Characterizing Twitter with Respondent-Driven Sampling

Mostafa Salehi, †Hamid R. Rabiee, Nasim Nabavi, Shayan Pooya
Department of Computer Engineering
Sharif University of Technology
Tehran, Iran
{mostafa_salehi, nabavi, pooya}@ce.sharif.edu, †rabiee@sharif.edu

Abstract— **Twitter as one of the most important microblogging online social networks has attracted more than 200 million users in recent years. Although there have been several attempts on characterizing the Twitter by using incomplete sampled data, they have not been very successful to estimate the characteristics of the whole network. In this paper, we characterize Twitter by sampling from its social graph and user behaviors through a random walk based sampling technique called Respondent-Driven Sampling (RDS). To the best of our knowledge, for the first time RDS method and its estimator are used in order to obtain uniform unbiased estimation of several key structural and behavioral properties of Twitter. We compared the performance of the proposed method with other sampling methods such as Metropolis-Hasting Random Walk (MHRW) and sampling from active users (Timeline) against the uniform sampling (UNI). In order to gather the required data, we have implemented four independent crawlers. Our experimental results indicate that the RDS method exhibits lower estimation errors to the sample in- and out- degree distribution compared to MHRW and Timeline. We also show that RDS is more suitable to sample the followers vs. followings ratio, and the correlation between followers/followings vs. tweets.**

Keywords- Twitter, Online Social Network, Sampling, Crawling, RDS, MHRW, Public Timeline, Uniform

I. INTRODUCTION

Twitter is the most popular microblogging Online Social Network (OSN). Currently Twitter has attracted more than 200 million users worldwide. The primary way of communication in Twitter is sending messages, called tweets (text-based messages shorter than 140 characters). Twitter users usually post many tweets about their daily activities. Users may subscribe to other author's tweets that is known as following, and subscribers are known as followers. By default, the tweets are public and everybody can see them, but a user may limit this feature so that only his/her followers can read his/her tweets. The Twitter network can be conceptualized as a directed graph, with the vertices of the graph representing the users and the directed edges of the graph representing the relationships (follower or following) between the users.

In recent years, a considerable amount of research has been done on the analysis of Twitter network characteristics [1], [2], [3], [4]. This studies range from the calculation of simple measurement summarizing structural properties, such as degree distribution to the extraction of complex relational patterns, such as followers vs. following ratio. One of the most difficult aspects in the characterization of the Twitter networks is its large size. In particular, size is an insuperable challenge for modeling this network, simulating its dynamical behavior or extracting common structural properties. The data that the most studies use to characterize Twitter is collected by a sampling method and it's only a portion of the Twitter network (contrary to what could be done [1] in 2007). The network resulting from such measurements may be thought of as a sample from the large Twitter network. These studies assert that characteristics of a sampled network graph are indicative of the same characteristics for the whole network graph. For instance, the properties for Twitter's social network such as degree distribution presented in [1], [2] and the characterization of Twitter users in [5], [4] are, in fact, the properties of the sampled graph, not the properties of the original graph. Such problems can be compensated for in many cases by using appropriate estimators [6].

The goal of this paper is to investigate the utility of various sampling methods for characterizing Twitter. Random Walk (RW) is a practical method to sample the social graphs by asking a user to identify several neighbours (for instance, followers in Twitter), one of whom is selected at random to be the next user, with the pattern continuing for a number of steps [7]. However, this method introduces a considerable selection bias [8].

In this study, we propose the use of Respondent Driven Sampling (RDS) for Twitter which is a RW-based sampling method in contrast to the Metropolis-Hasting Random Walk (MHRW) [3]. MHRW modifies the probabilities of next user selection in ordinary RW in order to have the uniform stationary distribution for visiting each user. This technique has been used for sampling from Twitter [3], Facebook [8], and Peer-to-Peer networks [9]. On the other hand, RDS uses the unmodified ordinary RW method to sample from a graph, but it corrects the selection bias by re-weighting the sampled values [10], [11], [12]. In the context of network sampling, this method has been used for Facebook [8], a variety of synthetically generated graphs [13], peer-to-peer networks [9]. We also collect a sample via the public timeline to sample currently active Twitter [3], [5], [4]. This sampling method has been used in most previous studies to characterize Twitter. We call this method Timeline. Moreover, we consider the UNI method [8] to collect a uniform sample by querying randomly generated user IDs. We use this sample as ground truth to evaluate other sampling techniques [8]. Such method can be costly when the user IDs space is sparse [14].

To the best of our knowledge, in this paper for the first time, we use the RDS method for sampling from the Twitter network and estimating its characteristics. Moreover, we use the Kolmogorov-Smirnov D-statistic to quantify the estimation error in MHRW, RDS and Timeline methods in comparison to the UNI method. In terms of results, we show that the accuracy of the RDS method for estimating the in- and out-degree distribution of the Twitter network is more than other methods. In addition, our empirical evaluations reveal that RDS outperforms MHRW and Timeline to estimate a number of user behavioural patterns including followers vs. followings ratio, correlation between followers and tweets, and correlation between followings and tweets.

