Computer Science and Application 计算机科学与应用, 2022, 12(11), 2633-2645
Published Online November 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.1211267

Hans汉斯

社交网络子图采样算法

李　准，雷晓颖

扬州大学，江苏　扬州

收稿日期：2022年10月26日；录用日期：2022年11月23日；发布日期：2022年11月30日

摘　要

在线社交网络(Online Social Networks, OSNs)数据量庞大，如何以低成本采样获取具有代表性的子图成为当前一个研究热点。现有的大部分采样算法仅仅体现在样本列表的无偏特性上，很多情况下，其采样样本构造的诱导子图难以代表原图结构。本文对各种类型的OSN提出了一种新的采样方法，该算法在原有的随机游走算法基础上重新计算了采样跳转概率，修正采样诱导子图的偏差，使其能够更出色地代表原图。同时，本文的采样算法通过计算权重的方式采集邻接节点，省去了自循环过程，从而大幅度提高了采样效率。实验结果表明，本论文提出的采样算法在度分布、聚类系数、传递性、同配性各个方面综合对比，采样获取的子图更加接近原图的属性结构。最后，该算法在大多数情况下，其性能与表现均优于现有采样算法。

关键词

在线社交网络(OSNs)，顶点采样，蒙特卡罗随机游走(MHRW)，无偏

Subgraph Sampling Algorithm of Social Network

Zhun Li, Xiaoying Lei

Yangzhou University, Yangzhou Jiangsu

Received: Oct. 26th, 2022; accepted: Nov. 23rd, 2022; published: Nov. 30th, 2022

Abstract

Online Social Networks (OSNs) maintain a huge amount of data, and how to sample the most representative subgraphs at a lower cost becomes an important research topic. Most of the existing sampling algorithms are only embodied in the unbiased characteristics of the sample list. In many cases, the induced subgraphs constructed from the sampled samples are difficult to represent the original graph structure. This paper proposes a new sampling method for various types of OSNs. This algorithm recalculates the sampling jump probability on the basis of the original random walk algorithm, and corrects the deviation of the sampling-induced subgraph, as a result, it can better represent the original graph. At the same time, the sampling algorithm in this paper collects the adjacent nodes by calculating the weight, eliminating the self-circulation process and greatly improv-

文章引用: 李准, 雷晓颖. 社交网络子图采样算法[J]. 计算机科学与应用, 2022, 12(11): 2633-2645.
DOI: 10.12677/csa.2022.1211267

ing the sampling efficiency. The experimental results show that the sampling algorithm proposed in this paper is more close to the original image attribute structure in terms of degree distribution, cluster, transitivity, and assortativity. Finally, in most cases, the algorithm perfects better than the existing sampling algorithm.

Keywords

Online Social Networks (OSNs), Vertex Sampling, Metropolis-Hastings Random Walk (MHRW), Unbiased

1. 引言

近些年来，在线社交网络(Online Social Networks, OSNs)成为人们日常交际、信息传播的重要平台，数以亿计的用户以及他们的朋友每天拥有海量的数据信息传播量。研究社交网络特性，分析社交网络应用，成为当前热门话题。一些研究者对 OSN 建模与度量，研究网络拓扑之间的结构，分析社交网络的影响力[1]，预测社交网络的发展趋势[2]。部分学者通过挖掘 OSN 传播性质与结构性质，研究其传播领域、用户行为与隐私等方面[3]。Viswanath 等人[4]通过区分弱连接与强连接对 Facebook 数据集的活动网络进行了时序上的研究，发现大部分社交联系强度都会随着时间逐渐变弱，然而社交网络图结构基本上不会变化。研究社交网络的结构属性、传播特性、演变规律以及用户之间的交互属性[5]，可以应用到现实生活诸多领域，如疫情风险监控、股市预测、推荐适合广告等各种场景[6] [7]。

截至 2021 年 6 月 30 日，Facebook 月度活跃用户(MAU)达到 29.0 亿人，Twitter 也有着数以亿计的用户量。然而面对如此庞大的 OSN，如果直接对其进行整体研究，将是一个不可估算的工程量。并且，为了保护用户个人信息，很多 OSN 公司都对部分数据进行了限制访问与隐私保护。因此，在有限的计算环境约束之下，如何才能获取一个代表性的子图样本，用以估算原图，成为目前正在研究的热门话题。

