On Sampling the Wisdom of Crowds: Random vs. Expert Sampling of the Twitter Stream

Saptarshi Ghosh
IIT Kharagpur, India
MPI-SWS, Germany

Muhammad Bilal Zafar
MPI-SWS, Germany

Parantapa Bhattacharya
IIT Kharagpur, India
MPI-SWS, Germany

Naveen Sharma
University of Washington, USA

Niloy Ganguly
IIT Kharagpur, India

Krishna P. Gummadi
MPI-SWS, Germany

ABSTRACT

Several applications today rely upon content streams crowd-sourced from online social networks. Since real-time processing of large amounts of data generated on these sites is difficult, analytics companies and researchers are increasingly resorting to sampling. In this paper, we investigate the crucial question of *how to sample the data generated by users in social networks*. The traditional method is to randomly sample all the data. We analyze a different sampling methodology, where content is gathered only from a relatively small subset ($< 1\%$) 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 diversity, timeliness, and trustworthiness of the information contained within them, and find important differences between the datasets. Our observations have major implications for applications such as topical search, trustworthy content recommendations, and breaking news detection.

Categories and Subject Descriptors: H.3.5 [On-line Information Services]: Web-based services; H.1.2 [User / Machine Systems]: Human information processing

General Terms: Experimentation, Measurement.

Keywords: Sampling content streams; Twitter; random sampling; sampling from experts; Twitter Lists.

1. INTRODUCTION

Crowd-sourced data generated in large social networking sites like Twitter and Facebook is valuable for a variety of data analytics applications ranging from content search and recommendations [5,9] to identifying breaking news [12]. However, sites like Twitter have several hundreds of millions of users tweeting many hundreds of millions of tweets every day [2], and the sheer volume of the entire tweet stream (known as the *firehose*) presents an enormous logistic problem for data analysts. Moreover, though some research studies [5,9] have used the *firehose*, very few organizations in the world have access to the *firehose*. So most analytics companies and researchers rely on sub-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 generated by the users in social networks?*

Today, most data analytics companies and researchers rely on *randomly* sampled tweets. Twitter supplies 10% randomly sampled tweets (known as the *gardenhose*) from its *firehose* for a fee, and 1% randomly sampled tweets for free. Random sampling is appealing for data analytics as the sampled tweets preserve the statistical properties of the global set of tweets, such as the fraction of tweets that are related to a given topic. Hence most studies have used a random sample of all tweets, e.g., to detect breaking news [12], and to map the content to various topical dimensions [11].

However, random sampling also preserves the large amount of unwanted information in the original tweets, such as spam and *non-topical*, 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 paper, we propose and analyze one such sampling methodology.

In contrast to random sampling, our new sampling method gathers content only from *topical experts*, i.e., Twitter users whose followers consider them to be knowledgeable on some topic. Recent studies have shown that topical experts are often the primary drivers of interesting discussions on Twitter [3]. 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 a good set of experts covering the wide range of topics that Twitter users are interested in. It is important to identify an extensive set of experts (covering popular, niche, local and global topics) to avoid being restrictive and falling prey to the long standing expert versus crowd debate [6]. We leverage a technique that we recently proposed [7, 13] to crowd-source ex-

CIKM'13, October 27–November 01, 2013, San Francisco, CA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2505615.

pert user detection in Twitter, to identify over half a million experts on a diverse set of topics of interest to Twitter users.

We gathered two samples of tweets – (i) the 1% randomly sampled tweet stream provided by Twitter, and (ii) tweets from the half million users who are identified as 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, trustworthiness, timeliness, and diversity (of sources and topics).

Note that there has been a long-standing debate on the expert-versus-crowd question as to which source of information is better in specific applications, such as understanding financial stocks [6]. However, there has not been any notable investigation on which is better for data mining / information retrieval applications such as topical search. Further, though several studies have used sampled content streams from Twitter, there has been little research on how the data streams obtained by different sampling approaches compare with each other. The only relevant study we know of is [10], which compared the full Twitter *firehose* with samples obtained through the "Twitter streaming API" Whereas [10] compares a sample with the full content stream, we compare two different methodologies of sampling content streams.