The organization of the paper is as follows. Section II discusses the related work. Section III describes the sampling techniques. Section IV summarizes the data collection process and the data sets. Section V provides a characterization of some key Twitter properties based on the aforementioned sampling techniques. Finally, section VI concludes the paper.

II. RELATED WORK

Graph sampling methods in network context are somewhat distinct from classical sampling methods. Random Walk (RW) [15] is one of the most important and widely used sampling methods in different kind of network contexts such as uniformly sampling Web pages from the Internet [16], content density in peer-to-peer networks [13], [9], [17], [18] degree distributions of the Facebook social graph [8] and large graphs in general [19]. An ordinary RW samples a graph by moving from a vertex, x, to a neighboring vertex, y, through an outgoing edge, $\{x, y\}$, chosen uniformly at random from x's neighbors. By this process edges and vertices are sampled. The probability of selecting the next vertex determines the probability that vertices are sampled. In any given connected and non-bipartite graph G, the probability of being at a vertex x converges at equilibrium to the stationary distribution $\pi(x) = degree(x) \,/\, 2|E|$, where $degree(x)$ and E are the degree of vertex x and are the set of edges of the network graph. Thus, the ordinary RW is biased towards vertex with higher degree.

Metropolis Hasting technique [20] can be used to modify the probabilities of the vertex selection in RW in order to have the uniform stationary distribution for visiting each vertex ($\pi(x) = 1/|V|$). This technique is a general Markov Chain Monte Carlo method [21] for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its stationary distribution. This approach, known as Metropolis-Hasting Random Walk (MHRW), has been applied to peer-to-peer networks [9], Facebook [8], and Twitter [22].

Alternatively, one can use the unmodified ordinary RW method to sample from a graph and correct the degree bias by re-weighting the sampled values. RDS method presents an approach to correct the bias [10]. Estimating the probability of visiting each vertex in RDS is based on a Markov Chain representation of the sampling process [11]. RDS method is an approach in the field of social sciences to sample and inference in hard-to-reach populations such as injection drug users [23]. In these populations, a sampling frame for the target population is not available. Sampling from online social network such as Twitter is analogous to the sampling of hidden population in the social sciences. In the context of graph sampling, RDS method has been used for Facebook [8], a variety of synthetically generated graphs [13], and Peer-to-Peer networks [9]. In this paper, we use RDS and MHRW as two methods which provide unbiased estimation of the network characteristics.

In some previous studies about the characterization of Twitter, the dataset was gathered via the two other techniques, namely the public timeline (to sample currently active user) [3], [5], [4] and BFS [2], [3], [4]. However, the estimation error in these methods has not been analysed so far for arbitrary graphs. Moreover, BFS leads to bias towards high degree vertices [24]. In this study, we use mainly the public timeline as a baseline for comparison.

III. SAMPLING METHODS AND ESTIMATORS

In general, there are two approaches for constructing and analyzing samples of networks in different fields; the graph sampling and the model based sampling.

Graph sampling can be categorized into two sub-groups according to the amount of available information about the network graph [19]; the coarse graining methods and graph exploration methods. In the coarse graining methods [25], the graph of network can be observed initially (i.e., the sampling frame is available), and its focus is on development of a smaller sub-graph of a network such that the sub-graph keeps at least some of the most relevant properties of the original network. In order to achieve this goal, most existing techniques are based on the idea of either grouping nodes together or removing some nodes. These techniques are very common in the statistical physics literature. On the other hand, reducing the network complexity by graph sampling may cause some information about the initial network to be lost. In the second sub-group, graph exploration techniques, the graph of network is unknown initially, except for some limited number of nodes (i.e., no sampling frame is available). The focus of these techniques is on deriving sample statistics for characteristics measured at each node of the sub-graph sample (e.g., node degree). Graph sampling techniques in this sub-group are based on the graph exploration methods, such as Random Walk. For unbiased estimation of the characteristics of an unknown network, traditional statistical inference has to be modified by assigning the appropriate weights to the values observed at each node in the sample. These techniques are very common in sociology and computer science.

The techniques in the second approach are model-based [26]. Here, we initially consider a probability model with unknown parameters for the target network. The focus of the techniques in this approach is to use the observed samples from the network to estimate the model parameters, and to extract the network characteristics based on this model. The main challenge in this approach is to specify probability models that adequately incorporate structures of the target network. Model-based approaches are very common in mathematics and statistics.

In Twitter Characterization, we assume that the graph of network is unknown initially. The objective of the network sampling in this study is not statistical inference about the parameters of a network model, but is extracting sample statistics for measured characteristics. Therefore, our focus in this paper is on the graph exploration methods. In the following, we describe various sampling methods and their corresponding approach to estimate characteristics of the Twitter network.

