An Improvement Direction for the Simple Random Walk Sampling: Adding Multi-homed Nodes and Reducing Inner Binate Nodes

Bo Jiao(✉), Ronghua Guo, Yican Jin, Xuejun Yuan, Zhe Han, and Fei Huang

Luoyang Electronic Equipment Test Center, Luoyang 471003, China
bjluoyang@hotmail.com

Abstract. Graph sampling is an important technology for the network visualization. In this paper, we use the normalized Laplacian spectrum to evaluate diverse biased sampling algorithms on Internet topologies, and numerically find that the simple random walk (SRW) sampling performs much better than other sampling algorithms (e.g., breadth first search, forest fire and random jump). Moreover, we analyze the deficiency of the SRW using the physical meaning of the normalized Laplacian spectrum on the size-independent Internet structure. Finally, we indicate that more multi-homed nodes should be added and more inner binate nodes should be reduced for better performance of the SRW sampling graphs on the normalized Laplacian spectrum which is a powerful tool for the study of size-independent structure embedded in evolving systems.

Keywords: Simple random walk · Normalized laplacian spectrum · Single-homed and multi-homed networks · Graph sampling

1 Introduction

The autonomous system level Internet topology evolves over time which induces that the network size (i.e., the node number) of the realistic Internet is very large. To more clearly realize the network visualization, we may need a sampling graph that has much smaller size and captures plenty of size-independent structures similar to the original large realistic network. There are two types of sampling algorithms: unbiased approach [1] and biased approach [2, 3, 4]. The unbiased approach strives to mathematically demonstrate that the sampling probability of each node is uniform, and is commonly applied when we can not observe all nodes of the original network (e.g., due to call limit of online social networks). On the other hand, the biased approach is biased towards high-degree nodes that will be sampled with higher probabilities, and has been widely evaluated using the performance comparison of graph metrics between the sampling and original networks. Leskovec et al. [2] used the Kolmogorov-Smirnov D-statistic to compare diverse biased sampling algorithms. Also, Xu et al. [3] used the

B. Jiao–The research field of Dr. Bo Jiao includes evolving network and spectral graph theory.

S. Wang and A. Zhou (Eds.): CollaborateCom 2016, LNICST 201, pp. 634–641, 2017.
DOI: 10.1007/978-3-319-59288-6_64

degree distribution and the clustering coefficient to derive an optimum solution in a hybrid biased sampling framework. However, these works focused on general scale-free networks. In other words, they have not considered the unique structure embedded in the Internet topology and can not be applied for the detailed analysis of the Internet sampling technology. In this paper, we apply diverse biased sampling algorithms in Internet topologies and evaluate them using the normalized Laplacian spectrum since our recent works [5, 6, 7, 8, 9] have demonstrated that the spectrum represents fruitful physical meanings for the unique size-independent structure of the Internet topology.

2 Background

2.1 Biased Sampling Algorithms

These algorithms can be classified into two categories: graph traversal and random walk. For connected Internet topologies, each node in the graph traversal is visited exactly once; examples include Breath First Search (BFS) and Forest Fire (FF). Whereas each node in the random walk can be revisited many times; examples include Simple Random Walk (SRW) and Random Jump (RJ).

The BFS is a classic graph traversal algorithm that starts from a randomly selected seed in the original network and progressively explores all neighbors [4]. At each new iteration of the BFS, the earliest explored but not-yet-visited node is selected next. Consequently, the BFS discovers first the nodes closest to the seed.

The FF is a randomized version of the BFS [4]. For every neighbor v of the current node, the FF decides explore v with probability p. When $p = 1$, the FF reduces to the BFS. It is possible that the FF dies out before it covers nodes with the expected number. Thus, we revive the FF from a random node already in the sample.

The SRW's walker starts from a randomly selected seed in the original network [2]. At each new iteration of the SRW, if the walker currently stays at node v, it randomly moves to a neighbor of node v with probability $1/d_v$ where d_v denotes the degree (i.e., the number of neighbors) of node v.

The RJ is similar to the SRW [3]. The only difference is that at each new iteration, with probability $c = 0.15$ (the value commonly used in literature), the RJ randomly jumps to any node in the original network and re-starts the random walk.

