, cunt, clothes, wash, joey, makeup	Entertainment
8	mine, hand, gym, weight, booty, butt, favor, workout, guitar, loss	Health & Fitness
9	xxx, hoes, dem, den, luv, sosa, drama, bout, knw, dats	
10	xxx, duck, vibe, stfu, karma, des, ton, par, mes, suis	
11	bout, idk, avi, bruh, lookin, disney, voice, lls, hood, shots	Entertainment
12	death, rain, sound, pizza, order, west, jay, bunch, dale, wifi	Lifestyle & Culture
13	mind, problem, bit, mouth, blow, attitude, tongue, wouldnt, mine, spongebob	Entertainment
14	fire, hell, nose, heaven, experience, earth, angel, brand, freak, form	Lifestyle & Culture
15	promise, leo, nights, joe, moon, brothers, stars, crack, twilight, jonas	Entertainment
16	wake, xxx, xxx, stomach, xxx, waste, clock, alarm, mins, xxx	
17	ada, yang, skype, jam, pin, lah, tak, meh, yah, hah	
18	hoje, tem, tempo, ela, mundo, ele, agora, gente, quero, seu	
19	music, internet, country, sun, share, web, sky, shine, diamond, rise	
20	niall, liam, louis, zayn, danielle, perrie, mia, eleanor, sono, taylor	Entertainment
21	town, london, bill, tom, lou, cars, chase, rick, cindy, george	Lifestyle & Culture
22	thewalkingdead, eye, action, episode, february, daughter, libra, walkingdead, heads, headcount	Entertainment
23	body, summer, count, seconds, limit, bomb, syria, camp, character, massage	Politics & Govt
24	touch, drop, price, teeth, space, record, foot, nails, feet, flow	
25	kiss, hug, hands, lips, swag, smell, pack, neck, arms, goodbye	Lifestyle & Culture

The last column identifies the relevant broad topic, as judged by the majority in an AMT survey (blank if no broad Topic could be identified). Some of the topics include abusive words, which have been replaced with "xxx."

in the other digest. We consider a certain LDA-topic as not found in the other digest if the similarity of this topic with every topic in the other digest (as measured by Equation (1)) is less than 0.1.[12] Table VII shows the 22 LDA-topics identified from the expert digest, which could not be matched to any topic identified from the random digest. For each LDA-topic, the top 10 terms are shown, along with the relevant "broad topic" as identified through the AMT survey described above (or blank, if no relevant broad topic could be identified for a certain LDA-topic). It is evident that the expert digest contains several news stories that are missing in the random digest, for example, stories related to politics (e.g., topic 15 in Table VII—conflict in Syria), health (e.g., topic 13—the healthcare debate in the United States), science & technology (e.g., topic 1—data management, topic 18—space research), sports (e.g., topic 6—National Football League in the United States, or topic 10—cricket test series between Australia and South Africa), and economy.

Conversely, Table VIII shows the 25 LDA-topics identified from the random digest that are not found in the expert digest. Almost all these topics either correspond to the large amounts of day-to-day conversation present in the Twitter random sample

[12]We also experimented with other threshold values between 0.05 and 0.3; however, the observations qualitatively remain same as reported in the main text.



(a) Hashtags (b) URLs

Fig. 5. The fraction of top K hashtags/URLs in one digest, which are present in the other digest (shown for different values of K).

(as also observed in Section 5.1) or are related to the two broad topics of Lifestyle & Culture, and Entertainment.

It is evident that if one collects only the random digest, one may potentially miss specialized information on various topics such as politics, sports, technology, and economy. On the other hand, if one collects only the expert digest, one may miss some amount of information on topics that are popular among the masses (e.g., television programs, celebrity news), but the information that is missed is of limited topical scope.

5.3. Comparing Hashtags and URLs in the Digests

The analysis in Section 5.1 suggested that hashtags and URLs are the primary indicators of meaningful topical content in the digests. Hence, in this section, we focus on comparing the hashtags and URLs (which we collectively refer to as "terms") contained in the two digests.

To differentiate between popular and not-so-popular terms, we rank the terms contained in a digest on the basis of their popularity (i.e., the number of distinct users who have posted the particular term in that digest).

Overlap between terms in the digests: We start by checking the overlap between the terms (hashtags / URLs) in the two digests. Figure 5(a) shows the fraction of the top K (i.e., most popular) hashtags in one digest that occur at least once in the other digest, for various values of K. For instance, we see that all the top $K = 1{,}000$ hashtags in the expert digest (represented by the solid red curve) appear in the random digest, whereas about 90% of the top 1,000 hashtags in the random digest (represented by the dashed blue curve) appear in the expert digest. Similarly, Figure 5(b) shows the corresponding fractions in case of URLs. We find that for the popular hashtags/URLs (i.e., for lower values of K), the fraction overlap is high, but the overlap falls off quickly for less popular content. Overall, less than 10% of all the terms contained in one digest occur in the other, whereas more than 90% of the top 1,000 (most popular) terms in either digest also appear in the other digest. Hence the information content of the two digests is quite similar amongst the popular content, but less similar amongst niche content.

In case of both hashtags and URLs, a larger fraction of the top terms in the expert digest is present in the random digest, than the fraction of the top terms in the random digest that is present in the expert digest. The overlap is especially poor for the top URLs in the random digest, for example, 3 out of the top 10 URLs in the random digest do not even appear in the expert digest (see Figure 5(b)).[13] This shows that the content

[13]The following URLs are all linked to online gaming sites (or related pages in Facebook): http://gigam.es/mtw_Tribez , http://gigam.es/etw_Tribez , and http://tiny.cc/lhs0mw.

(hashtags/URLs) that is popular in the expert digest is also relatively popular in the random digest, but the reverse is not always true. Later in this section, we discuss the popular content in one digest that are not popular (or even missing) in the other digest.

We also observe that, in general, the overlap is lower in case of URLs (Figure 5(b)) as compared to that for hashtags (Figure 5(a)), which can be explained as follows. Hashtags in Twitter are used to refer to current events or news stories [Yang et al. 2012b], and users who comment on a certain event/news story often use one or two specific hashtags. On the other hand, a popular news story/event is usually discussed in various websites/web pages (e.g., by different news media sites); therefore, there is a larger variety among the URLs tweeted by the users who wish to comment on the news story. Hence, it is natural to have higher overlap among the hashtags (than for URLs) even if the same news stories are being discussed in the two digests.

Correlation between popularity of common terms: Next, we focus on the terms (hashtags/URLs) that are common in both digests and check whether the popularity of a given term in one digest is correlated with its popularity in the other digest.