Our analysis reveals that the tweets sampled from experts are not only richer and more diverse in their information content, but also are more trustworthy, i.e., have fewer malicious URLs or spam content. Tweets sampled from experts also tend to capture breaking news stories a little earlier than randomly sampled tweets. We conclude that expert sampling is more efficient than random sampling for content-centric applications ranging from topical search / recommendation to breaking news detection.

2. METHODOLOGY

Our goal is to collect and compare tweets obtained through 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.

2.1 Sampling methodologies

Random sample: As mentioned earlier, the most commonly used methodology of sampling tweet streams is random sampling. We used the publicly available Twitter streaming API to gather the 1% random sample of all tweets provided by Twitter [14] over the month of December 2012.

Sampling from experts / popular users: Another possible methodology is to sample content from the experts / most popular users. To identify the experts / most popular users in Twitter, we started a long-running profile 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, we were able to crawl data for the first 50 million Twitter users. The experts / most popular users were identified from among these 50 million users, as described below.

There are several metrics to rank users in Twitter and hence to identify popular users [4, 7]. 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 whose data we could gather, using each of the above metrics, and found that there is a significant overlap among the top-ranked users according to the various metrics. For instance, there is 68.2% overlap between the top 500,000 users (i.e., top 1% of 50 million) 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).

Topic	Experts identified by List-based method
Music	Katy Perry, Lady Gaga, Justin Timberlake, coldplay, P!nk, Marshall Mathers
Physics	Institute of Physics, Physics World, Fermilab Today, CERN, astroparticle
Neurology	Neurology Today, AAN Public, Neurology Journal, Oliver Sacks, ArchNeurology
Environment	GreenPeace USA, NYTimes Environment, TreeHugger.com, National Wildlife,

Table 1: Examples of topical experts for specific topics, as identified by a List-based methodology proposed in our prior work [7, 13].

We decided to use the List-rank metric where users are ranked according to the number of Lists in which they are included. The Lists feature in Twitter is used to group together experts on common topics [1]; for instance, a user interested in music may create a List named 'Music' and add popular musicians like 'BritneySpears' and 'LadyGaga' as members of the List. Using Lists to identify popular users has an additional advantage – as we have shown in our prior studies [7, 13], List names and other meta-data can be used to infer the topical expertise of the members of the Lists. Similar to our prior study [7], 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 1 shows some example topics and some of 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 like any other method for identifying experts, the above methodology also has a few limitations. The primary one is that the set of identified experts is limited to users who joined Twitter early, and is hence biased towards certain countries where Twitter first became popular (e.g., the USA). However, our objective 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.

2.2 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 *random-digest* and *expert-digest* respectively. Table 2 gives the number of tweets and the number of distinct users who posted the tweets in both the digests, over three time durations – a day (December 3, 2012), a week (December 3 – 9, 2012) and the entire month of December 2012. In each of the

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
Sized random digest – subsampled	2,264,904	1,930,045	15,517,042	9,105,185	63,497,081	21,941,041

Table 2: Number of tweets and distinct users who tweeted in the three digests – the random 1% digest, the expert digest, and a sub-sampled random 1% digest containing the same number of tweets as the expert digest – across three different durations – a day, a week, and a month.

three durations, the random digest has about 1.8 times as many tweets as the expert digest. Put differently, our expert digest contains about 0.55% of all tweets posted on Twitter. To enable a fair comparison between the two digests, we (randomly) subsampled the random digest to contain the same number of tweets as the expert digest over each of the three durations; the statistics for the subsampled random digest are also shown in Table 2. In the rest of the paper, we always compare the expert digest with the similarly sized random digest.

