Sampling Online Social Networks: An Experimental Study of Twitter

Maksym Gabielkov, Ashwin Rao, Arnaud Legout

Sampling Online Social Networks: An Experimental Study of Twitter

Maksym Gabielkov
Inria
Sophia Antipolis, France
maksym.gabielkov@inria.fr

Ashwin Rao
Inria
Sophia Antipolis, France
ashwin.rao@inria.fr

Arnaud Legout
Inria
Sophia Antipolis, France
arnaud.legout@inria.fr

ABSTRACT

Online social networks (OSNs) are an important source of information for scientists in different fields such as computer science, sociology, economics, etc. However, it is hard to study OSNs as they are very large. For instance, Facebook has 1.28 billion active users in March 2014 and Twitter claims 255 million active users in April 2014. Also, companies take measures to prevent crawls of their OSNs and refrain from sharing their data with the research community. For these reasons, we argue that sampling techniques will be the best technique to study OSNs in the future.

In this work, we take an experimental approach to study the characteristics of well-known sampling techniques on a full social graph of Twitter crawled in 2012 [2]. Our contribution is to evaluate the behavior of these techniques on a real directed graph by considering two sampling scenarios: (a) obtaining most popular users (b) obtaining an unbiased sample of users, and to find the most suitable sampling techniques for each scenario.

Categories and Subject Descriptors

C.2.m [**Computer-communication Networks**]: Miscellaneous

General Terms

Measurement, Experimentation

Keywords

Twitter, social networks, sampling, social graph.

1. INTRODUCTION

The number of users of OSNs is constantly increasing. It will be harder to study OSNs as the data grows bigger. Indeed, companies take measures to prevent the crawls of their social networks, e.g., Twitter has discontinued the API 1.0 that supported anonymous requests and the use of already whitelisted machines, the new API 1.1 requires user authentication for each request making crawls harder and longer to perform.

Also, the Twitter social graph is very different from the graphs of classical social networks, e.g., Facebook, because it is directed. There is no notion of friendship on Twitter, users can only *follow* other users to subscribe for their tweets. The action of following does not require any confirmation from the person being followed. Directed graphs are harder to study than undirected ones.

In 2012 we collected the full graph of Twitter, resulting in a graph with 505 million nodes and 23 billion arcs [2]. We use this graph to see how the classical sampling techniques are working with a limited sampling budget, that is a limited number of nodes that can be sampled.

Our contribution is to show that classical sampling techniques have a large bias toward high degree nodes. So classical techniques can be used to identify such nodes, but not to perform a uniform sample of the directed social graph. We describe the techniques in Section 2 and present the result of the application of these techniques to our dataset in Section 3.

2. SAMPLING TECHNIQUES

For our experimental study, we considered two classical techniques of sampling. The social graph of Twitter is directed, therefore each technique can traverse the graph in three different directions (a) in the direction of *followers* (the arcs of the graph will be directed from followers to the users being followed), (b) in the direction of *followings* (the arcs will be directed in the opposite way), (c) *bidirectionaly* (arcs will be treated as edges).

Breadth-first search (BFS) is a natural way to sample social networks, because of the way the APIs of these networks are designed. Usually the APIs provide access to list of friends/followers/followings of a particular user of a social network, which corresponds to the adjacency list of the node in the graph. For example, Kwak *et al.* [3] used BFS with Perez Hilton as a seed to crawl the Twitter social graph in 2009. However, when the number of requests to the API is limited, the sample obtained with BFS will have a large bias.

Random Walk (RW) is a way of traversing the graph from node to node by picking a random neighbor. Random walks are known to visit the nodes with the probability proportional to the degree of the node. There is a large literature on random walks for undirected graph, but very few on directed ones.

SIGCOMM'14, August 17–22, 2014, Chicago, IL, USA.
ACM 978-1-4503-2836-4/14/08.
http://dx.doi.org/10.1145/2619239.2631452.

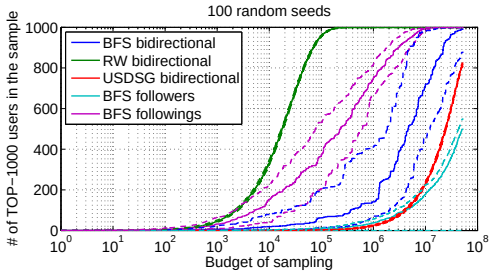

Figure 1: Sampling high degree users. The figure is based on 100 samples, solid line shows the median, dashed lines show the 25 and 75 percentiles.

