The Impact of Sampling on Big Data Analysis of Social Media: A Case Study on Flu and Ebola

Kuai Xu, Feng Wang
Arizona State University
{kuai.xu, fwang25}@asu.edu

Xiaohua Jia
City University of Hong Kong
csjia@cityu.edu.hk

Haiyan Wang
Arizona State University
haiyan.wang@asu.edu

***Abstract*—The explosive growth of online social networks in recent years have generated massive amount of data-sets in user behaviors, social graphs, and contents. Given the scale, heterogeneity, and diversity of such *big* data, sampling becomes a simple and intuitive approach to reduce the size of the data-sets for collecting, measuring, and understanding users, behaviors and traffic in online social networks. In this paper, we quantify the impact of random sampling on the analysis of online social networks with Twitter streaming data as a case study. In addition, we design different sampling strategies including community sampling and strata sampling, and evaluate their impact on a broad range of behavioral characteristics of online social networks. Our experimental results show that community sampling has the minimum impact on tweet distributions across users and the structure of retweeting graphs, while achieving the similar data reductions as random and stratified sampling.**

I. Introduction

The explosive growth of social media such as Twitter and Sina Weibo have shifted the communication paradigms of today's society. The unprecedented data, which are continuously generated from millions of active users, leads a number of challenges in data acquisitions, analysis and visualizations. Sampling is a widely used technique in big data analysis and social science where collecting and making sense of data from the entire population or data-sets is infeasible or too expensive.

Towards this end, we first characterize the impact of sampling on Twitter streaming data with Ebola and flu as case studies from a variety of perspectives including tweet volumes, tweet distributions, and user influence. Specifically, we compare the volumes of Ebola and flu tweets between sampling data-sets and full data-sets, and discover that the time granularity and topic popularity play significant roles on the sampling impact. In addition, our analysis reveals that the sampling process alters tweet distribution across users, thus affecting the quality and efficiency of the algorithms that rely on tweet distributions such as spammer detection. Finally, we unveil the significant impact of random sampling on user influence and the largest connected components of reweetting graphs.

In light of the significant impact of sampling on tweeting volume, tweet distributions, and user influence, we design stratified sampling and community sampling in addition to random sampling for evaluating the benefits and limitations of different sampling strategies. Our stratified sampling process first divides the user populations of Twitter based on a demographic feature such as geographical locations and languages, and then sample the tweets posted by the users of each subpopulation independently to include the tweets from all subpopulations in the final sampled data-sets. In contrast to stratified sampling, community sampling selects a particular set of communities from Twitter user population and only samples tweets from these pre-determined set of communities. For these three sampling design strategies, i.e., random sampling, stratified sampling and community sampling, our sampling process calculates the sampling rate for ensuring the sampled data-sets have the same numbers of tweets.

Our experimental results based on these sampling design strategies with real Ebola and flu tweets show that community sampling retains the tweet distributions across users in a similar fashion as full data-sets, while the other two sampling approaches are unable to capture the original tweet distributions. In addition. community sampling preserves larger connected components of retweet graphs than the others, reflecting the common interest of the users of the same community on the same hot topics or latest events. Based on our analysis, we recommend to use community sampling to preserve important characteristics of the original tweet data-sets while significantly reducing massive amount of tweet data for analysis.

The contributions of this paper are multi-fold as follows:

- We shed light on the impact of sampling on social media analysis of tweet volume, tweet distributions, information diffusion, and user influence.
- We develop big-data-aware sampling designs and strategies for social media analysis to reduce the size of tweet data for analysis and more importantly to retain the key characteristics in the original tweet data-sets.
- We evaluate the performance and cost different sampling designs with large scale social media data-sets and discover that community sampling has the minimum impact on tweet distributions across users and the structure of retweeting graphs, while achieving the similar data reductions as random sampling and stratified sampling.

II. Data Collection

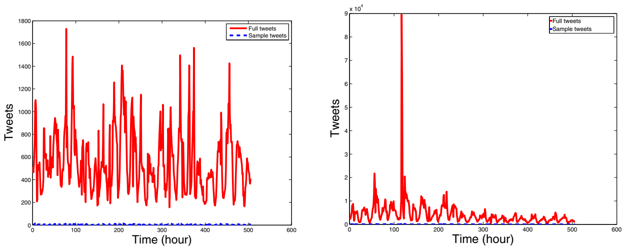
To characterize the impact of sampling on social media analysis, we leverage the public data streams via Twitter *sample* streaming API [1], which provides a random 1% of all public tweets (also called statuses). In addition, the

978-1-4799-5952-5/15/$31.00 ©2015 IEEE

filter streaming API could return all public tweets which match a small set of pre-defined keywords, user identifications (userids), or geographical locations.