2.2 Normalized Laplacian Spectrum

Let $G = (V,E)$ denote an undirected and simple graph where V and E are respectively node set and edge set, d_v denote the degree of node v in G, D denote the diagonal degree matrix of G, and $A = (a_{ij})$ denote the adjacency matrix where $a_{ij} = 1$ if $(v_i,v_j) \in E$ and $a_{ij} = 0$ otherwise. Then, the normalized Laplacian matrix of G is $L(G) = D^{-1/2}(D\text{-}A)D^{-1/2}$ [11], and the normalized Laplacian spectrum of G includes all eigenvalues of $L(G)$: $0 = y_1 \leq y_2 \leq \ldots \leq y_n \leq 2$ where n denotes the node number of G. The WSD is a metric defined as $\sum_{i=1,2,\ldots,n}(1 - y_i)^N$ [11] where 4 is the best

selection of N [8], and the ME1 quantifies the number of the eigenvalue 1. Thus, we can determine that the spectrum can be described by two weakly-related spectral metrics (i.e., the ME1 and the WSD).

If G represents the Internet topology, we demonstrated that the ME1 reflects the node classification of G [5, 9]. Specifically, node set V can be classified into three subsets [10]: $P(G) = \{v \in V \mid d_v = 1\}$ called pendants, $Q(G) = \{v \in V \mid \exists\, w,(v, w) \in E, w \in P(G)\}$ called quasi-pendants, and $R(G) = V \setminus (Q(G) \cup P(G))$. Let $Inner(G) = (V_I, E_I)$ denote the subgraph of G induced by $R(G)$, and $d_I(v)$ denote the degree of node v in $Inner(G)$. Then, node set V_I (i.e., $R(G)$) can be further classified into six subsets [5]: $PI = \{v \in V_I \mid d_I(v) = 1 \wedge \forall(v, w) \in E_I, d_I(w) > 1\}$ called inner pendants, $QI = \{v \in V_I \mid \exists\, w, (v,w) \in E_I, w \in PI\}$ called inner quasi-pendants, $RI = \{v \in V_I \mid d_I(v) \geq 2 \wedge \forall(v,w) \in E_I, w \in QI\}$ called inner restricted nodes, $BI = \{v \in V_I \mid d_I(v) = 1 \wedge \forall(v,w) \in E_I, d_I(w) = 1\}$ called inner binate nodes, $II = \{v \in V_I \mid d_I(v) = 0\}$ called inner isolated nodes and $OI = V_I \setminus (PI \cup QI \cup RI \cup BI \cup II)$ call-ed inner noise nodes. Additionally, we demonstrated that the ME1 is approximately equal to the periphery number minus the core number where $P(G)$, PI, RI, II and half of BI compose the periphery, $Q(G)$, QI and another half of BI compose the core, and OI is the system noise [9]. Also, we demonstrated that the WSD monotonically decreases as the network is transformed from singe-homed to multi-homed [9].

3 Real-World Data

We extract 24 real-world Internet graphs from AS-Caida dataset [12]. These graphs were explored from Jan 2004 to Nov 2007. We use these graphs to analyze sampling algorithms since they have stable performances of the ME1 and the WSD, as shown in Fig. 1. The network sizes of these graphs span the range from 16,301 to 26,475.

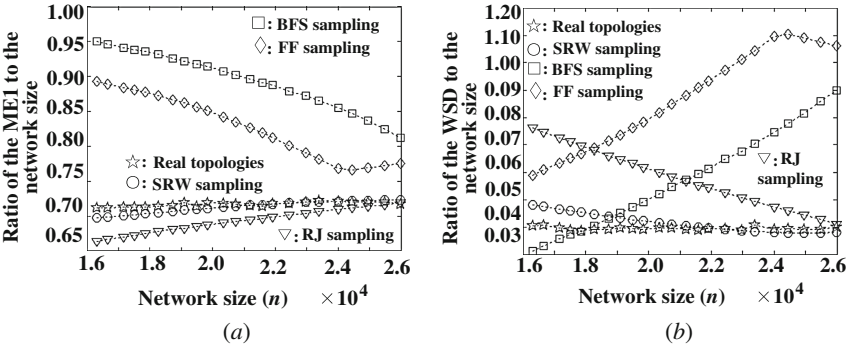

Fig. 1. Comparisons of the WE1 and the WSD on real-world and sampling graphs. (*a*) ME1/*n vs. n*. (*b*) WSD/*n vs. n*. The FF has $p = 0.7$ which is a good selection for $p \in \{0.5, 0.55, \ldots, 0.90\}$.

For each biased sampling algorithm, the largest real-world graph having 26,475 nodes is selected as the original network, and the network sizes of a sequence of sampling graphs are set to fall in the range from 16,301 to 25,988 which are consistent with those of the top 23 smallest real-world Internet graphs. For a certain network size of sampling graphs, each sampling algorithm runs ten times and the corresponding statistic is the average over ten realizations.