2. OSN 采样

通常把 OSN 建模为社交图，用 G = (V, E)表示，其中，顶点集合 V 表示 OSN 中的用户，边集合 E 表示用户之间的关系或连接，这种关系根据图的类型分为有向与无向。对于无向图，顶点 v 的邻接节点数量称为节点 v 的度，其中，v∈V。对于有向图，定义节点 v 出发指向其邻接节点的数量称之为 v 的出度，从其他邻接节点出发指向 v 的节点数量称为 v 的入度。

为了获取一个具有代表性的样本，学者们先后研究出各种算法进行测评。一开始最常使用广度优先搜索算法(Breath First Search, BFS)，Ahn 等人[8]通过采样获得网络图拓扑结果，分析了 OSN 中用户的交流状况与发展趋势。Ferrara 等人[9]对 Facebook 数据集进行采样，分析研究了社区网络的统计特征与个人的组织模式。同时，Chau 等人[10]研究并行爬行采样机制，可以大幅度提高网络效率，然而效率提高并不能改变采样结果的评测。经研究[11]比较 BFS 采集到的子图度数偏高，在采样比率不足时，度分布存在较大的误差[12]。部分学者使用用户均匀采样方法(Uniform Sampling of User IDs, UNI)，以获得对于网络的无偏估计，然而网络用户 id 的长度复杂性导致采样方法的命中率相当之低，且频繁的拒绝－接受采样又会导致带宽消耗较大，文献[13] [14]提出通过改进 UNI 自适应方法，提高采样命中率，需要研究网络 id 的大致分布有足够的研究。随机游走算法(Random Walk, RW)广泛运用于 OSN 采样，在计算机科学、数学物理和金融领域等多方面进行展示

与应用[15]，然而 RW 采样获取到的子图也是偏于高度节点[16]，部分学者通过改进随机游走或者重新估算样本的方法对网络进行采样。Salehi 等人[17]提出一种新的链接跟踪采样方法，将网络进行社区划分，然后分别在每个社区进随机游走，用以研究网络特性，需要事先了解网络结构使用聚类划分的策略[18]。文献[19]采用将静态网络拓展为时态网络的方法，基于个性化 PageRank，使用随机游走采样理论分析。Gjoka 等人[11][16]通过研究对比普通的游走算法、重新加权随机游走算法(Re-Weighted Random Walk, RWRW)、以及蒙特卡罗随机游走(Metropolis-Hastings Random Walk, MHRW)对社交网络图进行了采样与检测，研究表明，这两种方法均可获得一个不错的样本用以估算原图结构，然而 RW 容易陷入局部重复采样，MHRW 有较高的自环率，RWRW 本身可用的估算因子有限。Dong 等人[20]使用 USRS 采样算法，通过跳转之前将一部分节点的自环率分配给邻接节点中自环率大于 0 的节点，从而提高采样效率，USRS 不仅需要访问当前节点的邻接节点，还需要访问其邻接节点的邻接节点，计算出自环率的分配值，该过程较为复杂且易受到 OSN 访问权限的影响。Jin 等人[21]使用 AS 采样算法，在 MHRW 算法基础上加入了随机跳转策略，提高了采样效率，同时避免陷入局部子图死循环无法出来的情况。文献[22]用 UD 算法通过添加缓存区进行同度数节点替换的方法，实现采样率提高，然而在一些 OSN 图采样时，所得到的子图平均度数与聚类系数是均是明显偏低的。Rezvanian 等人[23]研究最短路径，随机选取网络的两个节点，将最短路径所涉及的节点记录，最后进行排序对网络进行研究与分析，以此获取的比较好的样本，最短路径需要实现知道网络结构以及计算过程较为复杂。

综合以上采样算法研究对比，各有优势，也同时有些不足之处，例如，采样效率、运行时间、算法复杂性、参数可调性、偏差性等等，在每个评测方面难以同时维持最佳性能效果[24]。实际对 OSN 采样的时候，又可能受到网络权限的访问限制，以及频繁的网络请求可能会导致采样难以继续。本论文基于随机游走算法，为了修正样本构造的诱导子图带来的误差问题，本文对节点的跳转权重重新进行了调整，弥补了某些情况下，MHRW 算法自环率较高的缺陷，在提高了采样效率的同时，依旧保持采集子图具有很好的代表性，能够还原原图。

3. 相关研究

3.1. 随机游走(RW)

随机游走是目前广泛运用于 OSN 的采样算法，选择若干初始种子节点，然后从当前节点出发，每次随机选择其中一个邻接节点作为下一个跳转的目标，每次跳转的概率为 $1/k_u$ (k_u 节点 u 的度数)，转移概率公式如下：

$$p_{uv}=\begin{cases}\dfrac{1}{k_u} & \text{if } v \text{ is a neighbor of } u\\ 0 & \text{otherwise}\end{cases} \tag{1}$$