Given the large set of important topics and influential users on Twitter social media, it is impractical to track all these topics and users. Instead, we choose two keywords: Flu and Ebola that are of interests to the general public due to the repeated seasonal influenza (flu) epidemiology and the Ebola virus epidemic in West Africa in 2014. Thus we continuously harvest every tweet mentioning “flu” or “Ebola”, which provides the full visibility on what Internet users talk about flu and Ebola on Twitter. We refer to the tweet streams collected via the *filter* streaming API as *full flu tweets* and *full Ebola tweets*. Our data collection starts since October 1, 2014, and currently continues to collects both data-sets.

To contrast the sampling and full tweet streams, we extract all tweets containing any of two keywords “flu” and “ebola” from the sample tweet streams, and refer to these two tweet streams as “sample flu tweets” and “sample Ebola tweets”. Clearly, the sample flu/Ebola tweets are a subset of full flu/bola tweets. Figure 1[a][b] show the volumes of flu tweets and Ebola tweets during the same week of 2014/10/02 - 2014/10/09, respectively. As shown in both figures, Twitter users actively share news stories on flu and Ebola and discuss the latest updates on the epidemics on the social media. The in-depth investigation reveals that the significant spike in Figure 1[b] are caused by massive retweets and comments when Brazil announced the first suspected Ebola case on October 10, 2014.



(a) Flu tweets

(b) Ebola tweets

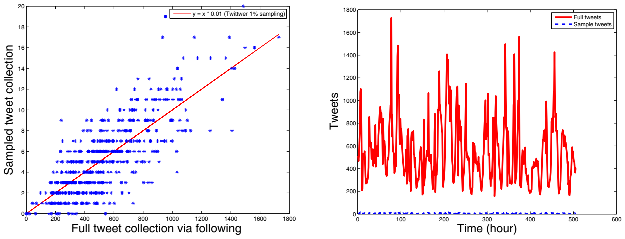
Fig. 1. The data collection of flu and Ebola tweets via random sampling and via following.

III. Quantifying Impact of Tweet Sampling

A. Sampling Impact on Tweeting Volumes

To study the impact of sampling on tweeting volumes, we first compare the projected tweeting volume based on the sampling principles with the observed tweeting volumes. Figure 2[a] shows the projected tweeting volumes on flu illustrated by the red line during October 2014, and the sampled and full tweet volumes on flu illustrated by the scatter plot, during the same time period. Based on the random sampling principle, the projected tweeting volume on flu, n, can be simply calculated as $n = N * p$, the product of the total number of full tweets on flu, N, and the sampling ratio, p, which is set as 1% in Twitter sampling streaming APIs.

As shown in Figure 2[a], it is very interesting to see that the actual observed tweeting volumes on flu for each hour deviates significantly from the projected volumes. In other words, the inference of full tweet volumes based on sampled tweet streams could potentially lead to inaccurate estimations. Figure 2[b] shows the inferred volumes on flu tweets and the actual full tweets on flu, confirming this conjecture.

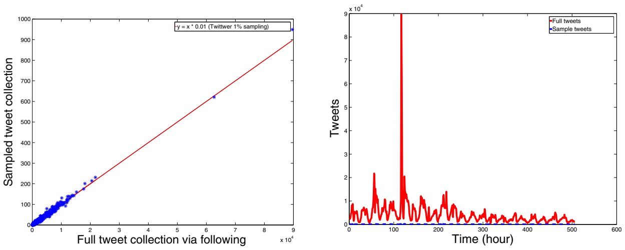


(a) Projected sampling vs. observed sampling tweets on flu

(b) Inferred vs. actual full tweets

Fig. 2. The impact of sampling on flu tweet volumes.