We use the well-known metric Pearson product-moment correlation coefficient (PCC) to measure the correlation between the popularities of the common terms in the two digests. Let V_E and V_R represent the vocabulary (set of all distinct terms) of the expert digest and the random digest, respectively, and let the popularity of a term t in the expert digest and the random digest be denoted by $t.p_E$ and $t.p_R$, respectively. Also, let the mean popularity of all terms in a digest be denoted by $\overline{p_E}$ for the expert digest, and $\overline{p_R}$ for the random digest. Then the PCC (which lies in the range $[-1, 1]$) between the popularities of terms in the two digests, is computed as follows.

$$PCC = \frac{\sum_{t \in V_E \cap V_R} (t.p_E - \overline{p_E})(t.p_R - \overline{p_R})}{\sqrt{\sum_{t \in V_E \cap V_R} (t.p_E - \overline{p_E})^2} \sqrt{\sum_{t \in V_E \cap V_R} (t.p_R - \overline{p_R})^2}} \tag{2}$$

We find only moderate correlation between term-popularities in the two digests—the PCC values are 0.328 for hashtags and 0.358 for URLs. Thus, even for the terms that are common between the two digests, significant differences between the popularities of the terms in the two digests exist. Furthermore, although the overlap among URLs in the two digests is lower than that for hashtags, the correlation between popularities is slightly higher for the URLs that are common between the two digests than for the common hashtags.

We now attempt to understand the differences in the terms contained in the two digests. The previous discussions show that both digests contain hashtags and URLs of widely varying popularity. Hence, in the rest of this section, we separately study the most popular content and the less popular content. We report results only for hashtags for brevity. Similar trends were observed for URLs, with the exception that the overlap between URLs in the two digests is consistently lower than that for hashtags (as explained earlier).

Comparing popular hashtags: As stated earlier, we ranked the hashtags according to the number of distinct users who have posted a hashtag in each digest. To focus on the most popular content in the digests, we considered the top 1,000 hashtags in each digest according to this ranking. As was shown in Figure 5(a), more than 90% of the top 1,000 hashtags in either digest is also contained in the other digest as well. However, some of the top hashtags in one digest are ranked much lower (or even absent) in the other digest. To better understand the differences among the popular content in the two digests, we studied those hashtags which are among the top 1,000 in one digest, but ranks beyond 10,000 in the other digest (or, are absent in the other digest).

Table IX. Examples of Top (Most Popular) 1,000 Hashtags in One Digest That Are Either Absent or Not Popular (Ranked Beyond Top 10,000) in the Other Digest

Top 1,000 hashtags in expert digest that are not popular/absent in random digest		Top 1,000 hashtags in random digest that are not popular/absent in expert digest	
Theme	Example hashtags	Theme	Example hashtags
Technology	TrustCloud, IBM, opensource, healthIT	Acquiring followers	refollow, instantfollow, followforfollow, teamfollowback, autofollow
Politics	doma, ssm, prop8 (all related to US Gov's stance on same-sex marriage.)	Twitter-specific memes	sadtweet, cantsleep, happytweet
Literature	nanowrimo (national novel writing month), aca, amreading	Games	farmville, iphonegames, androidgames
Ongoing events	leweb12 (conference of web technology business leaders)	Ongoing events	niallspanish, goodluck1d (all related to concert of One Direction band)
Business	b2b, crm, smallbusiness, custserv	Entertainment	LoveInParis (television series)

Among the top 1,000 hashtags in the expert digest, 54 are absent / appear beyond the top 10,000 in the random digest. Similarly, 170 hashtags among the top 1,000 in the random digest, are absent / appear beyond the top 10,000 in the expert digest. A sample of these hashtags have been shown in Table IX, along with a descriptive theme to which the hashtags are related.

We find that the comparison of popular hashtags in the two digests yields very similar observations to those obtained from the topic model analysis in Section 5.2. The hashtags from the expert digest in Table IX (which are absent/not popular in the random digest) are related to specialized topics in technology, politics, literature, and business, which might explain their lack of popularity amongst the general Twitter crowds. On the other hand, the hashtags that are very popular in the random digest, but are absent/not popular in the expert digest, are mostly related to acquiring followers in Twitter, Twitter-specific memes, and entertainment (games, music concerts, television programs). Intuitively, the experts would have lesser interest in these topics, which explains the much lower ranks of these hashtags in the expert digest.

Comparing hashtags in the long-tail: Finally, we attempt to characterize the less popular hashtags (i.e., the hashtags in the long-tail in the two digests). Since the tail of either digest is likely to contain large amounts of conversational/personal hashtags whose context would be difficult to identify algorithmically, we preferred to use human judgement to characterize these hashtags. We randomly selected 200 hashtags from each digest, from among the ones that have been posted by less than five distinct users. Two volunteers independently studied the selected hashtags, attempting to characterize them. The volunteers were instructed to see the tweets containing the hashtags in order to understand the context and also to use Web search on the hashtags if necessary. There was over 87% agreement between the observations of both volunteers, which is summarized in Table X.

Overall, the tails of both digests have content of similar nature; however, the relative fraction of specific types of content was widely different. In both digests, some of the hashtags were, in fact, screen names of Twitter users, and the volunteers could not understand the context of a significant fraction of hashtags (e.g., non-English ones); these are grouped together under category "others" in Table X. The tail of the expert digest was found to be much more topical as compared to that of the random digest—48% of the less popular hashtags in the expert digest were related to a variety of niche

Table X. Examples of Less Popular Hashtags (Posted by Less Than Five Distinct Users) in the Two Digests

Type of Hashtag	Hashtags in expert digest		Hashtags in random digest	
	%	Example hashtags (explanations)	%	Example hashtags (explanations)
Topical	**48%**	jfr (Japan Fire Rescue), fordham Football, rncareer (Registered Nursing Career), estampasenvivo (Live Photography), jgpfinal (Junior Grand Prix Final), anthrax	**13%**	mfad (a music album), sixtoes (a music band), Java6, bh_news (Boston Herald news), ghana4peace
Opinion, sentiment	**10%**	WeThink, Intuitive, SensitiveGuy	**23%**	naynay (dismissive gesture), mommyhelp (pretending to asking help from mother), yoyooo (happy)
Twitter-specific memes	**9%**	NotAtAllPredictable, fitness4mom, wishIknewAt17, justMentioning	**18%**	FavoriteDay, OurLegendaryHash-tags, SheIsAnnoying
Others	**33%**	Non-English/Twitter user-names used as hashtags/could not categorize	**46%**	Non-English/Twitter user-names used as hashtags/could not categorize

A large fraction of such hashtags in the expert digest are related to niche topics, whereas Twitter-specific memes and hashtags expressing opinion/sentiment account for most of these hashtags in the random digest.

topics (e.g., Japan Fire Rescue, Registered Nursing Career, Junior Grand Prix Final), as compared to only 13% in the random digest. On the other hand, the hashtags in the tail of the random digest were dominated by Twitter-specific memes, and hashtags expressing personal opinion or sentiment of the user posting the tweet.