简单的随机游走算法能够在短时间内构造一个诱导子图，然而这样的采样是偏向于高节点的[11][16]，其随机游走采样算法服从概率分布：

$$\pi_u=\frac{d_u}{2\cdot|E|} \tag{2}$$

每个节点的采样概率均与其度数成正比，该算法并不是无偏的，对于度数越高的节点，被采样到的概率越高。文献[20]显示，简单随机游走在每一步对于每个节点选中概率为 1/du，当随走过程收敛后，所采集到的样本平均度数与原图节点度数平方成正比：

$$\frac{\sum_{u\in S}k_u}{N_s}\langle k\rangle=\frac{\sum_{u\in G}k_u^2}{N} \tag{3}$$

其中，k_u 表示当前节点度数，S 表示采集到的样本，G 表示原图，d 表示原图的平均度数，N 表示节点数，N_s 表示样本节点数，$\langle k \rangle$ 表示原始网络图的平均度数。

3.2. 蒙特卡罗随机游走(MHRW)

基于 Metropolis-Hastings 对于随机游走进行一种改进。该算法从当前节点开始，选择一个邻接节点作为目标节点，产生 0~1 之间的随机数 p，若 p 小于等于当前节点度数与目标节点度数的比值，则跳转至目标节点，否则停留在当前节点，重复迭代以上过程。转移概率公式如下：

$$p_{uv} = \begin{cases} \min\left(\frac{1}{k_u}, \frac{1}{k_v}\right), v \text{ is a neighbor of } u \\ 1 - \sum_{v=u} \min\left(\frac{1}{k_u}, \frac{1}{k_v}\right), v = u \\ 0, \text{otherwise} \end{cases} \tag{4}$$

在任意时刻 t，当前游走到节点 u 时，其邻接节点 v_i 的跳转概率为$(P_{v1},P_{v2},\cdots,P_{vn})$。当 $K_v \le K_u$ 时，$P_v = 1/k_u$，当 $K_v > K_u$ 时，$P_v = 1/k_v$。在 MHRW 算法构建的状态转移矩阵中，从图中所有点到达节点 v 的概率之和为 1，并且对于图中的任意点 u 和 v，均有 $P_{uv} = P_{vu}$，该过程可以建模为马尔科夫链模型[20]，当 t 趋于无穷大时，每个节点的采样率均相等，其静态分布收敛于 $\pi_u = 1/|V|$，由此算法所采集到的样本能够精确估算原图度分布。然而高度自循率以及重复采样，大幅度降低了算法的采样效率，MHRW 算法高度重复的节点在生成诱导子图的时候只生效一次，有时并不能很好地代表原图。

