Hindawi
Security and Communication Networks
Volume 2022, Article ID 7003265, 7 pages
https://doi.org/10.1155/2022/7003265

Research Article

A Network Sampling Strategy Inspired by Epidemic Spreading

Qiang Dong,[1] **En-Yu Yu**,[1] **and Wen-Jun Li**[2]

[1]*School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China*
[2]*School of Software and Services Outsourcing, Suzhou Vocational Institute of Industrial Technology, Suzhou 215004, China*

Correspondence should be addressed to Qiang Dong; dongq@uestc.edu.cn

Received 3 November 2021; Accepted 31 December 2021; Published 28 February 2022

Academic Editor: Wei Wang

Nowadays, network sampling has become an indispensable premise and foundation for large-scale network analysis, and its effectiveness determines to a large extent the reliability and practicability of the subsequent network analysis results. In this paper, we propose a network sampling algorithm inspired by an epidemic spreading model named the contact process. The contact process is similar to the random walk process but different from it in two key points. First, at each time step, a randomly selected sampled node rather than the latest sampled node is responsible for recruiting a new node from its neighborhood. Second, the responsible node recruits one of its neighbor nodes with a probability inversely proportional to the degree of this neighbor node, instead of equal probability. Experiments on nine indiscriminately selected real-world networks show that our proposed sampling algorithm has a significant advantage in preserving two basic network properties, the degree distributions and clustering coefficient distributions of original networks, compared with seven classical sampling methods.

1. Introduction

In recent decades, the rapid development of storage technology has allowed online social network (OSN) providers to deposit almost all user-generated information every day. The analysis of OSNs is receiving remarkable research attention from both the academic and industrial communities. However, in some scenarios, some access restrictions are imposed on the network such that it is hard or infeasible for people to study the whole network. In other scenarios, the network is, however, available but too large to be stored and analyzed in a reasonable amount of memory and time. In the face of these problems, network sampling techniques have emerged to help us effectively and efficiently study and analyze real-world networks. The concept of network sampling can be simply described as follows. Given a network $G = (V, E)$ and the sampling ratio ρ, where $0 < \rho \ll 1$, the primary goal of network sampling is to construct a representative subnetwork $G_S = (V_S, E_S)$ which preserves the most important properties of the original network, where $V_S \subset V$, $E_S \subset E$, and $|V_S| = \rho * |V| = N_S$.

Nowadays, network sampling has become an indispensable premise and foundation for large-scale network analysis, and its effectiveness determines to a large extent the reliability and practicability of the subsequent network analysis results. Besides, network sampling also has a wide spectrum of applications, e.g., surveying hidden population in sociology, visualizing social graph, scaling down Internet AS graph, and graph sparsification [1].

A large number of sampling techniques have been proposed in the past few decades, designed for various purposes and for preserving different network properties [2]. These sampling techniques can be categorized into two groups: random selection and network exploration techniques. In the first group, nodes or links are recruited in the sample uniformly at random or proportional to some particular characteristic like degree or PageRank values [3]. In the second group, the sample network starts from a randomly selected seed node and is expanded following the local connections of previously sampled nodes.

Leskovec and Faloutsos [4] show that, among the typical network sampling methods, random walk (RW) and forest fire (FF) sampling methods have the best overall performance. Recently, Blagus et al. [3] empirically compared 11 representative network sampling methods on 12 real-world networks and concluded that breadth-first search (BFS) and

random walk with subgraph induction (RWI) sampling methods show the best overall performance in preserving the degree and clustering coefficient distribution of original networks. Next, we will briefly review these sampling methods, which will be used as comparing counterparts of our proposed algorithm.

Random walk (RW), forest fire (FF), and breadth-first search (BFS) are somewhat similar to each other. Initially, the sampled node set V_S and edge set E_S are both empty. At the first step, a randomly selected node is added into V_S. At each following step, for every node u added into V_S at the previous step, j randomly selected neighbor nodes of u, say $v_1, v_2, \ldots, v_j$, are added into V_S, and the corresponding edges (u, v_1), (u, v_2), $\ldots$, (u, v_j) are added into E_S. For RW, $j = 1$; for FF, j follows a geometric distribution with mean value $p/(1-p)$, where we set p to be 0.7 as suggested in [4]; for BFS, $j = k_u$, where k_u is the degree of node u. This step repeats until the sampling size is reached; that is, $|V_S| = \rho * |V|$. Another key different point is that the random walk is memoryless and a visited node has a probability of being visited again in the future, while the forest fire and breadth-first search never include the repeated nodes.