4 Comparison of ME1 and WSD on Sampling Algorithms

As shown in Fig. 1, the SRW performs much more stable on the normalized Laplacian spectrum compared to other biased sampling algorithms, and its corresponding curves are much closer to those of the real-world dataset. To subtly analyze the SRW, we compare the SRW with one mutation of the SRW, called Random Walk Flying Back (RWFB). The only difference between the SRW and the RWFB is that at each new iteration, with a probability c, the RWFB flies back to the original seed selected by the SRW and restarts the random walk [2]. As shown in Fig. 2, when $c = 0.1$, the RWFB performs much better than the SRW. The comparison between the SRW and the RWFB will be used to explore the improvement direction of this type of algorithms.

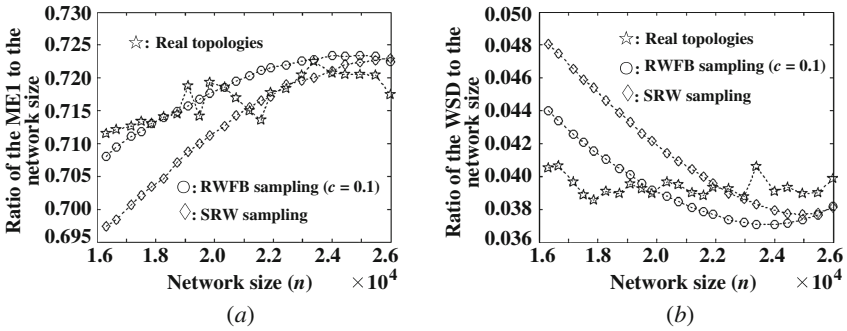

Fig. 2. Comparisons of the WE1 and the WSD on real-world and random walk sampling graphs. (*a*) ME1/*n vs. n*. (*b*) WSD/*n vs. n*.

5 Analysis for the Numerical Results

According to Sect. 2.2, for the Internet topology, node classification is an important feature reflected by the normalized Laplacian spectrum. Because the Internet topology has plenty of nodes with degree one, the pendant set $P(G)$ and the inner noise node set OI are respectively related to the largest and the smallest cardinalities. Thus, we analyze the two node set features of diverse sampling algorithms. The BFS is a graph traversal algorithm that constructs tree-like sampling graphs since each node is exactly visited once for this type of algorithms. Breath first principle induces small depths (where depth is defined as the maximum distance between root and leaves) of the tree-like sampling graphs. As is well known, trees with small depths have an extremely

large number of pendants (i.e., leaves with degree one). Additionally, this situation will lead to extremely small cardinalities of other node sets. Although the FF is a randomized version, its many performances remain similar to the BFS. Thus, for the BFS and the FF, their pendant numbers are extremely larger than that of the real-world dataset, and their inner noise node numbers are approximately equal to zero. The RJ is a hybrid algorithm of the SRW and the Random Node (RN) samplings. The SRW is biased towards high-degree nodes while the RN uniformly samples each node. Thus, the small-degree node number of the hybrid algorithm RJ is larger than that of the SRW. Note that the pendant set is an important component of small-degree nodes, which explains why the pendant number of the RJ is obviously larger than that of the real-world dataset. Therefore, the bad performances on node classification of the BFS, the FF and the RJ are critical reasons for the best performance of the SRW on the normalized Laplacian spectrum, as shown in Fig. 1.

According to Sect. 2.2, the Internet topology can be divided into eight node classifications. Specifically, pendant set $P(G)$, inner isolated node set II, quasi-pendant set $Q(G)$ and inner binate node set BI occupy the vast majority of the Internet nodes [9]. As shown in Fig. 3, we exhibit the evolving features of the four node sets on real-world and two random walk sampling graphs. Next, we will analyze the physical meaning embedded in Fig. 3 and investigate the improvement direction of the random walk sampling. Based on Fig. 3, for the RWFB, its pendant number, quasi-pendant number and inner binate node number are decreased and its inner isolated node number is increased in contrast to those of the SRW. The physical interpretation of these phenomena is presented in Fig. 4. In Fig. 4(*a*), each periphery node is attached to only one core node so these periphery nodes are single-homed. With the increasing of the links between periphery and core nodes, increasingly more nodes are transformed from single-homed to multi-homed, as shown in Fig. 4(*b*). As is well known, multi-homed nodes have better fault tolerance. Due to the rich club phenomenon, extremely few core nodes attract the majority of periphery nodes to connect with them. As shown in Fig. 4(*b*), inner binate nodes are generated by the small-degree core nodes which are connected with only one periphery node. If we remove the links between two inner binate nodes, in contrast to Fig. 4(*a*), all of the pendants, quasi-pendants and inner binate nodes will be reduced, and the inner isolated nodes will be added, as shown in Fig. 4(*c*). Thus, Fig. 4 explains why more added multi-homed nodes and more reduced inner binate nodes of the RWFB sampling graphs compared to those of the SRW induce the phenomenon shown in Fig. 3. The ME1 quantities the periphery number minus the core number [9], so the ME1 of Fig. 4(*c*) is obviously larger than that of Fig. 4(*a*) which verifies the phenomenon of Fig. 2(*a*). With the transformation from single-homed to multi-homed networks, the WSD monotonically decreases in general [9], so the WSD of Fig. 4(*c*) is commonly smaller than that of Fig. 4(*a*), which verifies the phenomenon of Fig. 2(*b*). Therefore, we can determine that adding multi-homed nodes and reducing inner binate nodes are the key reasons for the better performance of the RWFB in Fig. 2.