4. RWSS 算法

4.1. 问题描述

本文针对采样子图中的度数偏差以及采样效率问题展开研究。

首先，本文使用 MHRW 算法对 Twitter 数据集进行采样，比较采样节点在原图的度分布以及其生成诱导子图的度分布图。

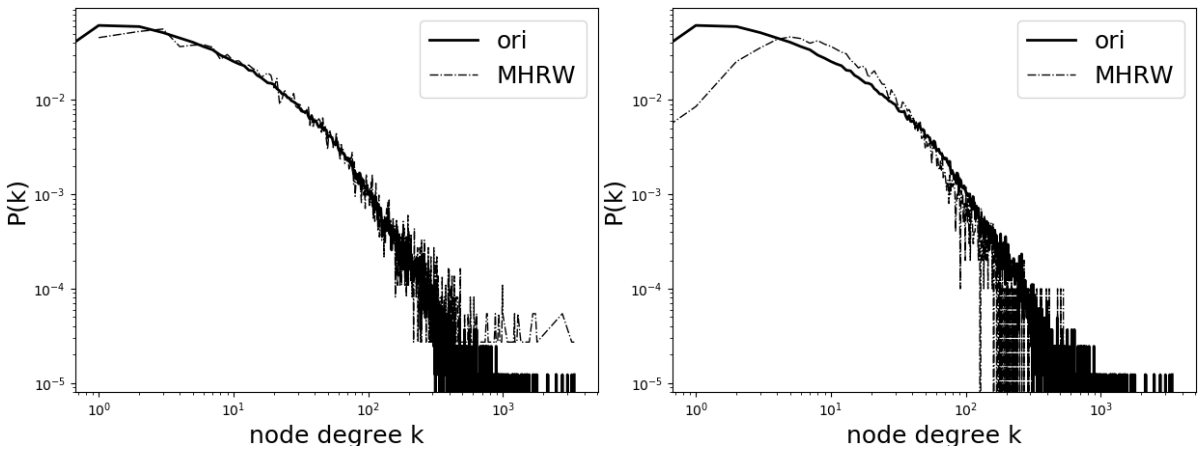


Figure 1. Degree distribution diagram of 10000 nodes in Twitter data set by MHRW algorithm: (a) Degree distribution diagram of sample nodes in the original graph; (b) Degree distribution diagram of induced subgraph constructed by sample
图 1. MHRW 算法在 Twitter 数据集采样一万节点的度分布图：(a) 样本节点在原图中的度分布图；(b) 样本所构造诱导子图的度分布图

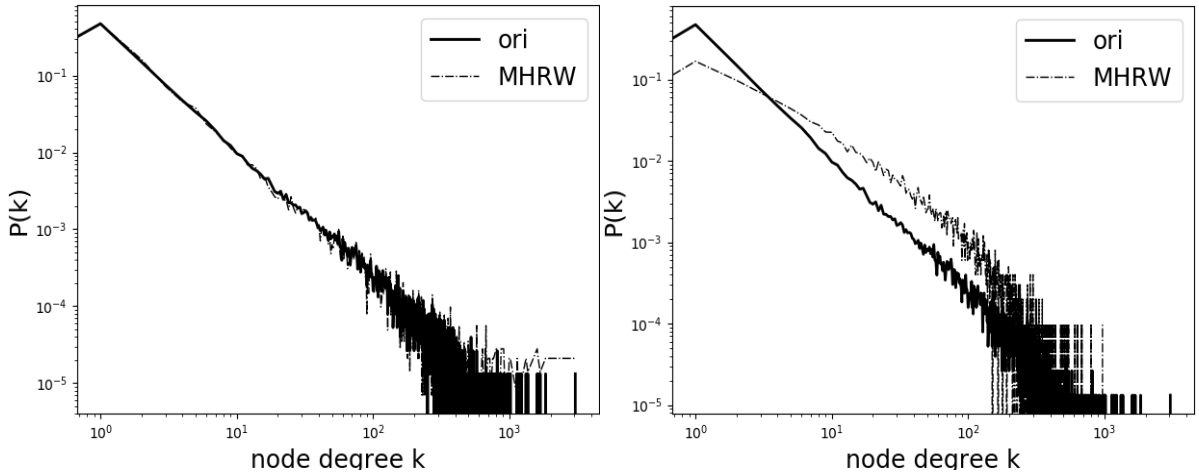


Figure 2. Degree distribution diagram of 10000 nodes in Epinions data set by MHRW algorithm: (a) Degree distribution diagram of sample nodes in the original graph; (b) Degree distribution diagram of induced subgraph constructed by sample
图 2. MHRW 算法在 Epinions 数据集采样一万节点的度分布图：(a) 样本节点在原图中的度分布图；(b) 样本所构造诱导子图的度分布图

对比图 1(a)、图 1(b)与图 2(a)、图 2(b)的度分布图，研究发现，MHRW 的无偏特性仅体现在节点的取样方面，所采集的样本度数在原图中均是无偏，然而在构建诱导子图时，多次自环以及重复游走过的节点只能生效一次，尤其是对于一些连通性相当差、度数过于偏低的图而言，过低的采样效率以及过度自循环引起的偏差估计，采集样本所构建的诱导子图，有时并不能很好地代表原图，如图 3 所示，对于 Epinions 数据集分别按一定比例进行采样，研究其采样效率与自环率。实验对比发现，这类图由于稳定偏高的自环率，导致采样效率持续很低，并且随着样本量逐渐增大，采样效率呈逐渐下降的趋势。

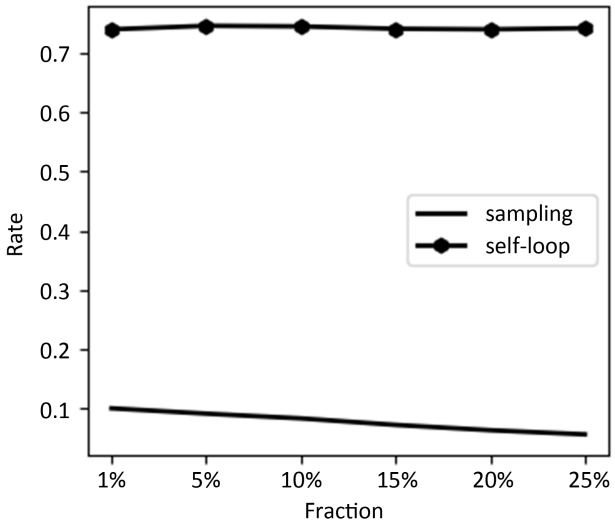


Figure 3. Sampling rate
图 3. 采样率

针对 MHRW 算法的样本构建子图过程中的度数偏差，本论文提出了一种采样高效、修正子图度数偏差的算法(Random Walk Subgraph Sampling, RWSS)。