Besides the abovementioned methods, Metropolis–Hastings random walk (MHRW) is demonstrated to be a well-performed sampling method in the literature [5, 6]. It achieves a uniform distribution of sampled nodes by the following transition probability:

$$p_{u,v}^{\mathrm{MH}} = \begin{cases} \min\left\{\dfrac{1}{k_u}, \dfrac{1}{k_v}\right\}, & \text{if } v \text{ is a neighbor of } v, \\ 1 - \displaystyle\sum_{w \neq u} p_{u,w}^{MH}, & \text{if } u = v, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

Blagus et al. [3] proposed analyzing the sampling methods with subgraph induction, where the final sample network is constructed from a generated sample and all the existing edges between any two nodes of this sample. They empirically show that RW, FF, and MHRW with subgraph induction, named RWI, FFI, and MHRWI, improve the performance of the corresponding methods without subgraph induction. Therefore, RWI, FFI, and MHRWI are also used as baseline algorithms in this paper.

In this paper, we propose a network sampling algorithm inspired by an epidemic spreading model, the contact process, and thus, it is called contact process sampling (CPS). It is similar to RW but has two key different points. First, at each time step, a randomly selected sampled node rather than the latest sampled node is responsible for recruiting a new node from its neighborhood. Second, a sampled node chooses one of its neighbor nodes with a probability inversely proportional to the degree of this neighbor node, instead of an equal probability.

The rest of this paper is organized as follows. Section 2 describes the contact process and introduces the CPS algorithm. Section 3 compares the sampling quality of CPS with the aforementioned well-performed sampling methods. Section 4 concludes the whole work and makes some remarks.

2. Proposed Model

The contact process, which was first proposed as a susceptible-infected-susceptible (SIS) model for epidemic spreading, has found wide applications in science and engineering [7]. A general contact process on a network is described as follows. Initially, a set of nodes are infected by a virus (or carry a piece of information), and other nodes on the network are susceptible (not infected). At each time step, an infected node is chosen at random, say node u. With probability p, the virus on node u dies, and node u becomes susceptible again; with probability $1 - p$, the virus on node u selects one neighbor of u to contact, say v. If v is already infected, nothing happens; if v is susceptible, it gets infected.

In such a contact process, the fraction of infected nodes on a given network in an ultimately steady state is dependent on two critical factors: the aforementioned death rate p and the contact probability $W(k)$, which is the probability that an infected node chooses a neighbor node of degree k to contact. Yang et al. [7] proved that, when p is smaller than the threshold value, if the contact probability $W(k)$ takes the form of $W(k) \sim k^{\beta}$, the fraction of infected nodes in the ultimately steady state is maximized when $\beta = -1$.

In this paper, we propose a network sampling algorithm named contact process sampling (CPS), which employs a process analogous to the contact process across the network. In order to get a sample network of N_S nodes as soon as possible, we eliminate the effect of the death rate p from our CPS model by setting it to be 0. To ensure the connectedness of a sample network, the CPS algorithm starts from only one node.

The CPS algorithm is presented by Algorithm 1. Initially, a randomly chosen node is recruited into the sample set. At each time step, a sampled node is chosen at random, say node u. Following the conclusion of Yang et al., node u chooses a neighbor node v to recruit into the sample set with probability $p_{u,v} = k_v^{-1} / \sum_{w \in \Gamma_u} k_w^{-1}$, where k_v is the degree of node v and Γ_u represents the set of u's neighbor nodes. This recruitment step repeats until the sample set contains N_S distinct nodes. Then, we construct the final sample network with these sampled nodes and the links which connect any two of these sampled nodes in the original network.