Although the RWFB performs better than the SRW on the normalized Laplacian spectrum, its stability of the evolving process on the spectrum is still unsatisfactory. Specially, as shown in Fig. 3, with the increasing of the size reduction ratio of the sampling graphs, the curves of the RWFB are father and father away from those of the real-world dataset. Additionally, the time complexity of the RWFB is very high since

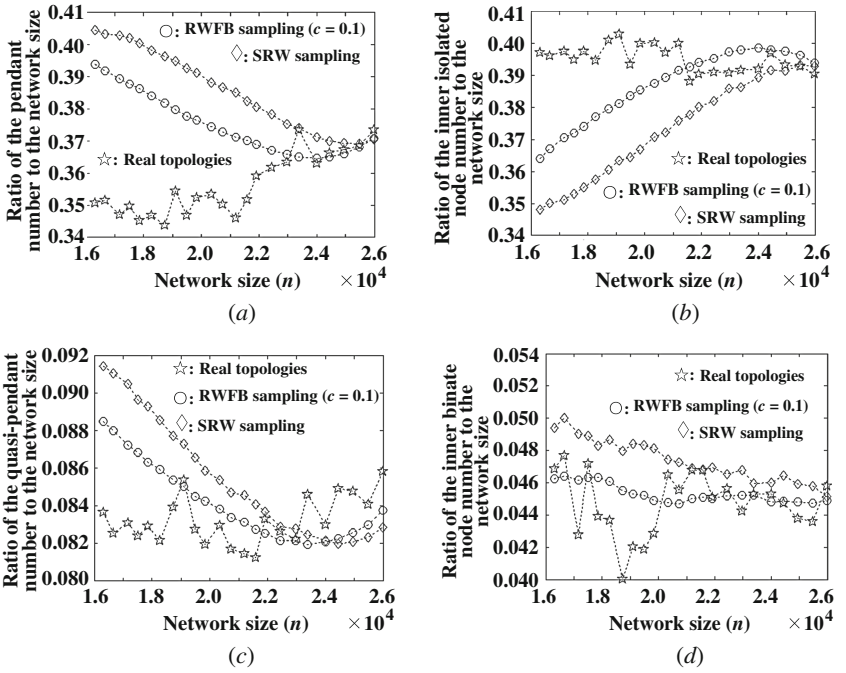

Fig. 3. Comparisons of the node classification on real-world and random walk sampling graphs. (*a*) pendant number/*n vs. n*. (*b*) inner isolated node number/*n vs. n*. (*c*) quasi-pendant number/*n vs. n*. (*d*) Inner binate node number/*n vs. n*.

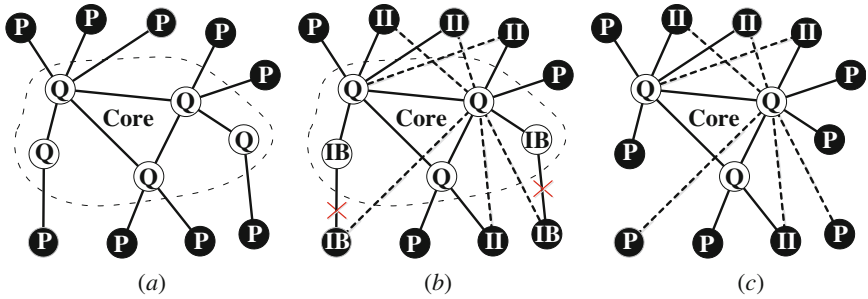

Fig. 4. Physical meaning embedded in Fig. 3. (*a*) A network with abundant single-homed nodes. (*b*) more multi-homed nodes are added. (*c*) More inner binate nodes are reduced. Note that white and black nodes respectively compose the core and periphery of the Internet, and P, Q, II and IB respectively denote pendant, quasi-pendant, inner isolated node and inner binate node.

flying back to the seed extremely increases the average visiting time of each node in the original network. However, based on the physical interpretation of Fig. 4, we can determine that adding multi-homed nodes and reducing inner binate nodes are valuable improvement directions for the simple random walk sampling algorithms.