4.2. 算法改善

根据上述研究结果，分析随机游走的采样偏差，简单随机游走算法中，跳转节点的选择概率与其度数成正比，在其采样之前，本文将其邻接节点的跳转概率序列(P_{v1},P_{v2},⋯,P_{vn})，P_{vi}=1/k_u，均除以其度数为跳转权重进行研究，得到新的序列($P_{v1'}$,$P_{v2'}$,⋯,$P_{vn'}$)，$P_{vi'}$ = 1/$k_u k_v$。其概率转移公式为：

$$p_{uv}=\begin{cases}\dfrac{1}{k_u k_v} & \text{if } v \text{ is a neighbor of } u\\ 0 & \text{otherwise}\end{cases} \tag{5}$$

本文使用 Twitter 以及 Epinions 数据集做实验比较。

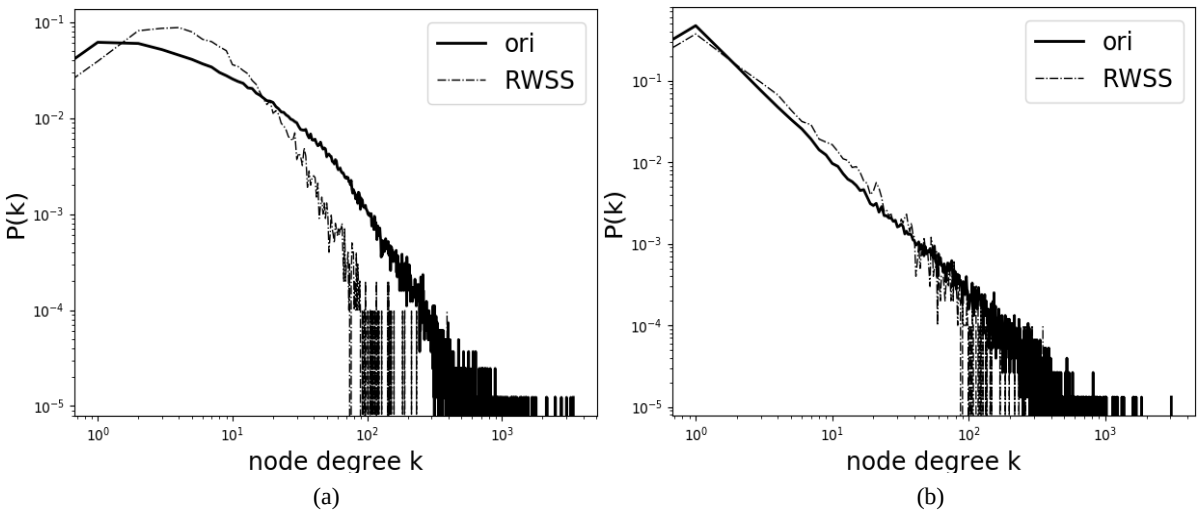


Figure 4. Degree distribution diagram of induced subgraph constructed by sample: (a) Twitter; (b) Epinions
图 4. 样本所构造诱导子图的度分布图：(a) Twitter；(b) Epinions

图 4 为使用公式(5)修改后的跳转概率，采样一万节点所展示的结果图，图 4(a)为采样 Twitter 数据集的样本构造诱导子图的度分布图，图 4(b)为采样 Epinions 数据集的样本构造诱导子图的度分布图。从图中可以看出，修改了跳转概率后的算法，严重倾向于低节点的采样，对于相当稀疏的图有着极佳的采样效果，但并不是适用于多种场景。

本文结合 RW 算法的采样偏差以及 MHRW 算法的无偏特性，提出一种适用于各种社交网络图的采样算法，通过本文采样算法获取的样本诱导子图，具有非常良好的代表性，能更显著体现原图特性。当 $k_v \geq k_u$ 时，其跳转概率与采样算法 MHRW 一致(P_{uv} = 1/k_v)，当 $k_v < k_u$ 时，由于先前的 MHRW 算法在其自循环过程中对很多低度数节点进行了多次重复采样，而这种重复采样在诱导子图的构图过程中是无效的，因为每个节点只能生效一次。由公式可知，MHRW 算法在 $k_v < k_u$ 时，其邻接节点的跳转概率为 P_{uv} = 1/k_u。