A. Metropolis Hastings Random Walk (MHRW)

As we mentioned earlier, an ordinary RW samples a graph by moving from a vertex, *x*, to a neighboring vertex, *y*, through an outgoing edge $\{x, y\}$, chosen uniformly at random from *x*'s neighbors. In each iteration of the MHRW method, as a Random Walk (RW) based sampling method, the walk selects the next vertex y uniformly at random among the neighbours of the current vertex x and then accept this selection with probability $\min(1, \mathrm{degree}(x)/\mathrm{degree}(y))$, in which the $\mathrm{degree}(x)$ is the number of *x*'s neighbors. Otherwise it stays at x. Therefore, the transition probability to vertex y is given by [27], [8]:

$$\mathrm{P}(\mathrm{x},\mathrm{y}) = \begin{cases} \frac{1}{\mathrm{degree}(\mathrm{x})} \min\left(1, \frac{\mathrm{degree}(\mathrm{x})}{\mathrm{degree}(\mathrm{y})}\right) & \text{if y neighbor of x} \\ 1 - \sum_{\mathrm{z} \neq \mathrm{x}} \mathrm{P}(\mathrm{x},\mathrm{y}) & \text{if y = x} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

This selection mechanism provides a way to sample nodes from the uniform distribution π, with $\pi(\mathrm{x}) = 1/|V|$. Thus, when the sample (sequence of sampled vertices $\mathrm{x_i}$) has been collected, we use sample mean $\bar{\mathrm{x}} = {}^{1}/_{N} \sum_{i=1}^{N} \mathrm{x_i}$ as an estimator and estimate the measured characteristic by the values observed in the sampled graph.

B. Respondent-Driven Sampling (RDS)

RDS is a new sampling method for hidden populations that is rapidly gaining in popularity. A population is hidden when there is no sampling frame, a list of all members we can sample from, such that samples may only be collected through iterative referrals from existing samples. RDS can be presented as Markov Chain Monte Carlo (MCMC) importance sampling. The problem of estimating vertex properties in Twitter social network is analogous to the sampling of hidden population in the social sciences.

In each iteration of the RDS method, as a Random Walk based sampling method, the walk selects randomly n (usually n=3 [28]) neighbours of the current vertex v as the next hop vertex. When n = 1, RDS is equivalent to the ordinary random walk. In RDS, vertices with more neighbours in the graph of network are more likely to be selected [29]. To adjust for this selection bias, RDS uses the Hansen-Hurwitz estimator [30], i.e. vertices are weighted inversely proportional to their network degree (i.e. the number of neighbours). Specifically, for the Twitter network, we consider in-degree or the number of followers to weight observed values. Suppose a stationary RW that has visited $\mathrm{V}' = \{\mathrm{x_i}\}$ sequence of sampled vertices. Each vertex is associated with a label Y. For instance, a label can be the number of followers of a user and its corresponding network characteristic is the fraction of user with Y_i followers. Let B_i contains all vertices x with label Y_i. We seek to calculate, $P(B_i)$, in the graph of network. The RDS estimator for $P(B_i)$ is defined to be [12]:

$$\widehat{\mathrm{P}}(\mathrm{B_i}) = \frac{\sum_{\mathrm{x} \in \mathrm{B_i}} {}^{1}/_{\mathrm{degree(x)}}}{\sum_{\mathrm{x} \in \mathrm{V}'} {}^{1}/_{\mathrm{degree(x)}}} \tag{2}$$

C. TimeLine

If the profile of a user is made public in the Twitter website, his/her activities appear in the public timeline of the recent updates. The third dataset used in this paper was collected by using the public timeline to sample currently active users. Samples were made by extracting the list of recent updates in the public timeline and selecting the set of users associated with the statuses in this list. Then, all information of these users was collected. Next, the public timeline was queried again to find the next set of active users. We call this method Timeline. This process is usually used in the previous studies for data gathering from Twitter network [3], [5], [4]. The sample mean was implicitly used in these studies to estimate the interesting characteristic of the Twitter. Although, the sample mean used can't be a valid estimator for Timeline, we used it in this paper to compare the Timeline method with the other sampling methods.

D. Uniform Sample (UNI)

Each user profile in Twitter is assigned a unique numerical ID. Therefore, one can sample users uniformly by querying randomly generated numerical IDs and discarding non-allocated IDs (non-existing users). This rejection sampling [31] guarantees a uniform sampling of the existing users. We gathered the forth dataset in this study by generating uniformly random 32-bit user IDs. In general, this approach can be resource intensive because the ID space is sparsely populated [32]. We use this method as the ground truth for assessing the quality of other sampling method. This process was used in [8] for data gathering from Facebook. Sample mean is used as an estimator in the UNI method.