Although only realistic autonomous system level Internet topologies with snapshots from Jan 2004 to Nov 2007 are analyzed in this paper, our recent studies [5, 9] verified that the physical meanings of the ME1 and the WSD hold for plenty of Internet evolving topologies. Specially, the core and periphery of the Internet (associated with the ME1) respectively are composed of the transit and stub nodes, which is consistent with the classical transit-stub model of the Internet [13]. Moreover, the transformation from single-homed to multi-homed (indicated by the WSD) reflects the Internet's requirement for better fault tolerance. Also, realistic Internet topologies derived from different data sources (e.g., AS-733, Oregon and AS-Caida) [12] keep plenty of similar size-independent structures [8]. Therefore, the derived results of this paper can be applied to more general cases of the Internet topology.

6 Conclusion

The normalized Laplacian spectrum is critical for evaluating graph sampling algorithms applied in the Internet visualization. In this paper, we use the spectrum to investigate the advantages and deficiencies of the SRW samplings and observe that the SRW and its mutation perform much better than other biased samplings. Additionally, based on the physical interpretation for the better performance of the RWFB, we indicate that adding multi-homed nodes and reducing inner binate nodes are important improvement directions for this type of SRW algorithms. In the future work, according to the improvement directions, we will design another mutation of the SRW which has better performance on the spectrum and higher runtime efficiency.

Acknowledgments. We would like to thank the anonymous reviewers for their comments that helped improve this paper. This paper is supported by the National Natural Science Foundation of China with Grant Nos. 61402485 and 61303061.

References

1. Lee, C.H., Xu, X., Eun, D.Y.: Beyond random walk and metropolis-hastings samplers: why you should not backtrack for unbiased graph sampling. ACM SIGMETRICS Perform. Eval. Rev. **40**, 319–330 (2012)
2. Leskovec, J., Faloutsos, C.: Sampling from large graphs. In: The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 631–636 (2006)
3. Xu, X., Lee, C.H.: A general framework of hybrid graph sampling for complex network analysis. In: 2014 Proceedings IEEE INFOCOM, pp. 2795–2803 (2014)
4. Kurant, M., Markopoulou, A., Thiran, P.: Towards unbiased BFS sampling. IEEE J. Sel. Areas Commun. **29**, 1799–1809 (2011)
5. Jiao, B., Zhou, Y., Du, J., et al.: Study on the stability of the topology interactive growth mechanism using graph spectra. IET Commun. **8**, 2845–2857 (2014)
6. Jiao, B., Nie, Y., Shi, J., et al.: Scaling of weighted spectral distribution in deterministic scale-free networks. Phys. A Stat. Mech. Appl. **451**, 632–645 (2016)

7. Jiao, B., Shi, J., Wu, X., et al.: Correlation between weighted spectral distribution and average path length in evolving networks. Chaos Interdisc. J. Nonlinear Sci. **26**, 023110 (2016)
8. Jiao, B., Nie, Y., Shi, J., et al.: Accurately and quickly calculating the weighted spectral distribution. Telecommun. Syst. **62**, 231–243 (2016)
9. Jiao, B., Shi, J.: Graph perturbations and corresponding spectral changes in internet topologies. Comput. Commun. **76**, 77–86 (2016)
10. Vukadinović, D., Huang, P., Erlebach, T.: On the spectrum and structure of internet topology graphs. In: Unger, H., Böhme, T., Mikler, A. (eds.) IICS 2002. LNCS, vol. 2346, pp. 83–95. Springer, Heidelberg (2002). doi:10.1007/3-540-48080-3_8
11. Fay, D., Haddadi, H., Thomason, A., et al.: Weighted spectral distribution for internet topology analysis: theory and applications. IEEE/ACM Trans. Networking **18**, 164–176 (2010)
12. Leskovec, J.: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data/
13. Calvert, K., Doar, M., Zegura, E.: Modeling internet topology. IEEE Trans. Commun. **35**, 160–163 (1997)