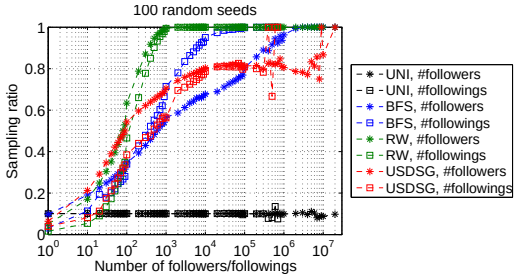

Figure 2: Sampling ratio for the bins with different number of followers/followings. The figure is based on 100 samples, lines show the median.

Unbiased sampling method for directed social graphs (USDSG) proposed by Wang *et al.* [4] is a modification of the random walk which discards a random jump to a node with a probability proportional to the degree of the node. The directed graph is treated as undirected, each arc is replaced by an edge.

3. RESULTS

We performed our experiments on the full Twitter social graph collected in 2012 that consists of 505 million accounts interconnected by 23 billion links [1]. For the purpose of this study, we took 100 random seeds and we performed each traversal 100 times starting from these seeds. We stopped the sampling when we sampled 10% of the graph (50.5 million nodes).

We do not present results of RW in the direction of followers and followings because RW stops when it encounters a node with no out-going arcs. In our experiments we observe such stops after sampling 50 nodes.

3.1 Popular users

To evaluate the ability of the techniques to sample popular users, we looked at the TOP 1000 most followed users in the sample, the result is presented on Figure 1. The best performance is achieved by RW, 342,000 nodes of the graph were crawled with RW to obtain all TOP 1000 most followed users. This behavior is explained by the fact that RW is biased towards high degree nodes. The BFS in all three directions has a high variance due to the position of the seed in the graph. Interestingly, BFS in the direction of followings can outperform RW when the budget is less than 2,000, because some of the Twitter users follow several popular accounts. So if we start the BFS from such a user, we can get with a limited budget a large number of popular users.

3.2 Unbiased sample

We now consider an unbiased sample in terms of the distribution of the number of followers and followings. To evaluate the quality of the distribution, we introduce a metric called *sampling ratio*. We distribute users in logarithmic bins based on the number of followers/followings they have. *Sampling ratio* is defined for each bin as the ratio of the number of sampled users in the bin to the total number of users in the bin. For the ground truth, we take a 10% uniform random sample of Twitter users (UNI on Figure 2).

Figure 2 presents the results for bidirectional versions of traversals. All the techiques we have tried have a high bias towards high degree nodes, including USDGS that fails to give an unbiased sample whereas it has been designed for that purpose.

We do not present BFS in the direction of followers and followings because some seeds do not have followers or followings resulting in an empty sample that can lead to the misinterpretation of the figure.

One may argue that the best way to obtain a uniform unbiased sample of the graph is to query the social network for randomly generated IDs (as shown in Figure 2 for UNI). However it is not always possible. For instance, the IDs on Facebook are assigned very sparsely, and only 75% of the account IDs in Twitter within the range $[0,\ 8 \times 10^8)$ correspond to valid accounts. Also companies may close the access to the IDs of the users to protect the privacy.

4. CONCLUSION

We have applied classical sampling techniques to the largest Twitter dataset ever collected. On the one hand, we showed that all classical sampling techniques introduce bias toward high degree nodes. This bias can completely change the results of the studies that rely on the partial crawl of the social graph. This motivates the need for a deeper study of the internal structure of social graphs to design an unbiased technique to sample directed OSNs.

On the other hand, the bias of these techniques towards high degree nodes gives a simple instrument to crawl high degree nodes, which correspond to popular users in the OSN.

More information about our study of Twitter can be found on our project page [1].

5. REFERENCES

[1] soTweet: Studying Twitter at Scale. `http://www-sop.inria.fr/members/Arnaud.Legout/Projects/sotweet.html`.

[2] M. Gabielkov, A. Rao, and A. Legout. Studying Social Networks at Scale: Macroscopic Anatomy of the Twitter Social Graph. In *Proc. of ACM Sigmetrics 2014*, Austin, TX, USA, Apr. 2014.

[3] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proc. of WWW'10*, Raleigh, NC, USA, May 2010.

[4] T. Wang, Y. Chen, Z. Zhang, P. Sun, B. Deng, and X. Li. Unbiased sampling in directed social graph. In *Proc. of SIGCOMM'10*, New Delhi, India, 2010.