2.3 Metrics to compare Data Samples

Our goal is to compare the expert and random digests to ascertain their utility for applications such as topical search and recommendation, breaking news detection, and so on. Hence we consider the following dimensions / metrics for comparing the tweet-samples.
(i) **Sources** – do the samples contain information posted by only the elite users, or also the voices of the crowd?
(ii) **Quality** – do the samples contain mostly conversational babble or useful information?
(iii) **Diversity** – what are the various topics which are covered by the samples?
(iv) **Trustworthiness** – what is the amount of spam (e.g., blacklisted URLs) included in the samples?, and
(v) **Timeliness** – how quickly is the information of a recent, unforeseen event (e.g., an accident or natural calamity) available in the samples?

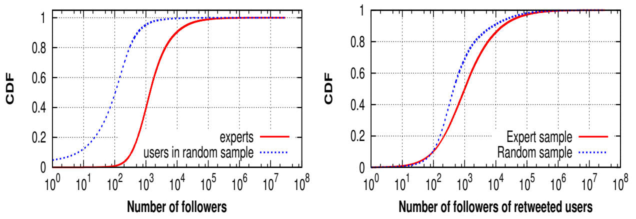
3. SOURCES OF CONTENT

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 (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?

3.1 Comparing sources of tweets

Comparing the number of users in the two digests (see Table 2), it is clear that the random digest includes tweets from a substantially larger (by one to two orders of magnitude) population of users than the expert digest. There is relatively little overlap between the users included in the two digests – less than 35% of the users in the expert digest are included in the corresponding random digest. Also, less than 0.15% of the tweets are common between the two digests.

Next, we compare the characteristics of the users encountered in the two digests. Figure 1(a) shows the distribution of the number of followers for the two sets of users. It is evident that the expert and random digests draw their tweets from very different Twitter user populations – compared to users in the random digest, experts are considerably more popular in Twitter, having significantly more followers.



(a) Users whose tweets are directly included **(b) Original sources of retweets**

Figure 1: Distribution of number of followers of users whose tweets are included in the random and expert digests: (a) users whose tweets are directly included, (b) original sources of retweets.

A direct consequence of the higher popularity of the experts is that the *experts' tweets are also significantly more popular*. The popularity of a tweet can be estimated by its *retweet-count*, i.e., the number of times the tweet is retweeted in the global Twitter network. We observe that a tweet in the expert digest has, on average, 6.6 times higher retweet-count as compared to a tweet from the random digest (details omitted for brevity).

3.2 Comparing original sources of retweets

The differences between user populations in the expert and random digests, while not surprising, raises a potential concern that by sampling tweets only from experts, one might miss useful and interesting information that is tweeted by the masses of non-expert users in Twitter. To check if this concern holds, we analyzed the *retweets* in the expert digest to see whether experts were retweeting (reposting / forwarding) only tweets from other popular Twitter users, or whether they are retweeting interesting information posted by ordinary users as well.

Table 3 shows the fraction of tweets in the two digests that are retweets. We see that both digests contain nearly the same fraction of retweets (18% in the expert digest and 21.6% in the random digest), which suggests that experts are also forwarding tweets from other users at the same rate as users in the random digest. Figure 1(b) shows the distribution of popularity (measured as number of followers) of users whose tweets were retweeted by experts. For comparison, we also plot the distribution of popularity of users whose tweets were retweeted by users included in the random digest. The popularity distributions for the two digests look fairly similar, implying that the experts are not limiting their retweets to their fellow experts. Rather, they retweet interesting information from both expert and non-expert users, just as a random Twitter user would retweet. In fact, our data suggests that experts themselves are interacting with non-expert crowds within Twitter, and they might be filtering / forwarding information from the crowds that they deem interesting and useful.

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

Table 3: Comparison of information content (retweets, hashtags, URLs) in the tweet-digests, collected during a day. In this section, the expert digest is compared with a similarly sized random digest (subsampled version of the Twitter 1% digest, which has the same number of tweets as the expert digest).

Category	Random digest	Expert digest
Conversational / babble	82.11 (11.5)	40.64 (22.4)
Opinion / Sentiment	8.42 (37.5)	10.16 (31.6)
Advertisement	1.05 (100)	5.88 (90.9)
Spam	1.58 (0.0)	0 (0.0)
Topical information	6.84 (53.9)	43.32 (90.12)

Table 4: Categories of tweets, as judged by evaluators in a survey conducted through AMT. The first number shows the percentage of tweets 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.