IV. DATA COLLECTION

We have implemented four independent crawlers for each method, namely RDS, MHRW, Timeline, and UNI, using the API functions provided by Twitter [33]. The RDS and MHRW methods start at a randomly selected user. At each step, the RDS method selects 3 random followings of the current user to be visited. We let each crawler to collect data during Sep. 29th to Oct. 12th, 2010 and gathered the full profile of each visited user, such as name and user ID, followers, followings, tweets and location. The collected datasets are summarized in Table I. The MHRW dataset contains 1881 unique user, which is less than the visited

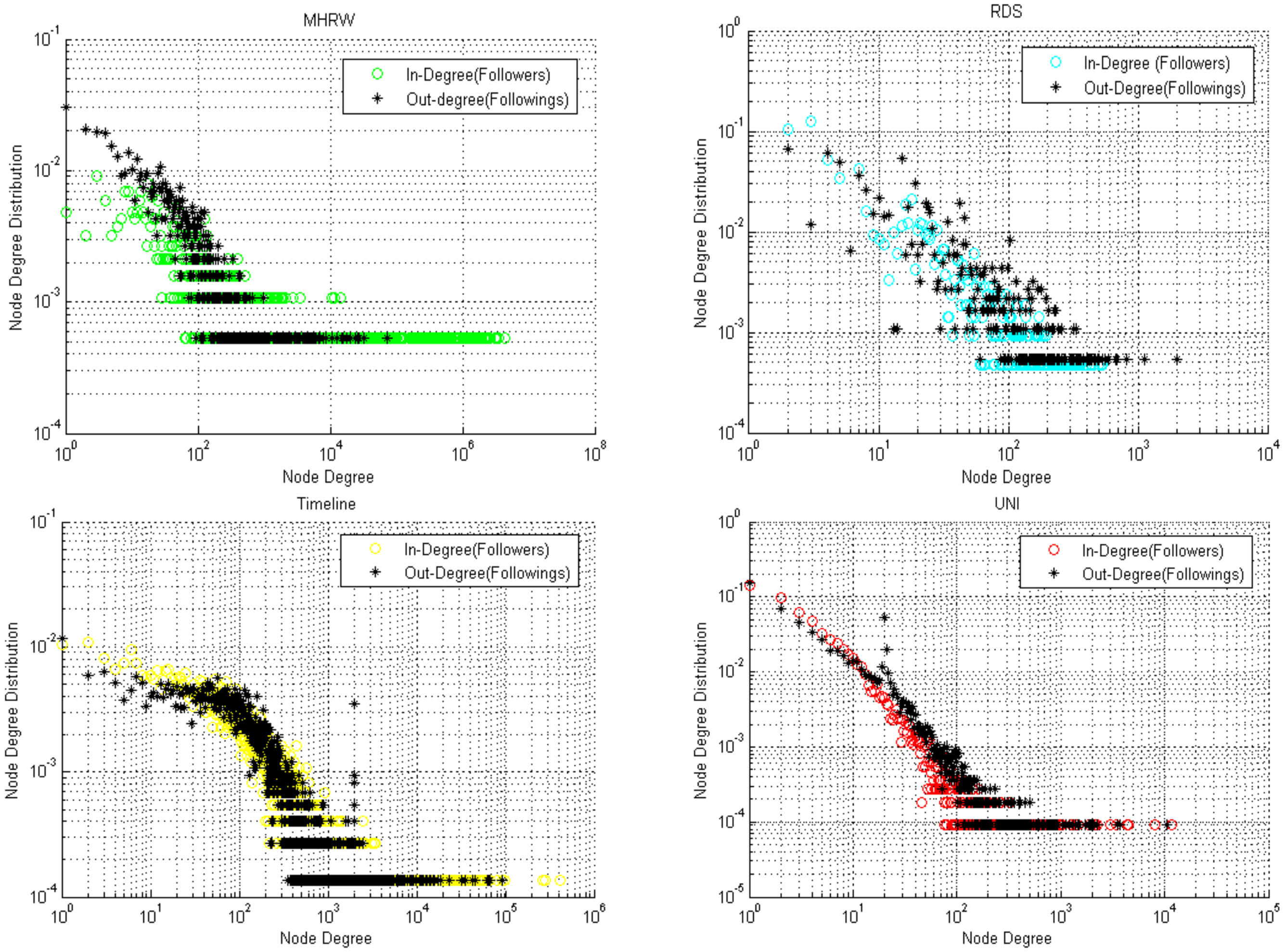

Figure 1. In- and out-degree distribution estimated by the four sampling methods, namely MHRW (top-left), RDS (top-right), Timeline (down-left), and UNI (down-right). All plots use log-log scale.

users in other datasets; this is because MHRW may repeat the same vertex in a walk. For the UNI method, we visited 11006 existing users by using user IDs which were picked uniformly at random from [0, 2^{32}-1].