3. Performance Evaluation

3.1. Datasets. Nine indiscriminately selected real-world networks from KONECT [8] are employed to test the performance of sampling models. They are all undirected and unweighted networks, and their basic statistics are presented in Table 1. The fill of a network is the proportion of edges to the total number of possible edges. The global clustering coefficient is defined as the probability that two incident edges are completed by a third edge to form a triangle. Assortativity is defined as the Pearson correlation coefficient between the degrees of connected nodes. For

Input: an undirect and unweighted graph $G = (V, E)$; the sample ratio ρ, where $0 < \rho < 1$;
Output: a sample graph $G_S = (V_S, E_S)$, where $V_S \subseteq V$, $E_S \subseteq E$ and $|V_S| = \rho * |V|$;
(1) Randomly select one node $u_0 \in V$, and let $V_S = \{u_0\}$ and $E_S = \varnothing$;
(2) **while** $|V_S| = \rho * |V|$ **do**
(3) Randomly select one node v from V_S;
(4) Select one node w from the neighborhood of v, with probability inversely proportional to the degree of w;
(5) $V_S = V_S \cup \{w\}$;
(6) **end while**
(7) **for** $(x, y) \in E$ **do**
(8) **if** $x \in V_S$ and $y \in V_S$ **then**
(9) $E_S = E_S \cup \{(x, y)\}$;
(10) **end if**
(11) **end for**
(12) **return** $G_S = (V_S, E_S)$;

Algorithm 1: Contact process sampling.

Table 1: Basic statistics of real-world networks used in this paper.

Network	Category	Nodes	Edges	Fill	Avg. degree	Global clust. (%)	Assortativity
PowerGrid	Infrastructure	4941	6594	5.40×10^{-4}	2.669	10.30	0.003 46
Amazon	Miscellaneous	334 863	925 872	1.65×10^{-5}	5.530	20.50	−0.05882
WordNet	Lexical	146 005	656 999	6.16×10^{-5}	9.000	9.58	−0.06233
AstroPh	Coauthorship	18 771	198 050	1.12×10^{-3}	21.102	31.80	0.205 13
Livemocha	Social	104 103	2193 083	4.05×10^{-4}	42.133	1.41	−0.14677
Gowalla	Social	196 591	950 327	4.92×10^{-5}	9.668	2.35	−0.02926
Brightkite	Social	58 228	214 078	1.26×10^{-4}	7.353	11.10	0.010 82
Douban	Social	154 908	327 162	2.73×10^{-5}	4.224	1.04	−0.18033
Flickr	Miscellaneous	105 938	2316 948	4.13×10^{-4}	43.742	40.20	0.246 85

other characteristics of these networks, the reader is referred to KONECT [8].

In this paper, we consider the sample ratio ranging from 0.2% to 20% of original networks (by step of 0.2% in 0.2% ~ 1% and 2% in 2% ~ 20%) as suggested in [3]. For each network, we perform 30 realizations of each sampling technique and each sample ratio. For each run of the exploration techniques, the sample starts from a randomly selected new seed node.

3.2. Evaluation Measures. For the evaluation of sampling algorithms, two well-known and widely used network statistics are used to measure the representativeness of the sampled network. They are degree distribution (DD) as a global statistical property and clustering coefficient distribution (CCD) as a local statistical property. The DD of a network refers to the probability distribution of degrees of all nodes in the network [9] and is represented by the fraction p_k of nodes of degree k, $k > 0$. The clustering coefficient of a node in a network is the proportion of that node's neighbors that are connected, and the CCD of a network refers to the probability distribution of the clustering coefficient of all nodes in the network [10].

We compare the DD and CCD of the sample network and the original network by the Kolmogorov–Smirnov D-statistic (KSD). KSD is used to measure the agreement of two cumulative distribution functions [11]: original distribution F_1 and estimated distribution F_2. It is defined as $\text{KSD} = \max_x\{|F_1(x) - F_2(x)|\}$, where x is over the range of the random variable. Clearly, it is a value between 0 and 1. The closer it is to zero, the higher is the similarity between the two distributions. Note that KSD does not address the issue of the scaling but rather compares the shape of the (normalized) distribution [4].