Summary: The detailed analyses in this section bring forth a number of interesting observations. The expert digest is found to contain much richer topical information as compared to the random digest—not only does the expert digest contain much higher number of hashtags and URLs, but it also covers a wider variety of topics (as shown by the topic model analysis). On the other hand, a very large majority of the tweets in the random digest are merely conversational or related to only few topics such as entertainment and lifestyle. Most of the very popular content in the expert digest is relatively popular in the random digest as well; however, some of the popular content in the expert digest contains specialized information on topics such as technology, politics, and business, which are much less popular in the random digest. On the other hand, a significant fraction of the popular content in the random digest is not at all popular in the expert digest (or, does not even appear in the expert digest)—these mostly correspond to information that is of limited topical interest, for example, day-to-day conversation, Twitter-specific memes, and methods of acquiring followers in Twitter. Furthermore, the less popular content in the expert stream is also considerably more topical (related to niche topics) than the less popular content in the random stream.

These observations show the far greater value of the expert tweet stream for information retrieval and data mining applications, such as topical search and recommendations. However, the popular content in the random digest is much more faithful to the interests of the masses (acquiring followers, games, entertainment, fashion, etc). Thus, selecting one digest over the other has its trade-offs and is best done depending on the specific task at hand.

6. TRUSTWORTHINESS

A major concern with crowd-sourced data is its trustworthiness. Since data sampled from Twitter is often used for search and recommendation [Choudhury et al. 2011a; Sankaranarayanan et al. 2009], it is important to ensure that the search/recommendation results do not contain spam or malicious URLs. In this section, we investigate the trustworthiness of the content in the expert and random digests.

6.1. Malicious Content and Unvetted Users

A number of recent studies have highlighted the growing problem with spam in Twitter [Grier et al. 2010; Thomas et al. 2011]. Hence, we start by analyzing the amount of malicious content (blacklisted URLs) in both expert and random digests.

Malicious URLs in the two digests: We check whether an URL in a tweet is malicious or blacklisted by attempting to fetch the corresponding web-page following the HTTP redirects, if any (e.g., if the URL has been shortened by some URL shortening service). Some URL shortening services such as bit.ly and tinyurl have implemented warnings for malicious URLs, attempting to fetch a malicious URL shortened by these services leads to an interstitial warning page. We check for such malicious URLs in case the URL is a shortened one. Additionally, the final landing URL is checked using the Google Safebrowsing API.[14]

We randomly selected 1 million URLs each from the two digests, and verified how many of them were blacklisted. We found 1,520 blacklisted URLs in the random digest, which were posted by 1,447 distinct users. On the other hand, we found only 129 blacklisted URLs posted by 46 distinct users in the expert digest. This implies that the random digest contains about 12 times the number of blacklisted URLs contained in the expert digest.

Unvetted users in the two digests: Next, we focus on the users who were found to have posted the blacklisted URLs in the two digests. For these users, we attempted to crawl their profile details and all the tweets posted in their lifetime; we also checked all the URLs posted by these users for blacklisted ones.

We found that out of the 1,447 users who posted a blacklisted URL in the random digest gathered in December 2012, 912 (i.e., 63.026%) were suspended by Twitter during the next 1 year (i.e., until December 2013), indicating that these users were probably spammers. Furthermore, out of these 912 suspended users, 468 had joined Twitter in December 2012 itself, which suggests that a considerable fraction of spammers/users posting blacklisted URLs are *recent joinees* who have not yet been thoroughly vetted. A considerable fraction of the remaining accounts (which have not been suspended until December 2013) have posted more than 100 blacklisted URLs in their lifetime, thus seriously undermining the trustworthiness of the tweets.

In contrast, none of the 46 experts who had posted a blacklisted URL have been suspended to date. In fact, we found that 32 of them have posted exactly one blacklisted URL in their entire lifetime, suggesting that these expert users may have posted the particular blacklisted URL unwittingly. The rest of the experts have posted multiple blacklisted URLs, but most such URLs point to articles on their personal websites, and those websites seem to have been compromised and blacklisted.

Note that Twitter already attempts to filter out spam/blacklisted URLs before providing the random sample stream. Despite this, there is a substantial amount of spam in the random sample. This confirms that real-time filtering of spam is very difficult, given that the blacklists are relatively much slower in adding blacklisted URLs as

[14]https://developers.google.com/safe-browsing/.

Table XI. Top 5 Users Who Dominate the Expert Digest and the Random Digest (i.e., Who Have a Very High Number of Tweets in the Digests)

Expert digest	Random digest
notiven: Venezuelan news feeds	**Favstar_Bot**: I tell you when your tweets are faved.
favstar50: collect your favs and retweets instantly	**currentcet** (tweets current CET time, every minute)
GetAFreelancer: World's largest outsourcing marketplace	**cashcoming**: Free tips to earn money online working from home
faithclubdotnet: God is real! Jesus is LORD!	**Jonellwx5**: Shop and Save at Beater Mixer
ContactoLatino: News headlines & opinion articles	**XboxSupport**: Guinness World Record Holder – Most Responsive Brand on Twitter!

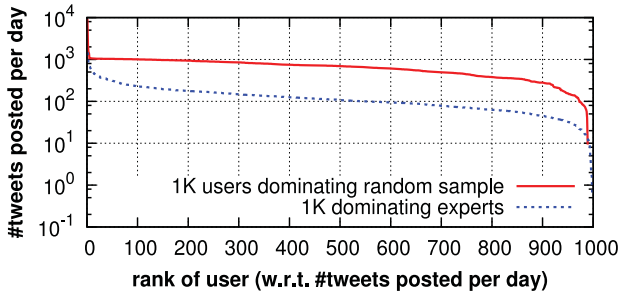


Fig. 6. Average number of tweets posted per day (over entire lifetime) by the top 1,000 users who dominated the random digest and the expert digest during December 2012.

compared to the rate at which such URLs are generated in Twitter [Grier et al. 2010]. Hence, an added advantage of sampling tweets from experts (who are well-established users) is that they are very less likely to post spam.