根据 MHRW 算法采样无偏特性，本文对于邻接节点度数小于当前节点的跳转概率进行权重参数调整 P_{uv} = α/k_v + (1 − α)/k_u，$\alpha \in [0,1]$。转移概率公式如下：

$$p_{nv}=\begin{cases}\dfrac{1}{k_v}, k_v \geq k_u\\ \alpha\dfrac{1}{k_v}+(1-\alpha)\dfrac{1}{k_u}, k_v < k_u\end{cases} \tag{6}$$

4.3. 参数确定

由公式(6)推算，当 $\alpha = 0$ 时，该算法完全等价于原先的 MHRW 算法，随着 α 值逐渐增大，算法将逐渐偏向于低度数节点的采样。接下来，本文描述如果获取一个最佳 α 值。对于一个完全不了解属性的社交网络图而言，是无法直接根据图形属性结构进行分析的，首先利用 MHRW 算法的无偏特性，对其进行无偏采样获取原图度分布，求出其平均度数：

$$\langle k \rangle = \sum_{i}^{N} k_i p_i \tag{7}$$

这里的 N 记为原图中节点度数类型数，$\langle k \rangle$ 表示平均度数，p_i 表示度数为 k_i 的节点数占网络中节点类型总数的比例。

根据 α 值取值范围，划分区间，进行多次采样测试调整参数，将采样的诱导子图平均度数 $\langle k' \rangle$ 与原图 $\langle k \rangle$ 作比较，以获取最佳 α 值，如图 5 所示，获取最佳参数值。

由于各个社交网络图的属性结构不一样，以及每次采样大小不同，在对多个图进行不同比例抽样时，α 取值也受到各个不同的社交网络图以及同一网络图的采样比例的影响。

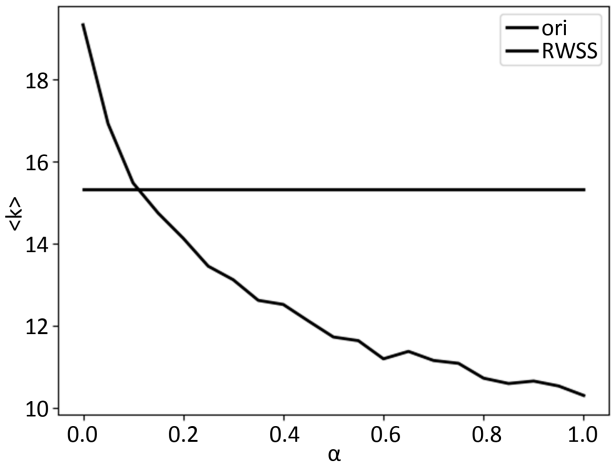


Figure 5. Test α parameter value
图 5. 测试 α 参数值

4.4. 采样流程

RWSS Sampling

1 Input: Graph G = (V, E)

2 get S by MHRW

3 calculate average degree <k> by S

4 get α by <k>

5 Select a seed vertex u from V as initial node

6 Add u to the S’

7 While stopping criterion not met Do

8 Select v by P_{uv} based on Formula (6)

9 Add v to the S’

10 u←v

11 end

12 generate G’ by S’

13 output S’, G’