3.3. Algorithm Comparison. The comparison of sampling techniques based on degree distribution is shown in Figure 1. We can see that, in most datasets, the techniques without subgraph induction (RW, FF, and MHRW) perform significantly different from other methods. This group of techniques approximates the degree distribution of the original networks with a larger deviation than others (except for PowerGrid and Douban). Therefore, this observation reinforces the conclusion of Blagus et al. [3] that the techniques with induction improve the performance of the corresponding techniques without it.

As for the performance of techniques with subgraph induction, the nine datasets can be categorized into two groups. In the first group of datasets (PowerGrid, Amazon, WordNet, AstroPh, and Livemocha), the techniques with subgraph induction perform similarly to each other, and our proposed CPS algorithm is the best in three datasets

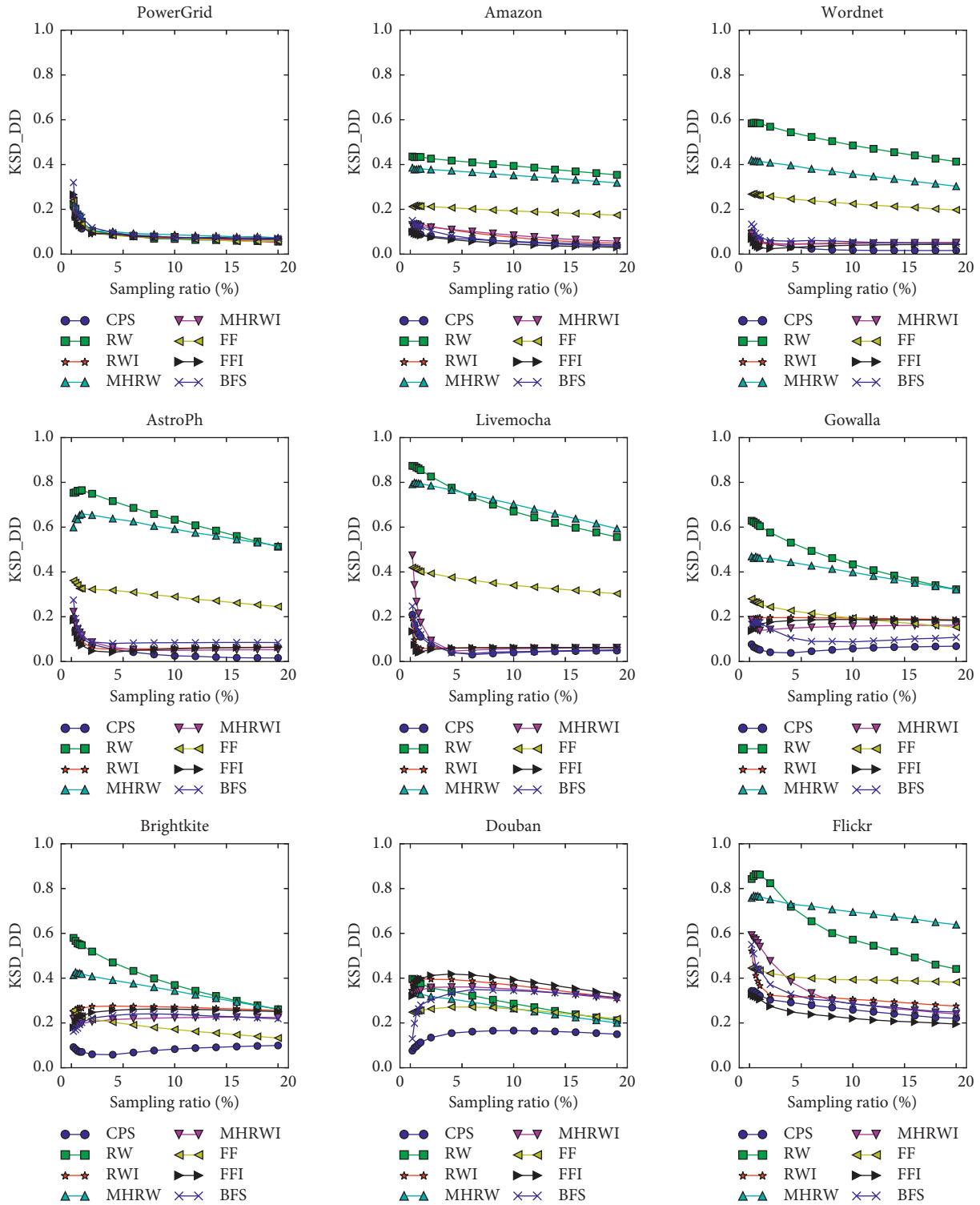

Figure 1: Comparison of sampling techniques based on degree distribution.