6.2. Spam/Promotional Campaigns to Distort Random Sampling

We now investigate whether the 1% and 10% randomly sampled tweet streams provided by Twitter encourage some specific users to tweet more, in the hope of increasing their chances of being represented in the resulting random digests. For this, we measured the number of times each individual user's tweets are captured in the two digests during the month of December 2012, and we identified those users whose tweets are sampled most number of times in the two digests.

Table XI shows some of the users who dominate the two digests, that is, who have a very high number of tweets in the digests. Most of the users who dominate the random digest are marketers or campaigners who wish to promote their business (e.g., *cashcoming*, *Jonellwx5*). In case of the experts who dominate the expert digest, we observe that some of them are news media sites (which frequently post current news, e.g., *notiven*, *ContactoLatino*), while others advertise for freelance workers or job openings (e.g., *GetAFreelancer*).

Note that it is quite natural that some users will tweet more than others, and these users will intuitively get sampled more frequently in the random digest. In fact, as just stated, some experts also tweet much more frequently than the others. However, it is important to note that although these experts post more frequently compared to other experts (and hence dominate the expert digest), their rate of posting tweets is much lower compared to the accounts which dominate the Twitter random sample. To demonstrate this, we plot in Figure 6, the average number of tweets posted per day (throughout their lifetime) by the top 1,000 users who dominated the random and

expert digests (in this figure, the 1,000 dominating users have been ranked with respect to the average number of tweets they post per day). Most of the top sampled users in the random digest contribute several hundreds of tweets every day; in fact, the 300 most dominant users in the random digest post close to the maximum allowed limit of 1,000 tweets per day. In contrast, the most sampled experts contribute an order of magnitude fewer tweets than the users who dominate the Twitter random sample.

Taken together, our findings show that (i) compared to expert digests, random digests include considerably more spammers and their spam content, and (ii) most of the spammers in the random digest are users who have joined Twitter recently and have not yet been vetted by Twitter's spam filters. Thus, unvetted users pose a fundamental problem for random sampling techniques. Furthermore, random sampling offers perverse incentives for spammers/marketers to tweet more frequently, as the probability of their tweets being sampled increases linearly with the number of tweets they post. On the other hand, content sampled from experts is free from such campaigns.

7. TIMELINESS OF INFORMATION

In this section, we analyze the timeliness or recency of information obtained from the two digests. In other words, we check that when an unforeseen event occurs (such as a natural calamity/accident), which digest gives the earliest information about the event. For this, we considered five such events that occurred in December 2012:

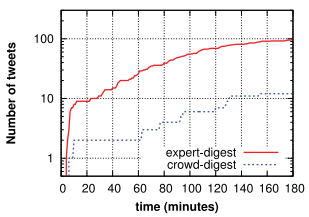
(1) The ship Baltic Ace sank in the North Sea after collision with another ship on December 5 (E1)
(2) A strong earthquake in Japan on December 7 (E2)
(3) Mexican-American singer Jenni Rivera died in an air crash on December 9 (E3)
(4) Shooting at Sandy Hook Elementary School in Connecticut on December 14 (E4)
(5) Nigerian State governor Patrick Yakowa died in a helicopter crash along with other politicians on December 15 (E5)

For each event, we identified the tweets related to the event that appeared in the expert digest and the random digest immediately after the event occurred. We used keyword-based search to identify the related tweets; for instance, keywords such as "Sandy Hook," "Sandyhook," and "shooting" were used to identify the tweets related to the incident E4 (shooting at Sandy Hook elementary school).

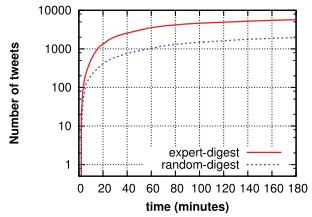
Figure 7 shows, for each event, the growth (with time) of the number of related tweets in the two digests during the first 3 hours after the event occurred/was first reported. For each of the five events, *a relevant tweet first appeared in the expert digest*. For the first three events, the first tweet in the random digest appeared within 2 minutes of the first tweet in the expert digest; however, the delay was longer (more than 10 minutes) for the events E4 and E5.

Table XII shows the first relevant tweet in the two digests, along with the user who posted the tweet and the time at which the tweet was posted. Interestingly, in each case, the first tweet in the random digest was a retweet of a tweet posted by an expert (or a popular user). In particular, for the events E1, E3, and E5, the first tweet in the random digest is a direct retweet of the first tweet in the expert digest. Also in case of E2 (earthquake in Japan), the first tweet observed in the random digest was a retweet of a tweet posted by @eew_jp_en (a Twitter account dedicated to Earthquake Warning), whereas the first tweet in the expert digest was a tweet posted by @eew_jp (the Japanese version of @eew_jp_en).

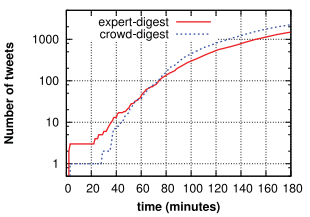
The aforementioned observations can be explained as follows. The set of experts includes a number of media sites/journalists (e.g., @Sahara Reporters) as well as Twitter accounts dedicated to emergency situations (e.g., @eew_jp @CTNotify). In some cases, these users post the earliest information about an unforeseen emergency, before the



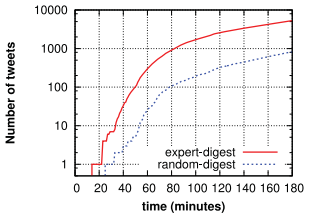
(a) E1: Ship sinks in North Sea



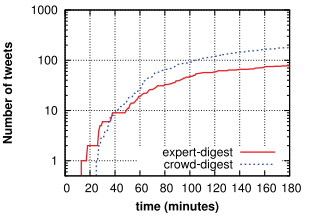
(b) E2: Earthquake in Japan



(c) E3: Us singer dies in air-crash



(d) E4: Shooting at Sandy Hook school



(e) E5: Nigerian politicians die in air-crash

Fig. 7. Number of tweets seen in the two digests immediately after an event occurs. Earliest information about the event was obtained in expert digest for all cases. First information in random digest appeared within 2 minutes for E1, E2, and E3, but there were delays of about 10 minutes for E4 and E5.