TABLE I. COLLECTED DATASETS FROM TWITTER BY DIFFERENT SAMPLING METHODS FROM SEP. 29TH TO OCT. 12TH, 2010

Total	***MHRW***	***RDS***	***Timeline***	***UNI***
Users	1881	2207	7412	11006
Edges	1880	5955	3171048	205078
Tweets	247233	345045	1342279	124574

V. CHARACTERIZATION RESULTS

In this section, we evaluate all candidate methods including RDS, MHRW, and Timeline by comparing them to UNI, in terms of accuracy in estimating the degree distribution and the user behaviour of Twitter.

A. Degree Distribution

Previous research has revealed that there exist some common structural properties such as small-world [34], scale-free [35], and high clustering attributes [36] in many real-world networks. These attributes have important effects on the behaviour and dynamical properties of networks [37], [38]. The pattern of connections between users on a social network, affect how people learn, form opinions, gather news, as well as affecting other phenomena, such as the spread of ideas. Unless we know something about the structural properties of these networks, we cannot hope to adequately understand how the corresponding system works.

The degree distribution, $P(k)$, is one of the most important structural properties that provides a natural summary of the connectivity in the graph of the network. More specifically, $P(k)$ is the fraction of vertices in a network with degree k. For directed networks there are an in-degree distribution and an out-degree distribution. In-degree distribution which is the distribution of the number of followers, and out-degree distribution which indicates the distribution of the number of followings are shown in Figure 1. As can be seen from this figure, one common aspect of all the four methods is the broad range over which the in- and out-degrees in the graphs vary. Both the in- and out-degree distributions indicate that a significant fraction of Twitter users have very low number of followers/followings. These correspond to the inactive members who have abandoned the Twitter shortly after creating their accounts (i.e., the tourists) [39] .

We use the Kolmogorov-Smirnov D-statistic, D , to measure the agreement between the degree distribution in candidate sampling methods (RDS, MHRW, and Timeline) and UNI method (Table II). The D-statistic is a relative measure to compare the distribution of the values in the two datasets [40]. It is defined as $\mathrm{D} = \max\{|\mathrm{F}'(\mathrm{x}) - \mathrm{F}(\mathrm{x})|\}$, where x is over the range of the interested characteristics; F and F' are the two empirical Cumulative Distribution functions (CDfs) of the data. A value of $D \leq \propto$ corresponds to no more than $\propto$ percentage point difference between CDFs. It is clear that smaller values of D indicate better fits of the sampling distribution to the UNI distribution.

While in the most previous works [3], [5], [4], [41], MHRW and Timeline methods have been used to characterize the Twitter network, the RDS method provides better estimation of the degree distributions compared to MHRW and Timeline methods, as shown in Table II.

TABLE II. KOLMOGOROV SMIRNOV D-STATISTIC FOR DEGREE DISTRIBUTION IN THREE CANDIDATE SAMPLING METHODS COMPARED TO UNIFORM SAMPLING METHOD (UNI)

Distribution	*MHRW*	*RDS*	*Timeline*
In-degree	0.81	0.32	0.78
Out-degree	0.53	0.19	0.71

B. *Followers vs. Following ratio*

The followers (those who follow you) vs. followings (those you follow) ratio is an important measurement that speaks a lot about the behaviour of a Twitter user and is also a measure of how well the users contributing to the Twitter. In addition, other users on Twitter may decide whether or not to follow you based on your followers/following ratio.

If you have 200 followers and you are following 100 people, your ratio is 2:1. A larger ratio means you have more followers than people you are following. Obviously, a smaller ratio means the opposite. In general, a ratio of less than 1.0 indicates that you are seeking information (from followings), but not getting much attention in return. Such behavior is typical of spammers, who contact everyone they can, and hope that some will follow them. Therefore, spammers have tiny ratios. A ratio of around 1.0 means you tend to exhibit reciprocity in your relationships, i.e. you are respected among your peers. A ratio of higher than 1.0 shows that you are a popular person and other users want to hear what you have to say, and you might be considered as a leader in your community. For example celebrities have enormous followers/following ratios.

To compute the dependence between followers and followings of a user in Twitter, we calculate the spearman rank correlation (r) between them [42]. The correlation value is a number between -1 and 1 with $r = 1$ corresponding to identical ranking and $r = -1$ corresponding to perfect inverse correlation. Finally, $r = 0$ corresponds to uncorrelated rankings. The statistical significance of r is tested by using the P-value. The P-value is the probability of getting a correlation as large as the observed value when the true correlation is zero.

The value of the calculated correlation gathered from the crawled data, is 0.21, 0.59, 0.78, and 0.66 for the MHRW, RDS, Timeline, and UNI sampling methods, respectively. The corresponding P-value for all correlations is near zero. Therefore, there exist significant high correlation values for all sampling methods. This indicates that the more the number of followings the more the number of followers. Hence, active users (i.e., those who have many followings) in the Twitter social networks also tend to be popular (i.e., those who have many followers). Moreover, the correlation for the RDS method is much closer to the correlation value of the UNI method, compared to the MHRW and Timeline sampling methods.