(WordNet, AstroPh, and Livemocha) and has a negligible difference from the best ones in the other two datasets. In the second group of datasets (Gowalla, Brightkite, Douban, and Flickr), the techniques with subgraph induction perform greatly different from each other. Our proposed CPS algorithm shows a significant advantage in three datasets (Gowalla, Brightkite, and Douban) and is the second-best in Flickr. In general, our proposed CPS algorithm is the best performing technique in preserving the degree distribution of the original networks.

The comparison of sampling techniques based on clustering coefficient distribution is shown in Figure 2. Similarly

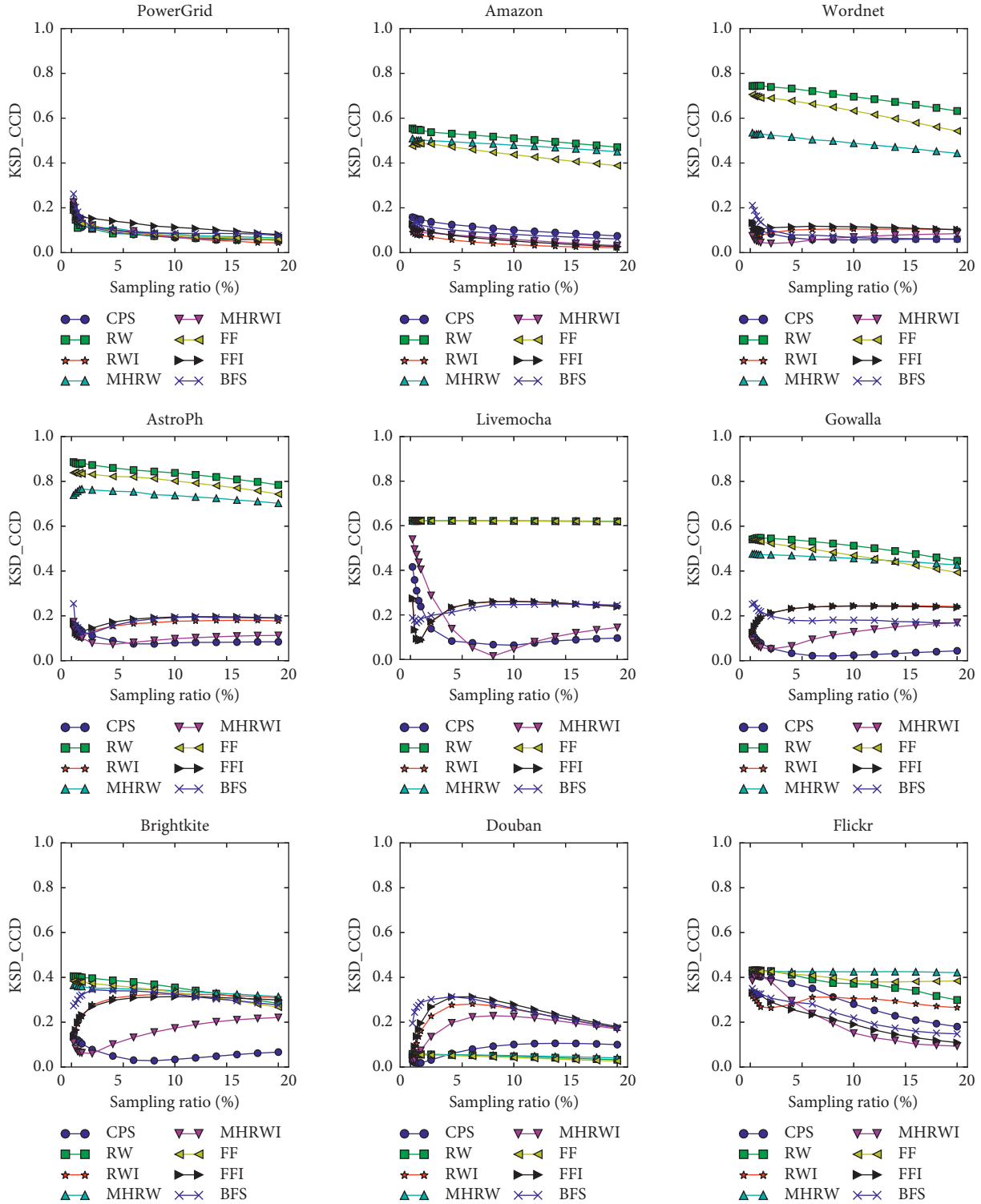

Figure 2: Comparison of sampling techniques based on clustering coefficient distribution.

to that of degree distribution, the techniques with subgraph induction perform better than the corresponding techniques without subgraph induction in most datasets (except for PowerGrid and Douban). The three techniques without subgraph induction, RW, FF, and MHRW, unless otherwise specified, are excluded from our following discussion.

The five techniques with subgraph induction have a similar declining shape of KSD plot in 3 datasets, PowerGrid, Amazon, and Flickr, where the CPS algorithm is comparable to other techniques. In contrast, in the other 6 networks, the KSD plots of RWI, FFI, and BFS algorithms begin to increase with the growth of sampling ratio, except for BFS in

Table 2: The D-statistics of degree distributions of 8 sampling algorithms on 9 real-world datasets, where the sampling ratio is 10%.

Network	RW	RWI	FF	FFI	MHRW	MHRWI	BFS	CPS
PowerGrid	**0.068**	0.071	0.075	0.075	0.086	0.077	0.072	0.072