Table XII. The First Tweets in the Expert Digest and Random Digest, Along with the Users Who Posted the Tweets and the Time at Which the Tweets Were Posted

Event	Extracts from first tweet in expert digest	Extracts from first tweet in random digest
E1	**(20:03:58)** *mpoppel*: Emergency services in the Netherlands responding to reports of sinking cargo vessel in North Sea	**(20:05:24)** *peacelily01*: RT @mpoppel: Emergency services in the Netherlands responding to reports of sinking cargo vessel in North Sea
E2	**(09:19:11)** *eew_jp*: *Tweet in Japanese giving exactly the same information as the first tweet in random digest*	**(09:19:28)** *PipemanYoda*: RT @eew_jp_en: Dec 7, 2012 17:18:29 JST Off the coast of Sanriku Depth: 10km Mag.: 6.6 JMA #earthquake
E3	**(18:21:14)** *Rodpac*: RT @IvanSamuelLuz: Se estrella el avi de @jennirivera, salii de Monterrey a la 1 con destino a Toluca.	**(18:22:52)** *Gusmer1*: RT @Rodpac: Se estrella el avi de @jennirivera, sali de Monterrey a la 1 con destino a Toluca.
E4	**(15:48:46)** *CTNotify*: RT @Rickbryce @CTNotify @CTPSscann newtown ct active shooter in school multi police units responding	**(15:59:08)** *ProvFireVideos*: RT @HeidiVoight Hearing unconfirmed reports incident in #Newtown #CT may be school shooting. Police on the way
E5	**(18:12:54)** *SaharaReporters*: Governor of Kaduna Yakowa, Former NSA Azazi Feared Dead In Crashed Naval Chopper	**(18:24:02)** *Rukayamohammed*: RT @SaharaReporters: Governor of Kaduna Yakowa, Former NSA Azazi Feared Dead In Crashed ...

The timestamps are according to Central European Time.

general population comes to know about the event. However, it is important to note that the first tweet that is posted in Twitter about an unforeseen event is not always from an expert/popular user. In some cases, the very first post comes from a common user, but the post does not get included in the random sample because of the low probability of any individual tweet being sampled. This post is then retweeted by an expert (media site/emergency-related account) who understands its importance; the expert's post, in turn, gets retweeted by a large number of her followers, and some of these retweets

appear in the Twitter random sample after some time. Such examples are seen for the events E3 (Singer Jenni Rivera's death, which was also stated in Section 4.2) and E4 (Sandy Hook school shooting). For these events, the first tweets were posted by users @IvanSamuelLuz and @Rickbryce, respectively, both of whom are nonexperts (having only 180 and 93 followers, respectively). Their tweets were retweeted by the experts @Rodpac (for E3) and @CTNotify (for E4), before the information was available in the random sample.

Thus, we observe that news about unforeseen events can be obtained marginally quickly in the expert digest as compared to the random digest. More importantly, the expert sampling often filters out important information posted by nonexperts (as was also observed in Section 4.2), whereas such posts are very likely to be missed by random sampling, which gives equal importance to all tweets irrespective of the value of the information contained in a tweet. Hence, the timely availability of a certain information in the Twitter random sample is often influenced by whether an expert picks up the information and posts it, thus increasing its popularity manifold.

8. REPRESENTATIVENESS OF OPINIONS

A lot of recent research has focused on inferring the opinions of the crowd on specific topics or events (e.g., movie releases, political elections), by analyzing the sentiments expressed in their tweets on the topic [Hannak et al. 2012; Tumasjan et al. 2010]. In this section, we investigate whether opinions mined from the expert and random digests on various topics are similar or different.

8.1. Inferring Sentiment from Tweets

Several algorithms have been proposed for inferring sentiment from English text [Liu 2006]. However, most tweets posted in Twitter do not use proper spellings or grammar. Since tweets can be at most 140 characters long, they often contain abbreviations and acronyms (e.g., OMG, LOL). Hence, sentiment inference algorithms trained on proper English text do not work as well when applied to Twitter messages [Hannak et al. 2012]. A popular choice for analyzing sentiment in tweets is to use algorithms based on token-lists, which contain a set of tokens (words) with a sentiment score attached to each [Bradley and Lang 1999]. However, existing token-lists do not work well with Twitter data, as the tweets frequently contain abbreviations, acronyms, and Twitter-specific features, such as hashtags and @user mentions. To address these challenges, prior attempts to analyze sentiment of tweets [Hannak et al. 2012] constructed a Twitter-specific token-list by using the tweets themselves.

We follow the methodology proposed in Hannak et al. [2012] to create a Twitter-specific token-list, and then analyze the sentiment of the tweets. We consider only English tweets by checking whether at least 70% of the words (tokens) contained in a tweet appear in the `wamerican-small` English dictionary. We initially derive a set of clearly positive and negative tweets, by considering only tweets that contain either a positive emoticon (smiley) `:)` or `:-)` or a negative emoticon (frowny face) `:(` or `:-(` (but not both types of emoticons). Note that prior research on sentiment analysis has shown the utility of emoticons, which frequently match the true sentiment of the writer [Derks et al. 2007]. We tokenize the tweets, ignoring user-mentions and URLs. We also ignore any token that did not appear at least 20 times in the dataset. To build our Twitter-specific token-list, we compute the relative fraction of times a particular token occurred in a positive tweet (i.e., a tweet containing a positive emoticon) and use this fraction as the sentiment score of that token. Hence each token has a score between 0 and 1, which indicates the propensity of the token being used in a positive context. Finally, we calculate sentiment score of a particular tweet by identifying tokens contained in

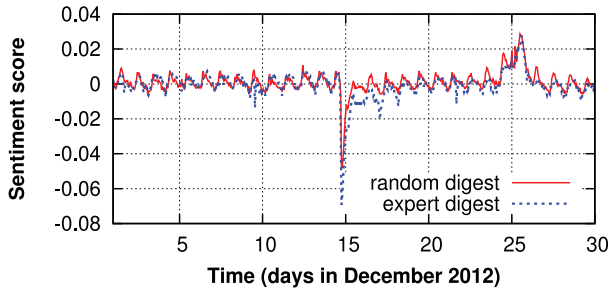


Fig. 8. Mean normalized sentiment scores of the set of tweets obtained in the two digests, during each individual hour in December 2012.

the tweet, and computing the weighted average of sentiment scores of the individual tokens in the tweet [Hannak et al. 2012].

We observed that the mean sentiment score for all tweets in the expert digest is 0.748 (standard deviation 0.060), and that for the random digest is 0.724 (standard deviation 0.074). In this section, we compare between the sentiments derived from the tweets related to specific topics / events in the two digests. In order to better compare the results across the two digests, we normalize the sentiment score of individual tweets. We normalize the sentiment score x of a tweet as $(x - \mu)/\sigma$ where μ and σ are respectively the mean sentiment score and the standard deviation of the digest from which the given tweet has been extracted. Finally, the sentiment score for a set of tweets is simply computed as the mean of the normalized sentiment scores for all tweets in the set; we refer to this score as the *mean normalized sentiment score* of the set of tweets.