4.5. 采样偏差

本文使用 MHRW 算法与 RWSS 算法分别在 Twitter 数据集与 Epinions 数据集各采集一万个节点，比较他们的度分布图。

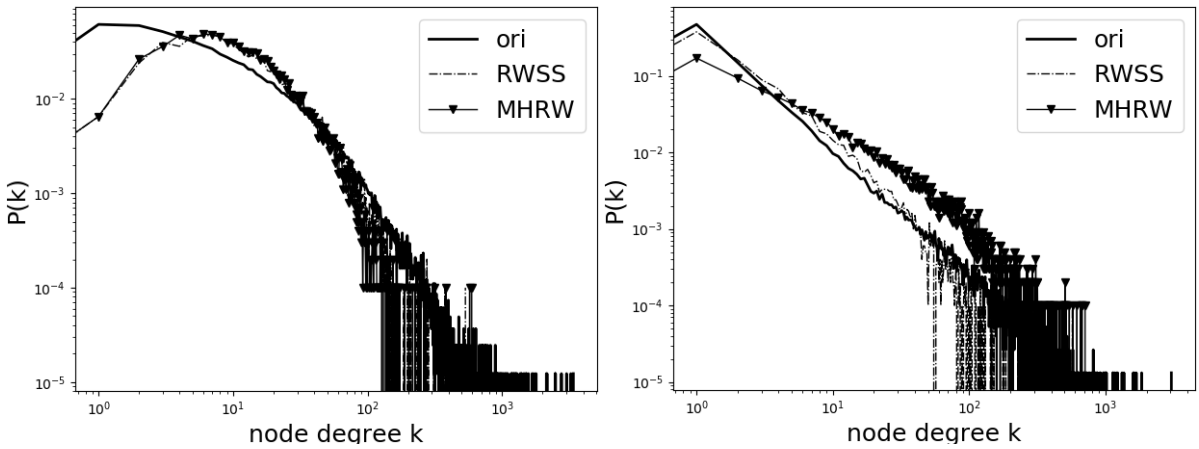


Figure 6. Degree distribution diagram of induced subgraph constructed by sample: (a) Twitter; (b) Epinions
图 6. 样本所构造诱导子图的度分布图：(a) Twitter；(b) Epinions

图 6(a)为采样 Twitter 数据集的样本，所构造诱导子图的度分布图，图 6(b)为采样 Epinions 数据集的样本，所构造诱导子图的度分布图。从图中可以看出，本文提出修正后的采样算法，不仅在较为稀疏的社交网络图上表现优异，对于之前采样表现不佳的社交网络图(如 Twitter)亦与 MHRW 算法相媲美。并且，通过后续研究发现，随着样本数量增多，本文提出的采样算法将会表现更加优异。

5. 评估实验

复杂网络的连通性和拓扑结构的测量对于分析、分类、建模和验证是必不可少的，本论文使用其中一部分指标进行检测采样效果[25]。先后研究了多个不同的数据集，本文分别测试了 RW、MHRW 以及 RWSS 算法，对于每个数据集，通过数十次不同的样本量进行对各种网络图进行采样，并比较测评采样子图结果。本文对社交网络图的度分布误差(*ks* 值校验)、聚类系数(cluster)、同配性(assortativity)、传递性(transitivity)进行了研究对比。其部分数据集如表 1 所示。

Table 1. Data set information
表 1. 数据集信息

数据集	节点数	边数	聚类系数	同配性	传递性
Twitter	81,306	1,342,310	0.5653	−0.039	0.1706
Epinions	75,879	405,740	0.1378	−0.141	0.0399
Tvshow	3892	17,262	0.3737	0.5604	0.5906
Public_Figure	7115	100,762	0.1409	0.2022	0.1666

5.1. 评估属性

本文采用 *ks* 值检测图的度分布，*ks* 是算法所采样子图节点度分布与原图节点累计度分布函数差值。如公式所示：

$$ks = \frac{\max\left|F(x) - F_s(x)\right|}{F(x)} \tag{8}$$

其中，$F(x)$是指原图累计分布函数，$F_s(x)$指采样子图的累计分布函数，*ks* 值越小代表采样子图的度分布与原图度分布越接近，反之，则表示采样子图与原图的度分布差距过大。

聚类系数用来表示一个网络集结成团的系数，对于单个节点来说，即该节点的邻接节点的连接程度。

同配性用来描述网络中，拥有某属性的节点倾向于与其相同属性链接的程度。本文使用度数作为校验结果。

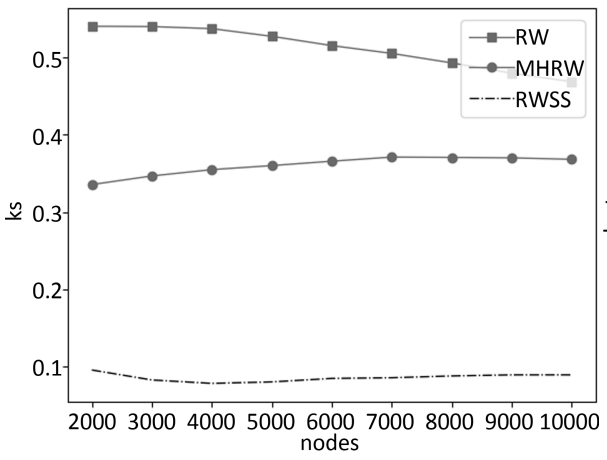
5.2. 实验验证

先前的学者使用 MHRW 对比研究过广度优先搜索算法(BFS)、滚雪球采样算法、森林火灾算法。研究结果[11]证明，由于 BFS 的高度节点偏差，MHRW 的采样结果在大多数情况下均是优于 BFS 以及其修正算法。综上，本文不再重复对比 BFS 以及其类似的修正算法。本文分别使用 RWSS、RW、MHRW 三种采样方法对数十种社交网络图进行了测评验证，研究结果表明，对于大部分情况下，本文所提算法的采样结果均是优于原先的社交网络采样算法。由于本文使用的 RWSS 在 Twitter 数据集上的最佳参数 α 取值为 0，此时采样效果等价于 MHRW 算法。故不再重复展示 Twitter 数据集以及类似取值 $\alpha = 0$ 数据集的研究结果。