Amazon	0.394	0.075	0.193	**0.045**	0.352	0.084	0.055	0.057
WordNet	0.485	0.050	0.225	0.040	0.358	0.048	0.055	**0.018**
AstroPh	0.633	0.059	0.290	0.056	0.591	0.050	0.084	**0.026**
Livemocha	0.670	0.062	0.340	0.062	0.702	0.057	0.044	**0.040**
Gowalla	0.434	0.194	0.194	0.188	0.398	0.158	0.088	**0.058**
Brightkite	0.369	0.271	0.171	0.261	0.343	0.223	0.238	**0.083**
Douban	0.286	0.368	0.263	0.393	0.265	0.351	0.344	**0.166**
Flickr	0.572	0.305	0.393	**0.220**	0.695	0.285	0.285	0.258

Table 3: The D-statistics of clustering coefficient distributions of 8 sampling algorithms on 9 real-world datasets, where the sampling ratio is 10%.

Network	RW	RWI	FF	FFI	MHRW	MHRWI	BFS	CPS
PowerGrid	0.072	0.069	0.076	0.113	0.082	0.070	0.086	**0.067**
Amazon	0.510	**0.036**	0.437	0.052	0.479	0.059	0.082	0.101
WordNet	0.696	0.105	0.633	0.115	0.488	0.070	0.067	**0.057**
AstroPh	0.838	0.176	0.802	0.195	0.737	0.098	0.192	**0.079**
Livemocha	0.622	0.261	0.621	0.261	0.621	**0.047**	0.246	0.064
Gowalla	0.513	0.245	0.469	0.243	0.456	0.129	0.181	**0.024**
Brightkite	0.355	0.326	0.334	0.314	0.340	0.174	0.325	**0.034**
Douban	0.048	0.259	**0.042**	0.279	0.051	0.226	0.261	0.101
Flickr	0.372	0.305	0.386	0.190	0.425	**0.151**	0.220	0.281

WordNet and Gowalla. Fortunately, the KSD plots of the CPS algorithm are always declining in these 6 networks, and the KSD values are very small compared with those of the RWI, FFI, and BFS algorithms. The performance of MHRWI is intermediate between that of CPS and those of RWI, FFI, and BFS, where the KSD value is closer to the former, and the shape of the KSD plot is similar to the latter. In general, the CPS algorithm has the best overall performance in preserving the clustering coefficient distribution of the original networks.