The high correlation between followers and followings in Twitter can be explained by the high number of symmetric edges. The high symmetry may be due to the tendency of users to reciprocate edges from other users who point to them. This process would result in receiving many incoming edges for active users. Moreover, it validates the results that we have observed.

Next, we compared the followers and followings of an individual user in the Twitter social network. Figure 2 illustrates the cumulative distributions of the followers vs. followings ratio for the four sampling methods. The RDS, Timeline, and UNI methods show a remarkable correspondence between followers and followings. However, the distribution for the MHRW method is markedly different; most users have considerably higher followers than followings, while a small fraction of users have significantly higher followers than followings. We used the D-statistic to quantify the difference between the distributions for the candidate sampling methods by considering UNI method as the ground truth. The values of D for the MHRW, RDS, and Timeline methods are 0.51, 0.21, and 0.38, respectively. Therefore, the RDS method is much more suitable for estimating the followers vs. following ratio than MHRW and Timeline methods.

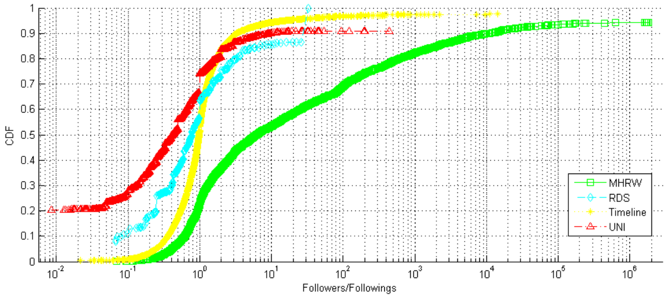

Figure 2. CDF of followers vs. followings ratio for all candidate sampling methods, namely MHRW, RDS, Timeline, and UNI.

C. Correlation between Followers and Tweets

In order to find the relationship between the number of the followers and tweets of an individual user, the spearman rank correlation (r) values between these parameters were calculated (Table III). Our results show significant high correlations for all sampling methods, i.e. the more the number of followers the more the number of posted tweets. Hence, popular users (i.e., those who have many followers) in Twitter, tend to more participate in tweeting. Perhaps when a user attracts a relatively large number of readers, she/he feels obligated to keep them entertained. Most likely, the more you Tweet the more followers you attract, provided you actually have something interesting to say. Moreover, the correlation values for the MHRW, RDS, and UNI methods are close to each other ($r \approx 0.50$). It is also shown that the Timeline method has the highest correlation value ($r = 0.72$).

TABLE III. THE SPEARMAN RANK CORRELATION BETWEEN THE FOLLOWERS/FOLLOWINGS AND THE POSTED TWEETS (THE CORRESPONDING P-VALUE FOR ALL CORRELATION IS NEAR ZERO.)

	MHRW	***RDS***	***Timeline***	***UNI***
Tweets and Followers	0.56	0.50	0.72	0.53
Tweets and Followings	0.45	0.41	0.51	0.52

Figure 3 shows the CDF of tweets vs. followers ratio for all four sampling methods. About 15% of selected users in UNI do not have any followers. Moreover, in the distribution for the Timeline method, most of the users have considerably higher tweets than the followers, while a small fraction of them have significantly higher tweets than followers. In addition, it is shown that the RDS, Timeline, and UNI methods have different distributions; about 50% of users have higher tweets than followers, others have lower tweets than followers. Furthermore, Figure 3 illustrates that the distribution of the tweets vs. followers ratios for UNI as ground truth is best matched by the RDS method. The values of D-statistic are 0.23, 0.18, and 0.56 for the MHRW, RDS, and Timeline, respectively. These results show that the RDS and MHRW methods are more suitable for estimating the tweets vs. followers ratio than the Timeline method.

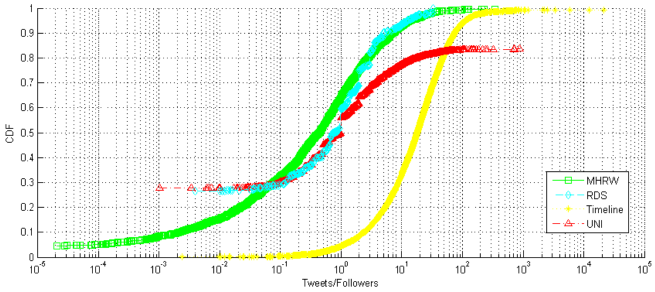

Figure3. CDF of tweets vs. followers ratio for all candidate sampling methods includes MHRW, RDS, Timeline, and UNI.

D. Correlation between Followings and Tweets

Similarly, we calculate the correlation between the number of the followings and tweets of an individual user in Twitter (Table III). The results show significant high correlations between followings and tweets for all sampling methods, i.e. the more the number of followings the more the number of posted tweets. Hence, active users (i.e., those who have many followings) tend to participate more in Twitter. Moreover, the correlation values for the MHRW, RDS, and UNI methods are close to each other (r is in the range of 0.4 to 0.5).