4. QUALITY & DIVERSITY OF CONTENT

In this section, we compare the quality and diversity of the information contained in the two digests. 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?

4.1 Comparing content quality in the digests

We first study what fraction of tweets in the two digests contain useful information. Judging whether a tweet contains useful information is inherently subjective. Hence, we judged the quality of information in the tweets through human feedback, using the Amazon Mechanical Turk service (AMT) where human volunteers judged the nature / topic of the tweets using a web-based feedback service.

For the AMT survey, we selected at random 200 English-language tweets each from the expert-digest and the random-digest. During the survey, each AMT worker was shown a tweet and 5 categories – conversational tweet, tweet containing sentiment / opinion but no topical information, advertisement, spam, and tweet containing useful topical information – and was asked to judge under which category the tweet falls. Each tweet was judged by at least 3 different workers. Table 4 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 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 4 also shows the percentage of tweets in each category that contained hashtags or URLs. We see that a large majority of the tweets that were judged to contain useful topical information 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.

Topic	Random digest	Expert digest
Entertainment	40.00	6.153
Sports	30.00	15.38
Science & Technology	20.00	15.38
Lifestyle & Culture	10.00	13.85
Government & Politics		16.92
Business & Economy		18.46
Education		3.08
Arts & Humanities		3.08
Environment & Weather		4.62
Health & Fitness		3.08

Table 5: Topics of the tweets which were judged (in AMT survey) to contain topical information. The numbers show the percentage of tweets in the two digests.

Table 3 also compares the number of hashtags and URLs contained in the expert digest and a similarly sized random digest (described in Section 2) gathered on December 3, 2012. 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.

4.2 Comparing topical diversity in the digests

In the AMT survey described above, whenever a worker judged a tweet to contain topical information, the worker was also asked to indicate the topic of the tweet from among a specified set of 10 topics, such as politics, sports, education, and so on (the complete list of topics is given in Table 5).[1] The results are summarized in Table 5. Among the tweets which were judged to contain topical information, 40% of such tweets in the random digest were related to *entertainment*, i.e., music, movies, and so on. Moreover, only four of the specified set of 10 topics were represented in the tweets from random digest. In contrast, the tweets in the expert digest covered all the specified topics, clearly showing that tweets in the expert digest cover a much wider diversity of topics as compared to the random sample.

4.3 Comparing diversity in popular content

In the analysis described above, we studied the quality of the information contained in the *entire* expert and random digests. However, the most important content in a digest is the content that is most popular; hence, we now study how similar or different is the popular content in the two digests.

Our findings reported above suggest that hashtags and URLs are the primary indicators of meaningful topical content in the digests. Hence, we focus on comparing the popular hashtags and URLs contained in the two digests. Below we only report on our results for popular hashtags, but the results for popular URLs were similar.

[1] We derived this specified set of topics from the Yahoo category directory (`dir.yahoo.com/` and `answers.yahoo.com`).

Top 1,000 hashtags in expert digest which are ranked beyond 10,000 in random digest		Top 1,000 hashtags in random digest which are ranked beyond 10,000 in expert digest	
Theme	**Example hashtags**	**Theme**	**Example hashtags**
Literature	nanowrimo (national novel writing month), aca, amreading	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
Ongoing events	leweb12 (conference of web technology business leaders)	Ongoing events	niallspanish, goodluck1d (all related to concert of One Direction band)
Technology	TrustCloud, IBM, opensource, healthIT	Games	farmville, iphonegames, androidgames
Business	b2b, crm, smallbusiness, custserv	Entertainment	LoveInParis (television series)

Table 6: Examples of top 1,000 hashtags in one digest, which are ranked beyond the top 10,000 in the other digest.

We ranked the hashtags according to the number of distinct users who have posted the hashtags in each digest. We observed that more than 90% of the top 1000 hashtags in either digest is contained in the other digest as well. Thus, we see a very high overlap between the most popular content in expert and random digests.