To quantitatively demonstrate the superiority of CPS to other methods, Tables 2 and 3 present the KSD values of DD and CCD produced by 8 sampling techniques on 9 datasets when the sampling ratio is 10% as suggested in [3], where the best and second best values for every dataset are highlighted in bold type. For DD, the CPS algorithm is the best in 6 out of 9 datasets, and for CCD, the CPS algorithm is the best in 5 out of 9 datasets. Other than CPS, no algorithm is ranked in the first position in more than 2 datasets, whether DD or CCD or both. Recall that the nine real-world network datasets are indiscriminately selected from KONECT [8]. We conclude that the CPS algorithm has a significant advantage in preserving degree distributions and clustering coefficient distributions of original networks.

4. Concluding Remarks

In this paper, we proposed a network sampling strategy inspired by the contact process and empirically validated its superior performance in preserving two important structural properties of original networks. Although it is a little similar to random walk sampling, two key different operations from RW make it produce better sample network than RW and several typical RW-variant sampling methods.

There is much work that remains to be done in the future. First of all, test of the CPS algorithm in preserving other properties of the original network would be useful to show possible limits of its applicability. Second, one should also investigate the characteristics of some datasets, typically Douban, where the sampling methods without subgraph induction perform better than the ones with subgraph induction. Finally, but not the least, the function approximation of a sample network is worthy of exploration. For example, the comparison of the epidemic spreading process on sample and original networks may be the topic of our next work [12].

Data Availability

The readers can access all the 9 datasets supporting the conclusions of the study from http://konect.cc/, and the details are listed as follows: PowerGrid, http://konect.cc/networks/opsahl-powergrid/; Amazon, http://konect.cc/networks/com-amazon/; WordNet, http://konect.cc/networks/wordnet-words/; AstroPh, http://konect.cc/networks/ca-AstroPh/; Livemocha, http://konect.cc/networks/livemocha/; Gowalla, http://konect.cc/networks/loc-gowalla_edges/; Brightkite, http://konect.cc/networks/loc-brightkite_edges/; Douban, http://konect.cc/networks/douban/; and Flickr, http://konect.cc/networks/flickrEdges/

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. Hu and W. C. Lau, "A survey and taxonomy of graph sampling," 2013, https://arxiv.org/abs/1308.5865.

[2] D. D. Heckathorn and C. J. Cameron, "Network sampling: from snowball and multiplicity to respondent-driven sampling," *Annual Review of Sociology*, vol. 43, no. 1, pp. 9.1–9.19, 2017.

[3] N. Blagus, L. Šubelj, and M. Bajec, "Empirical comparison of network sampling: h," *Physica A: Statistical Mechanics and Its Applications*, vol. 477, pp. 136–148, 2017.

[4] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 631–636, Seoul, South Korea, May 2006.

[5] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in facebook: a case study of unbiased sampling of OSNs," in *Proceedings of the 2010 IEEE INFOCOM*, pp. 1–9, IEEE, San Diego, CA, USA, March 2010.

[6] A. H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Respondent-driven sampling for characterizing unstructured overlays," in *Proceedings of the INFOCOM 2009*, pp. 2701–2705, IEEE, Rio de Janeiro, Brazil, April 2009.

[7] R. Yang, T. Zhou, Y. B. Xie, Y. C. Lai, and B. H. Wang, "Optimal contact process on complex networks," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 78, no. 6, Article ID 066109, 2008.

[8] J. Kunegis, "KONECT-the koblenz network collection," in *Proceedings of the 22nd international conference on World Wide Web companion*, pp. 1343–1350, Rio de Janeiro, Brazil, May 2013.

[9] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science (New York, N.Y.)*, vol. 286, no. 5439, pp. 509–512, 1999.

[10] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[11] N. K. Ahmed, J. Neville, and R. Kompella, "Network sampling: from static to streaming graphs," *ACM Transactions on Knowledge Discovery from Data*, vol. 8, no. 2, pp. 1–56, 2013.

[12] Z. S. Jalali, A. Rezvanian, and M. R. Meybodi, "Social network sampling using spanning trees," *International Journal of Modern Physics C*, vol. 27, no. 5, Article ID 1650052, 2016.