We run the similar analysis on Ebola tweets during the same time period. As illustrated in Figure 3[b], the projected sampling tweets and the observed sampling tweets on Ebola are very similar, thanks to the large volumes of tweets on Ebola during October 2014 when millions of people tweeted, retweeted, and commented on the latest statues of Ebola outbreaks on social media. As a result, the inferred and actual Ebola full tweets are also very close, reflecting the insignificant impact of random sampling on hot topics and latest events on social media.



(a) Projected sampling vs. observed sampling tweets on Ebola

(b) Inferred vs. actual Ebola full tweets

Fig. 3. The impact of sampling on tweet volumes at different time granularity.

In summary, our experimental results show that the impact of sampling on tweeting volumes is significant especially for less popular topics. In addition, we find that the impact differs for different time granularity. Specifically, the smaller time granularity, the larger the impact.

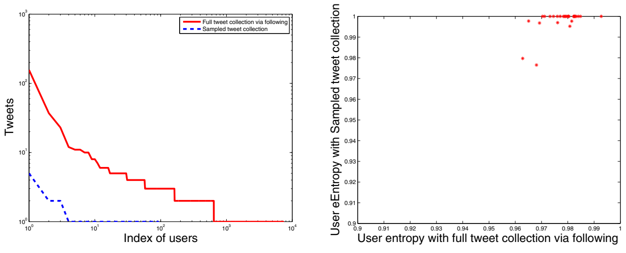
B. Impact on Tweet Distributions Across Users

The process of random sampling on tweet streams not only affects tweeting volumes, but also changes tweet distributions across users. Figure 4[a] shows tweet distributions across users who have tweeted flu during a 24-hour time window. Clearly, the distributions for full data-set and sampled data-set are very different, since the sampling process drops 99% tweets.

To quantify the difference of tweet distributions, we apply entropy concepts from information theory [2] to measure the distributions of tweeting activities on flu or Ebola across all users, say U_t and $|U_t| = m$, who have tweeted a given topic t during a given time window. Let x_i denote the number of tweets by a user $u_i \in U_t$ during the time period, and the total number of flu tweets n then becomes $n = \sum_{i=1}^{m} x_i$. Thus the probability of tweets from a given user $u_i \in U_t$ can be derived as $p(i) = x_i/n, x_i \in U_t, i = 1, 2, \cdots, m$, where m is the number of the unique users who have tweeted on a topic during the time window. The entropy of quantifying tweet distributions from the set of users is $H(X) = -\sum_{x_i \in U_t} p(x_i) \log p(x_i)$, while the standardized entropy is $H'(X) = \frac{H(X)}{H_{max}(X)} = \frac{H(X)}{\log m}$, where $H_{max}(X) = logm$ is theoretically the maximum value for $H(X)$.

The values of entropy and standardized entropy measure the randomness or uniqueness of tweet distributions across users who have tweeted on the same topic. If every user tweets once on a topic, e.g., flu, then the standardized entropy value becomes 1. On the other hand, if all the tweets are posted by the same user, the standardized entropy is calculated as 0. Thus comparing the standardized entropy of tweet distributions with sampling data-set and with the full data-set can effectively reveal how the sampling process change the underlying tweet distributions for users who have tweeted on hot topics such as flu or Ebola.

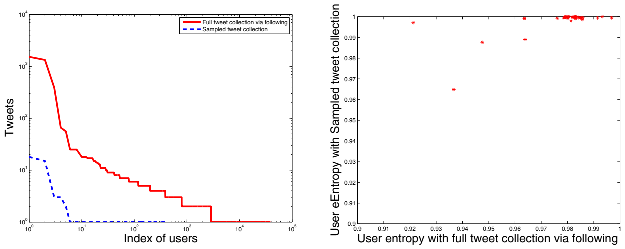
Figure 4[b] shows a scatter plot of standardized entropies for users with full data-sets as well as sampled data-sets. Apparently, the user entropies in the sampling data-sets tend to be higher than those of full tweet collection via following. Upon close examinations, we find that the sampling process does not include many tweets for users who have tweeted multiple tweets, which shifts uneven tweet distributions in the original full data-set towards a more random fashion, leading to higher entropies.



(a) Tweet distributions of users tweeting flu in one day
(b) Entropy of users tweeting flu in one day

Fig. 4. Tweet distributions of users tweeting flu in one day.