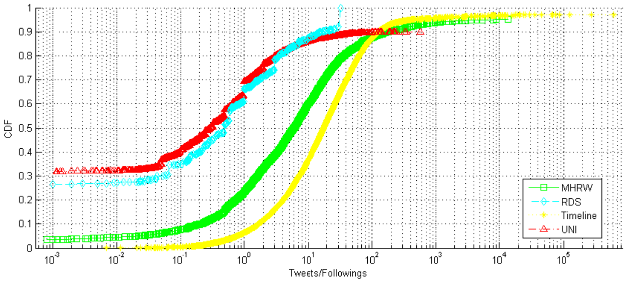

Figure 4. CDF of tweets vs. followings ratio for all candidate sampling methods includes MHRW, RDS, Timeline, and UNI.

Figure 4 shows the CDF of tweets vs. followings ratio for all four sampling methods. About 10% of the selected users in UNI do not have any followings. Moreover, in the distribution for the MHRW and Timeline methods, most of the users have considerably higher tweets than followings, while a small fraction of them have higher tweets than followings. In addition, it is shown that the RDS and UNI methods have different distributions; about 60% of users have higher tweets than followings, while others have lower tweets than followings. Furthermore, Figure 4 shows that the distribution of the tweets vs. followings ratios for UNI, as ground truth, is best matched by the RDS method. The values of D-statistic are 0.46, 0.18, and 0.65 for the MHRW, RDS, and Timeline, respectively. These results show that the RDS method is more accurate than the competing sampling methods for estimating the tweets vs. followings ratio.

VI. CONCLUSIONS

In this study, we have presented RDS as a powerful technique for sampling from the Twitter network. The performance of the RDS method was compared with MHRW, Timeline, and UNI (as uniform sampling) methods by gathering the required data by implementing four different types of crawlers. The results of our analysis revealed that RDS outperforms the competing sampling methods. It performed remarkably better in estimation of the in- and out-degree distribution (closer performance to the uniform sampling). It was also more efficient in terms of accuracy for sampling the followers vs. followings ratio and the correlation between followers/followings and tweets. In our future work, we would focus on the sampling from disconnected/connected components of the Twitter network by using multiple random walks.

ACKNOWLEDGMENT

The authors would like to thank Mr. Naderi of Sharif University for his help in preparing the manuscript.

REFERENCES

[1] A. Java, X. Song, T. Finin, and B. Tseng, “Why we twitter: understanding microblogging usage and communities,” *Web-KDD and SNA-KDD*, California: ACM, 2007, pp. 56-65.

[2] S. Ghosh, G. Korlam, and N. Ganguly, “The Effects of Restrictions on Number of Connections in OSNs: A Case-Study on Twitter,” *WOSN*, USA, 2010.

[3] B. Krishnamurthy, P. Gill, and M. Arlitt, “A few chirps about twitter,” *WOSN*, USA: ACM, 2008, pp.19-24.

[4] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a Social Network or a News Media?,” *World Wide Web Conference*, ACM Press, 2010.

[5] A. Hughes and L. Palen, “Twitter Adoption and Use in Mass Convergence and Emergency Events,” *ISCRAM Conference*, Gothenburg, Sweden, 2009.

[6] E. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*, Springer, 2009.

[7] A. Klovdahl, Urban Social Networks: Some Methodological Problems and Possibilities, The Small World: A Volume of Recent Research Commemorating Ithiel De Sola Pool, Stanley Milgram, and Theodore Newcombe. (1989) 176.

[8] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou, “Walking in Facebook: A Case Study of Unbiased Sampling of OSN,” *Technical Report on arXiv:cs.NI/0906.0060*, May 2009. IEEE INFOCOM, San Diego, 2010.

[9] D. Stutzbach, R. Rejaie, N. Duffield, S. Sen, and W. Willinger, “On unbiased sampling for unstructured peer-to-peer networks,” *IMC*, Brazil: ACM, 2006, pp. 27-40.

[10] D. Heckathorn, “Respondent-driven sampling: a new approach to the study of hidden populations,” *Social problems*, vol. 44, 1997, pp. 174-199.

[11] M. Carlo, S. Goel, Matthew, J. Salganik, “Respondent-driven sampling as Markov Chain, ” *Statistics in Medicine,* 28:2202-2229, 2009.

[12] E. Volz and D. Heckathorn, “Probability based estimation theory for respondent-driven sampling,” *Journal of Official Statistics*, vol.24, 2008, pp.79.

[13] A. Rasti, R. Rejaie, N. Duffield, D. Stutzbach, and W. Willinger, “Evaluating sampling techniques for large dynamic graphs,” *Technical Report CIS-TR-08-01, University of Oregon*, Sep. 2008.