8.2. Comparing Opinions from Expert vs. Random Digests

We start by analyzing how the cumulative sentiment of all tweets in the expert and random digests vary over time. We consider the sets of tweets obtained in the two digests during each hour in December 2012, and compute the mean normalized sentiment score of the sets. Figure 8 shows the mean normalized sentiment scores of the random digest and the expert digest, during each hour in December 2012. There is a clear diurnal variation in overall sentiment in both digests. Similar diurnal variation in overall sentiment in Twitter was observed earlier in Hannak et al. [2012]. Furthermore, we see that the sentiment scores of the experts show good agreement with that of the random sample across time in general; in fact, the Pearson correlation coefficient between the hourly sentiment scores is as high as 0.886. The sentiment scores also show similar variation corresponding to major events in the offline world. For instance, the dip in both sentiment scores during December 14–15, 2012 coincides with the tragic shooting incident in Sandy Hook Elementary School, while both sentiment scores rise during December 25–26 on the occasion of Christmas. Thus, both expert and random digests yield similar results when they are used to infer the overall mood of the Twitter user population.

Next, we compare opinions mined from expert and random digests about specific topics and events that generated considerable discussion in Twitter.

Mining opinions about major events/topics: We now look into a set of major events and topics that were discussed heavily in December 2012. Similar to the analysis in Section 7, tweets relevant to a specific event were identified through matching a set of keywords related to the event. We looked at events for which there were at least 500 relevant English tweets in both digests. The selected set of events and topics are (i) the

Table XIII. Mean Normalized Sentiment Scores of Tweets Related to Specific Events/Topics, in the Two Digests

Event/topic	Random digest	Expert digest
Shooting at Sandy Hook Elementary School	–1.817	–2.255
Fiscal cliff situation in the United States	0.282	–0.006
Right to Work law in the United States	–0.122	–0.339
Gang rape incident in Delhi, India	–0.993	–1.586
Syrian civil war	–0.482	–1.106
Egyptian protests	–0.094	–0.456

shooting at Sandy Hook Elementary School, (ii) the U.S. fiscal cliff situation, (iii) the Right to Work law in the United States, (iv) the gang rape incident in Delhi, India, (v) the Syrian civil war, and (vi) political protests in Egypt.

Table XIII gives the mean normalized sentiment scores of the tweets related to each event/topic, in the two digests. We find that for most of the topics/events (except for the fiscal cliff situation in the United States), the overall sentiment of the random digest is correlated with that of the expert digest, and is largely negative for the chosen set of topics/events. Although the sentiment scores of the entire random digest and the entire expert digest match well (as shown in Figure 8), we find that when specific topics/events are considered, experts express more pronounced feelings (i.e., more positive or negative), while those of the random digest are much closer to being neutral.

The only notable difference is in the topic of the US fiscal cliff. This topic refers to the automatic decrease in budget deficit in the United States, that could have occurred at the beginning of 2013 due to increase of taxes and lowering of government spending. The short-term impact of the subject roused heavy debate in the US media. For this particular topic, we find that expert opinion is more or less neutral (with very slight negative polarity), whereas that of the random sample digest is largely positive. We manually studied 100 randomly selected tweets from both digests which mention the topic of fiscal cliff. We find that 30% of the tweets in the random digest show positive sentiment, as compared to only 15% of the tweets in the expert digest. These tweets are mostly informative, providing links to websites on understanding the subject and related news. Importantly, the random digest has a much lower fraction (4%) of tweets with negative emotion, compared to 17% of the tweets expressing negative opinion in the expert digest on this particular topic. These tweets display negative emotions on the inability of the US senate and the president to come to a decisive solution to the issue. The rest of the tweets in both digests convey no definite positive or negative sentiments. We believe that this difference in opinion between the two digests is arising due to the higher level of awareness of the experts on this impending event.

Mining opinions about movies: Finally, we focus on tweets about major Hollywood movies that were released during November–December 2012. We extracted tweets related to each movie by looking for keywords and hashtags related to the movie in the tweet text. We selected six movies that had at least 500 related English tweets in both the expert and random digests: *Life of Pi*, *The Hobbit*, *Les Miserables*, *Skyfall*, *This Is 40*, and *Django Unchained.*For each movie, we computed the sentiment scores for the movie from the relevant English tweets in both the digests.

We observe that for a given movie, the distributions of normalized sentiment scores of individual tweets in both the digests usually follow Gaussian distributions. Since the distributions of sentiment scores for two movies often overlap, we use the following approach to decide whether the sentiment scores for two movies are significantly different. We obtain the 95% confidence range for the normalized sentiment score of the set of tweets related to a given movie, as $(\bar{x} \pm 1.96\, s/\sqrt{n})$, where n, $\bar{x}$ and s are respectively

Table XIV. 95% Confidence Intervals for the Mean Normalized Sentiment Score of Tweets Related to Hollywood Movies, from the Two Digests

Movie	Random digest	Expert digest
Life of Pi	0.331–0.305	0.296–0.268
The Hobbit	0.356–0.330	0.206–0.188
Les Miserables	0.350–0.288	0.190–0.144
Skyfall	0.348–0.262	0.189–0.127
This Is 40	0.197–0.147	0.092–0.054
Django Unchained	0.093––0.009	−0.227––0.295

The six movies are divided into two groups, where the sentiment score for each of the movies within a group is indistinguishable from that of at least one other movie in the same group, since their 95% confidence ranges overlap in one or both of the digests.

the size, the mean, and the standard deviation of the normalized sentiment scores of the given set of tweets. Finally, we do not distinguish between normalized sentiment scores for two movies, if their confidence ranges overlap.

Table XIV shows the mean normalized sentiment scores for the selected movies in the two digests, along with the 95% confidence range (obtained as stated earlier). Though the absolute values of the sentiment scores are higher in the random digest than in the expert digest, we find that the relative rankings of the movies in the two digests match very well. In both the digests, the six movies are clearly categorized into two groups. *Life of Pi*, *The Hobbit*, *Les Miserables*, and *Skyfall* evoked high positive sentiments in both digests with very close values; the 95% confidence interval for each of these four movies overlaps with that of some other movie in this group in at least one of the digests. Hence, we do not distinguish between their normalized sentiment scores. On the other hand, *This Is 40* and *Django Unchained* have relatively lower sentiment scores in both the digests. Thus, we find that both expert and random digests yield similar results when they are used to infer public opinion about movies. The only notable difference between the sentiment in the two digests is for *Django Unchained*, for which the sentiment scores are neutral in the crowd digest and largely negative in the expert digest. Manual observation of the related tweets reveals that the negative sentiments in the expert digest come predominantly in the context of users discussing the very negative subject of the movie (excessive depiction of slavery, racism, brutality) and not the quality of the movie itself.