Figure 5 shows similar observations on tweet distributions for users who tweeted Ebola during the same time window. Our observations of sampling impact on tweet distributions have very important implications. For example, spammer detection algorithms, which rely on correlating multiple tweets from the same spammers, might not have sufficient data points for analysis, thus are unable to effectively detect spammers in social media.

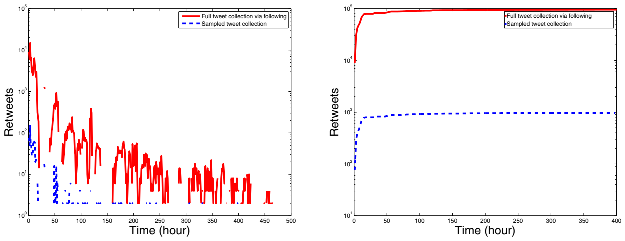


(a) Tweet distributions of users tweeting Ebola in one day
(b) User entropy of tweeting Ebola in one day

Fig. 5. Tweet distributions of users tweeting Ebola in one day.

C. Impact on User Influence

Next we characterize the impact of random sampling on information diffusion, user influence, and retweet graphs. Figure 6[a] illustrates the time-series of retweeting activities for one Ebola tweet over time for full retweeting data-set and for random sampling data-set, respectively. As shown in Figure 6[a], the time-series pattern of retweeting activity after random sampling does not fully capture the original pattern of the full data-sets for the same tweet. Similar observations hold for the retweeting behaviors of other tweets as well. On the other hand, the cumulative statistics on the retweeting activities for random sampling data-sets is very similar to that of full data-set, as reflected in Figure 6[b]. In addition,



(a) Retweets over time
(b) Cumulative retweets over time

Fig. 6. The impact of sampling on information diffusion.

We characterize the impact of sampling on user influence by analyzing the retweeting graphs, which are built via extracting the user information from the retweets of a given tweet. For example, if a user A retweets a given retweet from the user B, we consider a directed edge $A \rightarrow B$ between two vertices A and B in the directed retweet graph. As B's tweet is being retweeted, we consider B as one of the influencers. It is important to note that a retweet user might not be an influencer if no users have retweeted any of his or her retweets.

Table I lists the impact of sampling on retweet statistics such as retweet users, influencers, user influence quantified by the number of retweeting messages, and largest connected component (cc) of retweet graphs for 5 selected flu and Ebola tweets. The retweet user of sampling data-set is approximately

1% of the full data-set, while the influencer of sampling data is much larger than 1%.

The sampling impact on user influence and largest connected component of retweet graphs is significant. Across all 5 tweets, the average influences in the sampled data-set is much smaller than that of full retweeting data-set. More importantly, the retweet graphs in the sampled data-set are disconnected due to the process of random sampling which do not include all the edges of retweeting relationships, thus leading to many subgraphs in the retweet graphs. The largest connected component for retweet graphs of these 5 tweets derived from sampling data-sets only account for 15% - 52% of all nodes in that graphs, while the retweets graphs built from the full data- set are connected, thus the largest connected components are the retweet graphs themselves.

IV. Sampling Designs

A. Random Sampling

Random sampling is a widely used straightforward sampling technique. Assuming a random sampling rate of p, any tweet from Twitter data streams has an equal probability p of being selected into the samples. In the case of Twitter streaming API, p is set as 0.01 by Twitter API service.

To verify the sampling rate of Twitter streaming API, we perform a simple random sampling with the same sampling rate on the full tweet streams we have collected on flu and Ebola. Our experimental results show that the characteristics such as tweet volumes, user distributions, and temporal patterns from both our sampling process and Twitter streaming API are very similar. The only difference lies in that our offline sampling process has the ability of choosing any sampling rate for studying the benefits and tradeoff of the random sampling rate.

B. Stratified Sampling

Online social networks typically have a very diverse set of users with different demographics, e.g., ages, genders, educations, incomes, geographical locations. A potential limitation of simple random sampling is that certain demographics, e.g., geographical locations with a relatively small number of users might not be observed in the data samples due to a low volume of tweets in the data streams.

Towards this end we consider to use stratified sampling technique [3] as one alternative sampling strategy to sample Twitter data streams. Specifically, we divide Twitter users based one demographic attribute, i.e., geographical location, into different *stratum*. For each stratum, we employ the same sampling rate for randomly choosing tweets. However, unlike from simple random sampling, our stratified sampling process considers the size of each *stratum* during the selection process in order to ensure the inclusion of data samples from small stratums as well.