However, some of the top hashtags in one digest are ranked much lower 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 1000 in one digest, but ranks beyond 10,000 in the other digest. Among the top 1,000 hashtags in the expert digest, 54 appear beyond the top 10,000 in the random digest. Similarly, 170 hashtags among the top 1,000 in the random digest, appear beyond the top 10,000 in the expert digest. A sample of these hashtags have been shown in Table 6. We note that the hashtags from the expert digest (which are 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 not popular in the expert digest, are mostly related to acquiring followers in Twitter, Twitter-specific memes, and games. Intuitively, the experts who are well known in their respective fields, would have lesser interest in these topics.

Hence, we conclude that the expert digest contains significantly more useful topical information on a more diverse set of topics, as compared to the random digest. On the other hand, the popular content in the random digest (which is not popular in the expert digest) is largely about Twitter-specific memes and methods of acquiring followers in Twitter, which are of limited topical interest. Hence, in general, the expert digest shows much better promise for information retrieval applications, such as topical search or recommendations.

5. TRUSTWORTHINESS & TIMELINESS

As crowdsourced data from Twitter is increasingly used for purposes such as identifying breaking news, some of the major concerns are the trustworthiness and timeliness of such data. In this scenario, the specific questions we address are (i) which digest contains larger amount of spam (e.g., blacklisted URLs)?, and (ii) when an unforeseen event occurs (e.g., a natural calamity or an accident), which digest gives the earliest information about the event?

5.1 Malicious content and unvetted users

We start by analyzing the amount of malicious content (blacklisted URLs) and unvetted users in the two digests.

Malicious URLs in the two digests: We check whether an URL in a tweet is malicious 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 interstitial warning pages for malicious URLs. We check for such pages in case the URL is a shortened one. Additionally, we check the final landing URL using the Google Safebrowsing API [2].

We randomly selected 1 million URLs each from the two digests, and checked how many of them were blacklisted. We found 1520 blacklisted URLs in the random digest, which were posted by 1,447 (i.e., 0.140%) distinct users. On the other hand, we found only 129 blacklisted URLs posted by 46 (i.e., 0.022%) 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: Of the 1,447 users who have posted at least one blacklisted URL in the random digest gathered in December 2012, 501 (i.e., 34.6%) were suspended by Twitter till January 16, 2013. Further, out of these 501 suspended users, 468 (i.e., 93%) 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 significant fraction of the remaining accounts (which have not been suspended by Twitter) have posted more than 100 blacklisted URLs in their lifetime, implying that it is very likely that these users are actually spammers whom Twitter has not been able to identify.

In contrast, *none* of the 46 experts who had posted a blacklisted URL have been suspended, suggesting that the expert users may have posted the few blacklisted URLs unwittingly. To confirm this hypothesis, we analyzed the historical tweet data for these 46 experts. We found that 32 of them had posted exactly one blacklisted URL. The rest of the experts 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. In spite of 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 very slow in identifying blacklisted URLs [8]. Hence an added advantage of sampling tweets from experts (who are well-established users) is that they are very less likely to post spam.

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

Event	Extracts from first tweet in expert-digest	Extracts from first tweet in random-digest
(E1) Singer Jenni Rivera dies in air-crash	**(18:21:14)** *Rodpac*: IvanSamuelLuz: Rodpac: Se estrella el aviÃşn de @jennirivera, salii de Monterrey a la 1 con destino a Toluca.	**(18:22:52)** *Gusmer1*: RT Rodpac: Se estrella el aviÃşn de @jennirivera, sali de Monterrey a la 1 con destino a Toluca.
(E2) Shooting at Sandy Hook Elementary School	**(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
(E3) Nigerian politicians die in helicopter crash	**(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 ...

Table 7: 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. The timestamps are according to Central European Time.