9. CONCLUDING DISCUSSION

A lot of prior work has focused on mining useful information from the randomly sampled Twitter streams. However, few, if any, have questioned or examined the random sampling methodology used to generate the streams. The primary contribution of this work lies in investigating an alternate method for sampling the content being generated in Twitter, one where we only sample content generated by a large and diverse set of expert users.

While we risk losing certain statistical properties of the tweet population by moving from random to expert sampling, we find that the move offers a number of advantages for applications that rely on the content contained within the tweets. The tweets become more rich in information content, and more popular content gets sampled. We find that important tweets posted by nonexperts also appear in the expert sample as experts retweet them. Another important gain is in terms of trustworthiness of the sampled data; compared to random sampling, expert sampling yields digests that are mostly free from blacklisted URLs and are much less biased by automated marketing accounts tweeting continuously. Expert sampling also captures breaking news stories and important events in a timely fashion, often marginally earlier than random

sampling. Even when mining for opinions, we find expert digest to be representative of the opinions found in the random digest. These properties of expert sampling make it a valuable methodology for generating content for several important content-centric applications. We discuss three example applications in the following text.

Topical search and breaking news detection: The tweet stream from experts can be used to search for timely information on topics, which can be as general as "politics" to as specific as "Australian Open." Extracting relevant information related to a topic from the hundreds of millions of tweets posted daily poses a formidable challenge. Since expert tweets are rich in topical content, expert sampling could provide a concise, easy-to-analyze digest that still captures much of the important information related to a topic. Also, breaking news about an unforeseen event may be available in the Twitter random sample only after a popular expert picks up the information (possibly from a nonexpert) and then posts it.

Trustworthy content recommendations: As we increasingly rely on crowd-sourced reports from Twitter about various events, trust becomes an important concern. Expert sampling offers a way to sample trustworthy tweets, which contain far fewer spam/malicious URLs compared to random digests. Note that the gain in trust is not achieved by relying on just a handful of experts—we are still collecting tweets from half a million users who have been indirectly vetted by the crowd and can provide varied views on diverse topics.

Opinion mining: The aggregated opinions/sentiments expressed by experts tend to mirror those mined from a random sampling of the entire Twitter crowd. However, the expert community tends to voice their opinion more strongly, that is, they tend to emote positive or negative sentiments more emphatically. When mining opinions regarding a topic, assertive opinions from experts may help to infer the overall sentiment more clearly and early.

Thus, our analysis of random versus expert sampling of data in large social networks reveals interesting trade-offs. Though expert sampling is more suitable than random sampling for data mining/information retrieval applications, random sampling also has certain advantages. Random sampling preserves the statistical properties of the entire data set, and automatically adapts to the growth and changes of the network, while expert sampling does not. For instance, expert sampling does not capture the large amounts of conversational tweets that are posted in Twitter because such tweets are deemed less important by experts. Furthermore, there are certain practical challenges in collecting the expert sample over long durations of time. For instance, the set of experts might require periodic recrawling as new experts join the social network or as old experts become inactive. Again, if the expert sampling methodology becomes popular, the sample could be spammed by malicious users creating fake Lists and infiltrating the set of experts. One potential way to prevent such infiltration would be to consider only trustworthy Lists (i.e., Lists created by trustworthy users) while identifying experts, similar to how algorithms such as Trustrank [Gyöngyi et al. 2004] are used to identify trusted websites in the Web-graph.

Given their relative merits, we believe that both random and expert sampling techniques would be needed in future, for example, employing hybrid sampling strategies which would enhance randomly sampled tweets streams with the views of the experts. Such approaches would again lead to additional issues such as figuring out the right balance between experts' content and randomly sampled content for the specific target application. However, as of today, most research and applications are centered only around random sampling, and we conclude by calling for equal focus on expert sampling of social network data.

ACKNOWLEDGMENTS

The authors thank the editors and the anonymous reviewers whose constructive suggestions helped to improve the article.

REFERENCES

Xavier Amatriain, Neal Lathia, Josep M. Pujol, Haewoon Kwak, and Nuria Oliver. 2009. The wisdom of the few: A collaborative filtering approach based on expert opinions from the web. In *Proceedings of ACM International SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*. ACM, New York, NY, 532–539.

Sebastien Ardon, Amitabha Bagchi, Anirban Mahanti, Amit Ruhela, Aaditeshwar Seth, Rudra Mohan Tripathy, and Sipat Triukose. 2013. Spatio-temporal and events based analysis of topic popularity in Twitter. In *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM'13)*. ACM, New York, NY, 219–228.

Sitaram Asur and Bernardo A. Huberman. 2010. Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE Computer Society, Washington, DC, 492–499.

Parantapa Bhattacharya, Saptarshi Ghosh, Juhi Kulshrestha, Mainack Mondal, Muhammad Bilal Zafar, Niloy Ganguly, and Krishna P. Gummadi. 2014. Deep Twitter diving: Exploring topical groups in microblogs at scale. In *Proceedings of ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW'14)*. ACM, New York, NY, 197–210.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3 (March 2003), 993–1022.

M. M. Bradley and P. J. Lang. 1999. Affective norms for english words (ANEW): Instruction manual and affective ratings. *Technical Report C-1, Center for Research in Psychophysiology, University of Florida* (1999).

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of International Conference on World Wide Web (WWW)*. ACM, New York, NY, USA, 107–117.

E. J. Candes and M. B. Wakin. 2008. An introduction to compressive sampling. *IEEE Signal Processing Magazine* 25, 2 (2008), 21–30.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and Krishna P. Gummadi. 2010. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM'10)*. AAAI Press.

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

Munmun De Choudhury, Scott Counts, and Mary Czerwinski. 2011b. Identifying relevant social media content: Leveraging information diversity and user cognition. In *Proceedings of ACM Conference on Hypertext and Social Media*. ACM, New York, NY, 161–170.

Munmun De Choudhury, Yu-Ru Lin, Hari Sundaram, K. Selcuk Candan, Lexing Xie, and Aisling Kelliher. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media? In *Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM'10)*. The AAAI Press.

Daantje Derks, Arjan E. R. Bos, and Jasper von Grumbkow. 2007. Emoticons and social interaction on the internet: The importance of social context. *Computers in Human Behavior* 23, 1 (2007), 842–849.

Eugene F. Fama. 1970. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance* 25, 2 (1970), 383–417.

Ove Frank. 1978. Sampling and estimation in large social networks. *Social Networks* 1, 1 (1978), 91–101.