Given the size of twitter data streams N, assuming a sampling rate p, a sampling process will lead to the sampled data set $n = N * p$. Let S denote the set of *stratums* of all Twitter data streams with each stratum clusters tweets posted by users with the same demographic attribute. For a stratum with m tweets such that $m < \frac{1}{p}$, it is likely that none of tweets in this small stratum is selected with the probability of p. To ensure the complete coverage of all stratum, our stratified sampling process will randomly select q tweets from such stratums, denoted as S_1. Let s_i, where i = 1, 2, ..., $|S_1|$ denote the size of each stratum in S_1.

Let S_2 denote the set of stratums, each of which has at least m' tweets such that $m' \geq \frac{1}{p}$ tweets. Thus, the total number of tweets from S_1 stratums is $q * |S_1|$. Note that if we use the same probability p to select tweets from S_2, the final sampled tweets will be large than n due to the stratified sampling on S_1. Thus, we reduce the sampling rate on S_2 to p' such that $(N - \sum_{i=1}^{|S_1|} s_i) * p' = n - q * |S_1|$. In other words, the total number of sampled tweet streams remains the same n, which is the same as simple random sampling process.

C. Community Sampling

As illustrated in the previous section, the impact of sampling on the retweet graphs shows that many retweet activities, represented as edges in the retweet graphs are not captured due to the random sampling process. In this paper, we propose to explore community sampling, a non-probabilistic sampling approach, to preserve retweeting activities and user interactions during the sampling process.

The essential idea of community sampling is to focus on the entire social communities from Twitter users, e.g., users from a certain city during the sampling process. In particularly, community sampling selects tweet streams posted by all users in one or a few communities with a much higher probability, e..g, $p >> 0.01$.

Let C denote the set of all communities of Twitter users, which are sortied in a non-increasing order based on the number of tweets posted by the users in the same community. The community sampling process first searches communities with at least n tweets where $n = N * 1\%$, and then locates the smallest community, i.e., C_j. Lastly, we sample tweets from C_j with a probability of p such that $|C_j| * p = n$.

If none of the communities has a size over n, we locate the first set of k communities C_1, C_2, ..., C_k such that $\sum_{i=1}^{k-1} |C_i| < n$ and $\sum_{i=1}^{k} |C_i| \geq n$. Similarly, we sample the tweets from these k communities with a probability of p such that $\sum_{i=1}^{k} |C_i| * p = n$.

In the next section, we will evaluate the impact of different sampling strategies on understanding the characteristics of Twitter data streams.

V. Experiments and Evaluations of Sampling Designs via Flu and Ebola Tweet Streams

A. Impact Difference on Tweet Distributions

To study the impact of sampling on tweet distribution across users, we focus on all tweets containing the keyword "Ebola" during the entire month of October 2014. As our data collection facility collects all tweets containing Ebola via the *following* streaming APIs, this full data-set of Ebola tweets

TABLE I
THE IMPACT OF SAMPLING ON RETWEET STATISTICS.

tweet ID	retweet users			influencers			average influence			largest cc of retweet graph	
	full	sample	ratio	full	sample	ratio	full	sample	ratio	full	sample
1	49340	557	1.2%	2157	50	2.3%	22.10	10.16	45%	100%	15.79%
2	46626	514	1.1%	851	43	5.0%	54.27	10.98	20%	100%	25.29%
3	24265	220	0.9%	782	19	2.4%	30.28	10.58	34%	100%	22.72%
4	14739	186	1.2%	664	27	4.0%	21.37	5.93	27%	100%	50.53%
5	13776	130	0.9%	555	15	2.7%	24.04	7.67	31%	100%	35.38%

serves as the ground truth and the benchmark of the tweet distribution across users.

For ensuring the fairness, the number of tweets for all sampling strategies is set to the same. Specially, we first set 1% as the sampling rate for random sampling and stratified sampling, and obtain the set of randomly selected tweets during October 2014. Subsequently, we search for one or a few communities which tweeted or retweeted on Ebola for approximately the same number during October 2014.

Figure 7 illustrates the tweet distributions of users with random sampling, stratified sampling, and community sampling as well as the full data set. It is very interesting to observe that community sampling achieves the closest tweet distribution across users among all three sampling strategies, while random sampling and stratified sampling have similar distributions which significantly deviate from the original distributions in the full data-set.

