Can Sampling Preserve Application Adoption Process over OSN Graphs?

Mohammad Rezaur Rahman, Chen-Nee Chuah
{mrrahman, chuah}@ucdavis.edu

Abstract

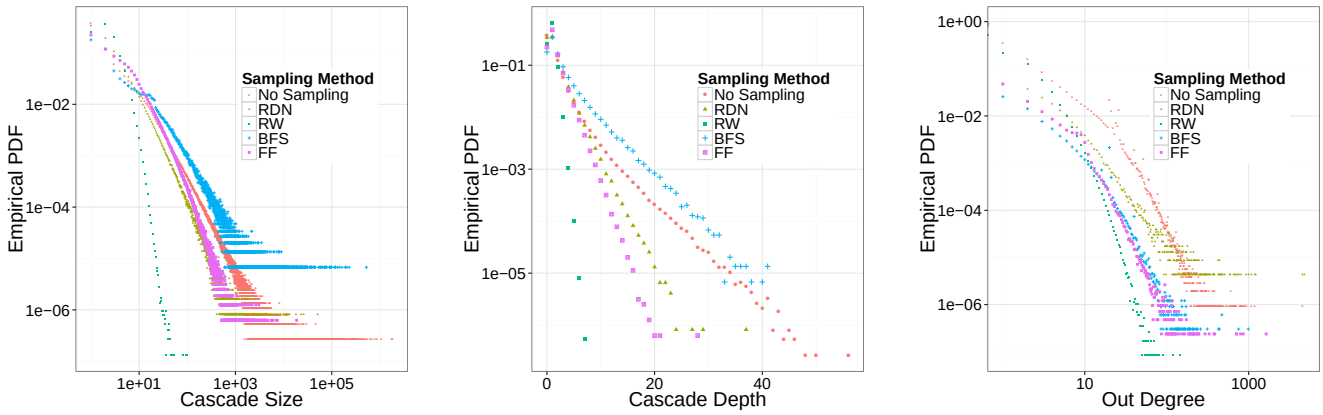
Online social network (OSN)-based applications often rely on user interactions to propagate information or to recruit more users. Understanding the *adoption* or *cascade* process of an idea, a product, or a new application over OSN graph is of great interest to advertisers, application developers, and OSN providers. Such adoption or information cascade process is an example of 'function' on OSN graphs. In this work, we investigate if existing graph sampling techniques known to preserve static graph properties can be equally effective in preserving dynamic 'functions' taking place over the OSN graphs.

There is a rich literature on sampling techniques that preserve static properties of social network graphs [1]–[4]. Leskovec et al. [1] examined how well various sampling methods preserve nine different graph properties, and measured the performance of those methods using D-statistic. It was found that Forest Fire algorithm performed best in preserving most of the graph properties [1]. Random walk is another commonly deployed technique in graph sampling [4]. Although different crawling and sampling techniques start with a single randomly chosen node, Ribeiro and Towsley [3] proposed to use multiple uniformly sampled starting nodes to randomly walk on the graph to prevent the walker from being trapped inside a small area of the whole graph.

As a first step towards understanding how well graph sampling techniques preserve 'functions' on networks, we consider OSN-application adoption process as a case study. In our previous work [5], we have performed detailed characterization of the adoption cascade of a popular Facebook gifting application, iHeart[1], in terms of the following properties: a) Cascade size distribution, b) Cascade depth distribution, and c) Out-degree distribution of the cascade seeds. In this work, we examine to what extent can existing sampling strategies capture a representative subset of user population to preserve some of the cascades properties mentioned above.

We explored the following graph sampling techniques with a modification to allow them to run on different communities of the application adoption process: a) Random Degree Node (RDN): nodes are chosen randomly with degree weight, b) Random Walk (RW): multiple uniformly chosen walkers walk on the graph, c) Breadth First Search (BFS): starts with multiple uniformly chosen initial nodes instead of a single node, and d) Forest Fire (FF): starts with multiple uniformly chosen initial nodes instead of a single starting node discussed by Leskovec et al. [1]; forward burning probability is 0.7. These sampling algorithms have been shown to preserve the degree distribution with some corrections for biasness [2], [6]. However, due to the fact that random node sampling is incapable of preserving connectivity [2], none of these algorithms seems to preserve any of the above application adoption properties, as illustrated in Fig. 1.

Fig. 1(a) shows that FF preserves the cascade size distribution closely, but overestimates the very small cascades and underestimate the very large ones. On the other hand, RDN, BFS, and RW all perform very poorly in preserving the cascade size distributions. Unfortunately, all methods except BFS underestimate the cascade depth distribution 1(b), and out-degree distribution of the seeds 1(c) is not preserved by any of the strategies. Our observations demonstrate the need for better graph sampling and streaming techniques that can preserve dynamic features associated with specific 'functions' on networks, as opposed to traditional graph properties.



(a) Cascade size distribution (b) Cascade depth distribution (c) Seeds' outdegree distribution

Fig. 1: Comparing application adoption properties with 10% sample size.

REFERENCES

[1] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, ser. KDD '06. New York, NY, USA: ACM, 2006, pp. 631–636. [Online]. Available: http://doi.acm.org/10.1145/1150402.1150479

[2] S. Lee, P. Kim, and H. Jeong, "Statistical properties of sampled networks," *Physical Review E*, vol. 73, no. 1, p. 016102, 2006.

[3] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proceedings of the 10th annual conference on Internet measurement*. ACM, 2010, pp. 390–403.

[4] L. A. A. LovÃ¡sz, "Random Walks on Graphs: A Survey," 1993. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.39.2847

[5] M. R. Rahman, P.-A. Noël, C.-N. Chuah, B. Krishnamurthy, R. M. D'Souza, and S. F. Wu, "Peeking into the invitation-based adoption process of osn-based applications," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 1, pp. 20–27, Dec. 2013. [Online]. Available: http://doi.acm.org/10.1145/2567561.2567565

[6] M. Kurant, A. Markopoulou, and P. Thiran, "Towards Unbiased BFS Sampling," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 9, pp. 1799–1809, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1109/jsac.2011.111005

[1]Detail of the data and the analysis can be found in [5]