Saptarshi Ghosh, Naveen Sharma, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. 2012a. Cognos: Crowdsourcing search for topic experts in microblogs. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 575–584.

Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. 2012b. Understanding and combating link farming in the Twitter social network. In *Proceedings of International Conference on World Wide Web (WWW'12)*. ACM, New York, NY, 61–70.

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 ACM International Conference on Conference on Information & Knowledge Management (CIKM)*. ACM, New York, NY, USA, 1739–1744.

Minas Gjoka, Maciej Kurant, Carter T. Butts, and Athina Markopoulou. 2010. Walking in Facebook: A case study of unbiased sampling of OSNs. In *Proceedings of IEEE Conference on Information Communications (INFOCOM'10)*. IEEE Press, Piscataway, NJ, 2498–2506.

Sandra Gonzalez-Bailon, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014. Assessing the bias in samples of large online networks. *Social Networks* 38 (July 2014), 16–27.

Catherine Grady and Matthew Lease. 2010. Crowdsourcing document relevance assessment with Mechanical Turk. In *Proceedings of NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT 2010)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 172–179.

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

Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. @spam: The underground on 140 characters or less. In *Proceedings of ACM Conference on Computer and Communications Security (CCS'10)*. ACM, New York, NY, 27–37.

Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. 2004. Combating web spam with trustrank. In *Proceedings of International Conference on Very Large Data Bases (VLDB) - Volume 30*. VLDB Endowment, 576–587.

Aniko Hannak, Eric Anderson, Lisa Feldman Barrett, Sune Lehmann, Alan Mislove, and Mirek Riedewald. 2012. Tweetin' in the rain: Exploring societal-scale effects of weather on mood. In *Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM'12)*. AAAI Press, Dublin, Ireland.

Liran Katzir, Edo Liberty, and Oren Somekh. 2011. Estimating sizes of social networks via biased sampling. In *Proceedings of International Conference on World Wide Web (WWW'11)*. ACM, New York, NY, 597–606.

W. Kellogg. 2006. Information rates in sampling and quantization. *IEEE Transactions on Information Theory* 13, 3 (2006), 506–511.

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

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

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

Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. 2012. TwiNER: Named entity recognition in targeted Twitter stream. In *Proceedings of International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, 721–730.

Jimmy Lin, Rion Snow, and William Morgan. 2011. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 422–429.

lists-howtouse. 2013. Twitter Help Center—Using Twitter Lists. Retrieved from https://support.twitter.com/articles/76460-using-twitter-lists.

Bing Liu. 2006. *Web Data Mining: Exploring Hyperlinks, Contents and Usage Data*. Springer-Verlag.

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

Fred Morstatter, Jürgen Pfeffer, and Huan Liu. 2014. When is it biased?: Assessing the representativeness of Twitter's streaming API. In *Proceedings of International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 555–556.

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's firehose. In *Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM'13)*. AAAI Press.

Xuan-Hieu Phan and Cam-Tu Nguyen. 2007. GibbsLDA++: A C/C++ Implementation of Latent Dirichlet Allocation (LDA). Retrieved from http://gibbslda.sourceforge.net/.

R. M. Poses, C. Bekes, R. L. Winkler, W. E. Scott, and F. J. Copare. 1990. Are two (inexperienced) heads better than one (experienced) head? Averaging house officers' prognostic judgments for critically ill patients. *Archives of Internal Medicine* 150, 9 (Sept. 1990), 1874–1878.

Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM'10)*. AAAI Press.

J. Romberg. 2008. Imaging via compressive sampling. *Signal Processing Magazine, IEEE* 25, 2 (2008), 14–20.

Paat Rusmevichientong, David M. Pennock, Steve Lawrence, and C. Lee Giles. 2001. Methods for sampling pages uniformly from the world wide web. In *Proceedings of the AAAI Symposium on Using Uncertainty within Computation*. AAAI Press, 121–128.

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

Jagan Sankaranarayanan, Hanan Samet, Benjamin E. Teitler, Michael D. Lieberman, and Jon Sperling. 2009. TwitterStand: News in tweets. In *Proceedings of ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS'09)*. ACM, New York, NY, 42–51.

Naveen Kumar Sharma, Saptarshi Ghosh, Fabricio Benevenuto, Niloy Ganguly, and Krishna Gummadi. 2012. Inferring who-is-who in the Twitter social network. *ACM SIGCOMM Computer Communication Review* 42, 4 (Sept. 2012), 533–538.

spritzer-gnip-blog. 2011. Guide to the Twitter API—Part 3 of 3: An Overview of Twitter's Streaming API. Retrieved from http://blog.gnip.com/tag/spritzer/.

Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. 2011. #TwitterSearch: A comparison of microblog search and web search. In *Proceedings of International ACM Conference on Web Search and Data Mining (WSDM'11)*. ACM, New York, NY, 35–44.

Kurt Thomas, Chris Grier, Vern Paxson, and Dawn Song. 2011. Suspended accounts in retrospect: An analysis of Twitter spam. In *Proceedings of ACM Internet Measurement Conference (IMC'11)*. ACM, New York, NY, 243–258.

A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM'10)*. AAAI Press, 178–185.

twitter-rate-limit. 2013. Rate Limiting—Twitter Developers. Retrieved from https://dev.twitter.com/docs/rate-limiting.

Twitter-stats. 2014. Twitter Statistics—Statistics Brain. Retrieved from http://www.statisticbrain.com/twitter-statistics/.

Twitter-stream-api. 2012. GET Statuses/Sample—Twitter Developers. Retrieved from https://dev.twitter.com/docs/api/1/get/statuses/sample.

Claudia Wagner, Vera Liao, Peter Pirolli, Les Nelson, and Markus Strohmaier. 2012. It's not in their tweets: Modeling Topical expertise of Twitter users. In *Proceedings of AASE/IEEE International Conference on Social Computing (SocialCom'12)*. 91–100.

Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Who says what to whom on Twitter. In *Proceedings of International Conference on World Wide Web (WWW'11)*. ACM, New York, NY, 705–714.

Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. 2012b. We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of International Conference on World Wide Web (WWW'12)*. ACM, New York, NY, 261–270.

Xintian Yang, Amol Ghoting, Yiye Ruan, and Srinivasan Parthasarathy. 2012a. A framework for summarizing and analyzing Twitter feeds. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 370–378.

Zhijun Yin, Liangliang Cao, Jiawei Han, Chengxiang Zhai, and Thomas Huang. 2011. Geographical topic discovery and comparison. In *Proceedings of International Conference on World Wide Web (WWW'11)*. ACM, New York, NY, 247–256.

Received February 2014; revised October 2014; accepted March 2015