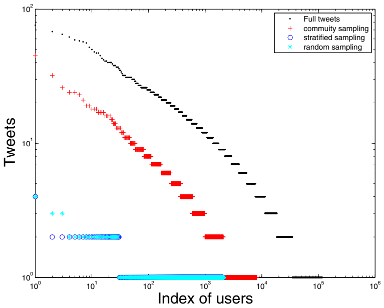


Fig. 7. The tweet distributions of users with different sampling designs.

B. Impact Difference on User Influence

In addition to characterizing impact difference on tweet distributions across users, we also analyze impact difference of sampling design strategies on user influence. Through analyzing five popular tweets on flu or Ebola topics, Table II summarizes the statistics of their retweet users, influencers, average influence and largest connected components of retweet graphs after sampling with three different strategies: community sampling, stratified sampling, and random sampling. As shown in Table II, all three sampling designs lead to similar numbers of retweet users, influencers, and average influence.

However, community sampling achieves the largest connected component of retweet graphs after sampling for all five cases. Our in-depth analysis shows that the underlying reason lies in the nature of community sampling process, which includes the tweets and retweets of the entire communities thus preserving the relationships of community social graphs and retweeting activities. On the other hand, stratified sampling and random sampling have dropped many retweeting and following relationships during the sampling process thus breaking retweet graphs into smaller connected components.

In summary, our experimental results with real flu and Ebola tweets show that community sampling retains the tweet distribution across users in a similar fashion as full data-set, while random sampling and stratified sampling are unable to retain the original tweet distributions in the full data set. In addition, community sampling successfully preserves the largest connected components in the retweet graphs formed by users who retweet the same tweet, reflecting their common interests on a certain topic or event. Therefore, we believe that community sampling is an excellent candidate for reducing massive amount of social media data while preserving important characteristics in the original and full data-set.

VI. Related Work

The rapid growth of online social networks such as Facebook, Twitter and Sina Weibo has received significant attentions from the research community. A rich body of literature studies have focuses on user behaviors [4], social graphs [5], information diffusions [6], trending topics [7], community detections [8], security and privacy in social networks [9]. The sheer *volume* of users and contents in online social networks, a *variety* of analytical metrics such as user relationships, interactions, influence, and topics, the streaming nature and high *velocity* of user-generated posts, contents and tweets have produced *big data* in online social networks [10].

Given the challenges of analyzing big social media data, sampling becomes a natural and intuitive approach to reduce the size for data collection, processing and analysis in online social networks. A number of studies focus on sampling social network graphs which often contain millions or event billions of nodes and edges [11], [12], [13]. For example, [11] explores Metropolis-Hasting random walk (MHRW) and a reweighted random walk (RWRW) to collect an unbiased sample of Facebook users through crawling its social graph, while [12] presents a heuristic for sampling large, static, undirected and crawling-based graphs via a stratified weighted random walk which considers both the graph structure and node properties. In addition, [13] introduces a variety of sampling-based algorithms to quickly approximate the neighborhood of a given user in a social network via a random set of sampled nodes in the neighborhood.

TABLE II
THE IMPACT OF DIFFERENT SAMPLING STRATEGIES ON USER INFLUENCES OVER ONLINE SOCIAL NETWORKS.

tweet ID	retweet users			influencers			average influence			largest cc of retweet graph		
	community	stratified	random	community	stratified	random	community	stratified	random	community	stratified	random
1	462	526	557	48	44	50	8.85	10.98	10.16	26.66%	18.25%	15.79%
2	482	519	514	48	49	43	9.08	9.59	10.98	38.82%	27.36%	25.29%
3	241	261	220	26	22	19	8.50	10.86	10.58	36.73%	26.81%	22.72%
4	147	165	186	18	22	27	7.33	6.50	5.93	51.61%	46.66%	50.53%
5	241	152	130	26	18	15	8.50	7.44	7.67	50.00%	33.55%	35.38%

In parallel to our research, a few studies have also investigated the impact of sampling on analyzing and characterizing online social networks [14], [15], [16]. For example, [15] studies how much sampling affects hashtags and topic analysis, network graph analysis and geolocations of tweets via comparing Twitter sampling stream and Twitter Firehose which includes every published tweet, while [16] empirically characterizes the impact of attribute and topology based sampling approaches on the discovery of information diffusion over Twitter data streams. In [14], Ghosh et al. reveal major differences on topic diversity, information timeliness, and content quality from two tweets streams, which are collected from random sampling and a group of expert users, respectively.