[14] B. Ribeiro, W. Gauvin, B. Liu, D. Towsley, “On MySpace account spans and double Pareto-like distribution of friends,” *IEEE Infocom 2010 Network Science Workshop*, March, 2010

[15] L. Lovász, “Random Walks on Graphs: A Survey.” *Combinatorics, Paul Erdos is Eighty,* vol.2, 1993, pp.1-46

[16] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, “On near-uniform URL sampling,” *Comput. Networks*, Amsterdam, The Netherlands: North-Holland Publishing Co., 2000, pp.295-308.

[17] C. Gkantsidis, M. Mihail, and A. Saberi, “Random walks in peer-to-peer networks: algorithms and evaluation,” Perform. Eval., vol. 63, Mar. 2006, pp. 241-263.

[18] L. Massoulié, E.L. Merrer, A. Kermarrec, A. Ganesh, “Peer counting and sampling in overlay networks: random walk methods,” *ACM Symposium on Principles of Distributed Computing*, Denver, Colorado, USA, 2006, pp. 123-132.

[19] J. Leskovec and C. Faloutsos, “Sampling from large graphs,” *Proceedings of the 12th ACM SIGKDD*, ACM Press, 2006, pp. 631-636.

[20] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller, “Equation of state calculation by fast computing machines,” *J. Chem. Phys*. 21 (1953) 1087-1092.

[21] S. Chib, E. Greenberg, “Understanding the Metropolis-Hastings Algorithm, ” *The American Statistician*. 49 (1995) 327-335.

[22] B. Krishnamurthy, P. Gill, and M. Arlitt, “A few chirps about twitter,” *WOSN*, Seattle, ACM, 2008, pp. 19-24.

[23] A. Abdul-Quader, D. Heckathorn, C. McKnight, H. Bramson, C. Nemeth, K. Sabin, K. Gallagher, and D. Des Jarlais, “Effectiveness of respondent-driven sampling for recruiting drug users in New York City: Findings from a pilot study,” *Journal of Urban Health*, vol. 83, 2006, pp. 459-476.

[24] L. Becchetti, C. Castillo, D. Donato, and A. Fazzone, “A Comparison of Sampling Techniques for Web Graph Characterization,” *LinkKDD*, Philadelphia, 2006.

[25] D. Gfeller and Paolo, “Spectral Coarse Graining of Complex Networks,” Physical Review Letters, vol. 99, 2007.

[26] M.S. Handcock and K. Gile, “Modeling Social Networks with Sampled Data,” Annals of Applied Statistics, 2007.

[27] W. Gilks, S. Richardson, and D. Spiegelhalter, “Markov Chain Monte Carlo in Practice,” *Chapman and Hall*, 1996, pp. 1-19.

[28] M. Kurant, A. Markopoulou, and P. Thiran, “On the bias of BFS,” *International Teletraffic Congress (ITC 22),* 2010.

[29] M. Salganik, D. Heckathorn, “Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling, ” Sociological Methodology. 34 (2004) 193-239.

[30] M. Hansen and Hurwitz, “On the theory of sampling from finite populations,” *Annals of Mathematical Statistics*, vol. 14, 1943.

[31] A. Leon-Garcia, Probability, Statistics, and Random Processes For Electrical Engineering (3rd Edition), Prentice Hall, 2008. A. Leon-Garcia, *Probability, Statistics, and Random Processes For Electrical Engineering (3rd Edition)*, Prentice Hall, 2008.

[32] B. Ribeiro, D. Towsley, “Estimating and Sampling Graphs with Multidimensional Random Walks,” *ACM SIGCOMM Internet Measurement Conference*, Nov, 2010.

[33] twitter4j, Http://Www.twitter4j.org/

[34] D. Watts, S. Strogatz, “Collective dynamics of small-world networks,” Nature. 393 (1998) 442, 440.

[35] A. Barabasi, R. Albert, “Emergence of scaling in random networks,” Science. 286 (1999) 512, 509.

[36] M. Newman, “Scientific collaboration networks. I. Network construction and fundamental results,” Physical Review E. 64 (2001) 016131.

[37] M. Newman, A. Barabasi, and D. Watts, *The Structure and Dynamics of Networks*, Princeton University Press, 2006.

[38] M. Salehi, R. Hamid R., and M. Jalili, “Motif structure and cooperation in real-world complex networks,” *Physica A*, vol.389, Dec. 2010, pp.5521-5529.

[39] R. Rejaie, M. Torkjazi, M. Valafar, W. Willinger, “Sizing up online social networks, ” IEEE Network. 24 (2010) 32-37.

[40] M. Goldstein, S. Morris, and G. Yen, “Problems with Fitting to the Power-Law Distribution,” *The European Physical Journal B,* vol.41, 2004, pp.255-258.

[41] B. Huberman, D. Romero, F. Wu, “Social networks that matter: Twitter under the microscope, ” First Monday. 14 (2009).

[42] J. Myers and A. Well, *Research Design and Statistical Analysis*, Lawrence Erlbaum, 2003.