5.2 Timeliness of information

Finally, we analyze the timeliness or recency of information, i.e., when an unforeseen event occurs, which digest gives the earliest information about the event. For this, we considered three such events which occurred in December 2012 – (E1) Mexican-American singer Jenni Rivera died in an air-crash on December 9, (E2) Shooting at Sandy Hook Elementary School in Connecticut, USA on December 14, and (E3) Nigerian State governor Patrick Yakowa died in a helicopter crash on December 15.

For each event, we identified the relevant tweets that appeared in the expert-digest and the random-digest immediately after the event occurred, through keyword-based search.[3] Table 7 gives 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.

For each of these events, *the earliest relevant tweet appeared in the expert-digest*. For E1, the first tweet in the random-digest appeared within two minutes of the first tweet in the expert-digest; however, the delay was longer (more than 10 minutes) for the events E2 and E3. More interestingly, Table 7 shows that in each case, the first tweet in the random-digest was a retweet of a tweet posted by an expert or a popular user. Specifically for the events E1 and E3, the first tweet in the random-digest is a direct retweet of the first tweet in the expert-digest.

These observations can be explained by the fact that the set of experts includes a number of media sites / journalists who usually post the earliest tweets about any event. These tweets immediately get retweeted by a large number of users who follow these media sites / journalists, and hence appear in the Twitter random sample after some time when a sufficiently large number of users have retweeted the tweet.

6. CONCLUDING DISCUSSION

The primary contribution of this paper lies in investigating an alternative strategy for sampling content streams generated in OSNs, which is different from the universally used random sampling. Our analysis of expert vs random sampling reveals that the experts' tweets are significantly richer in information content (whereas close to 90% of the random sample is devoid of topical information), cover more diverse topics, and more popular content. Experts' tweets are also more trustworthy (contain much less spam) and they often capture breaking news stories marginally earlier than random sampling. These properties of expert sampling make it a valuable methodology for generating content for several important content-centric applications, such as topical search, trustworthy content recommendation, breaking news detection, and so on.

On the other hand, random sampling preserves certain important statistical properties of the entire data set, which expert sampling does not. For example, expert sampling does not capture conversational tweets that might be deemed as less important by experts. Given their relative merits, we conclude by calling for equal focus on random and expert sampling of social network data.

Acknowledgment: This research was supported in part by a grant from the Indo-German Max Planck Centre for Computer Science (IMPECS).

7. REFERENCES

[1] Twitter Help Center: How to Use Twitter Lists. `http://tinyurl.com/UseTwitterLists`.

[2] Twitter now averaging 400 million tweets daily. `http://tinyurl.com/TweetsPerDay`, Jun 2012.

[3] S. Ardon et al. Spatio-Temporal Analysis of Topic Popularity in Twitter. *arXiv:1111.2904 [cs.SI]*, 2011.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proc. ICWSM*, May 2010.

[5] M. D. Choudhury, S. Counts, and M. Czerwinski. Find me the right content! diversity-based sampling of social media spaces for topic-centric search. In *Proc. ICWSM*, 2011.

[6] E. F. Fama. Efficient capital markets: a review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

[7] S. Ghosh, N. Sharma, F. Benevenuto, N. Ganguly, and K. Gummadi. Cognos: crowdsourcing search for topic experts in microblogs. In *Proc. ACM SIGIR*, 2012.

[8] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *Proc. ACM CCS*, 2010.

[9] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proc. ACM SIGKDD*, 2011.

[10] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proc. ICWSM*, 2013.

[11] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *Proc. ICWSM*, 2010.

[12] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: news in tweets. In *Proc. ACM SIGSPATIAL Conf. on Advances in Geographic Information Systems*, 2009.

[13] N. Sharma, S. Ghosh, F. Benevenuto, N. Ganguly, and K. Gummadi. Inferring Who-is-Who in the Twitter Social Network. In *Workshop on Online Social Networks*, 2012.

[14] Limit on Streaming Tweets | Twitter Developers. `https://dev.twitter.com/discussions/6789`.

[3] For instance, keywords such as 'Sandy Hook', 'Sandyhook', 'shooting' were used for the incident related to shooting at Sandy Hook elementary school.