VII. CONCLUSIONS AND FUTURE WORK

The explosive growth of online social networks such as Twitter and Sina Weibo has produced massive amount of data in social graphs, user-generated contents and network traffic. The big data in online social networks has driven many researchers to explore sampling strategies to make sense of online social networks. However, little is known about the impact of sampling on the results and conclusions of online social networks analysis. In this study, we collect sampled data-set and full data-set of two important topics, i.e., flu and Ebola, from Twitter data streams, and quantify the impact of random sampling on tweet volume estimations, tweet distributions across users, and user interactions, and user influence. In addition, we design community sampling and stratified sampling for understanding the tradeoff and benefits of different sampling strategies on analyzing social networks data. Our experimental results demonstrate that community sampling, while achieving similar data reductions as random sampling and stratified sampling, delivers the best performance in terms of preserving tweet distributions across users, estimating user influence measured by retweeting and commenting activities, and characterizing retweet graphs among users. Our future work lies in developing adaptive and synthetic sampling approaches to reduce the size of social network data while minimizing the impact of sampling on the accuracy of social network characterizations.

ACKNOWLEDGMENT

This work was supported in part by National Science Foundation grant CNS #1218212 and by RGC of Hong Kong Project No. CityU 114713.

REFERENCES

[1] Twitter, "The Streaming APIs," https://dev.twitter.com/streaming/public.

[2] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.

[3] G. Henry, *Practical Sampling, Applied Social Research Methods Series*. SAGE Publications, 1990.

[4] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing User Behavior in Online Social Networks," in *Proceedings of ACM SIGCOMM International Measurement Conference*, November 2009.

[5] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of ACM SIGCOMM conference on Internet measurement*, October 2007.

[6] F. Wang, H. Wang, K. Xu, J. Wu, and X. Jia, "Characterizing Information Diffusion in Online Social Networks with Linear Diffusive Model," in *Proceedings of IEEE International Conference on Distributed Computing Systems (ICDCS)*, Philadelphia, PA, July 2013.

[7] S. Asur, B. Huberman, G. Szabo, and C. Wang, "Trends in Social Media: Persistence and Decay," in *Proceedings of International Conference on Weblogs and Social Media (ICWSM)*, July 2011.

[8] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in Social Media," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 515–554, May 2012.

[9] C. Zhang, J. Sun, X. Zhu, and Y. Fang, "Privacy and security for online social networks: challenges and opportunities," *IEEE Network*, vol. 24, no. 4, pp. 13 – 18, July 2010.

[10] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of International Conference on World Wide Web (WWW)*, April 2010.

[11] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou, "Walking in Facebook: A Case Study of Unbiased Sampling of OSNs," in *Proceedings of IEEE INFOCOM*, March 2010.

[12] M. Kurant, M. Gjoka, C. Butts, and A. Markopoulou, "Walking on a Graph with a Magnifying Glass: Stratified Sampling via Weighted Random Walks," in *Proceedings of ACM SIGMETRICS*, June 2011.

[13] M. Papagelis, G. Das, and N. Koudas, "Sampling Online Social Networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, pp. 662–676, March 2013.

[14] S. Ghosh, M. B. Zafar, P. Bhattacharya, N. Sharma, N. Ganguly, and K. P. Gummadi, "On Sampling the Wisdom of Crowds: Random vs. Expert Sampling of the Twitter Stream," in *Proceedings of ACM International Conference on Information and Knowledge Management (CIKM)*, October 2013.

[15] F. Morstatter, J Pfeffer, H. Liu, and K. Carley, "Is the Sample Good Enough? Comparing Data from Twitters Streaming API with Twitters Firehose," in *Proceedings of International Conference on Weblogs and Social Media (ICWSM)*, July 2013.

[16] M. Choudhury, Y. Lin, H. Sundaram, K. Candan, L. Xie, and A. Kelliher, "How Does the Data Sampling Strategy Impact the Discovery of Information Diffusion in Social Media?" in *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, May 2010.