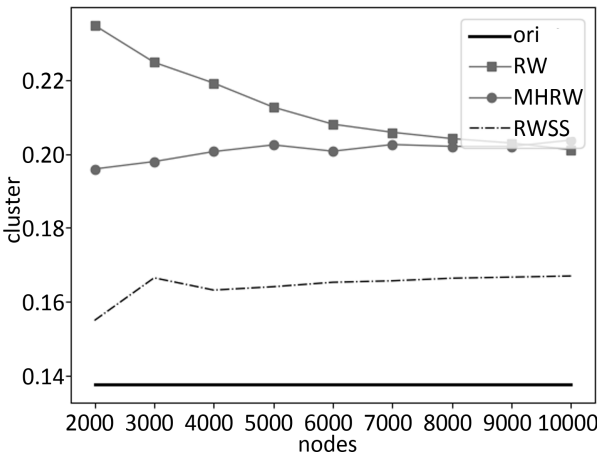
本文将其中效果较为显著的社交网络图(表 1)的数据集进行度分布、聚类系数、同配性以及传递性采样评测。对于每种图，根据该图大小，按照一定比例的节点采样采样 50 次以上，将其样本生成诱导子图，所求属性的平均值与原图进行比较。

测试实验结果如图 7~9 所示，其中，横坐标为采样节点数，图(a)的纵坐标为度分布垂直距离 *ks* 平均值，图(b)的纵坐标为聚类系平均值，图(c)的纵坐标为同配性平均值，图(d)的纵坐标为传递性平均值。ori 为对应的原图数值属性。

图 7 展示了平均度数、聚类系数、同配性、传递性均偏低的 Epinions 数据集所构建的社交网络图采样结果，其中参数 α 取值 0.2，实验显示，RWSS 采样算法在平均度数与聚类系数的表现效果均远远优于



(a)



(b)

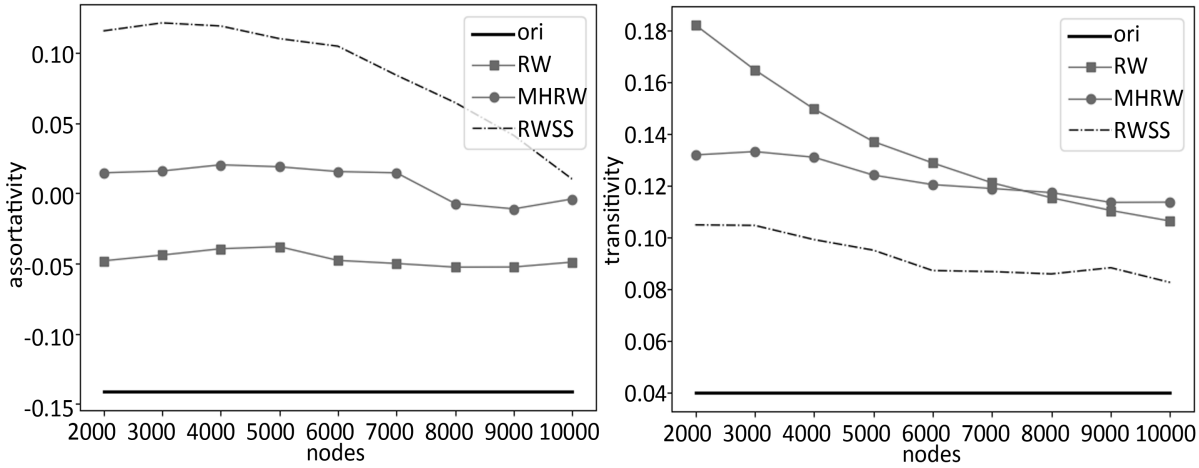


Figure 7. Epinions: (a) Degree distribution vertical distance *ks*; (b) Cluster; (c) Assortativity; (d) Transitivity
图 7. Epinions：(a) 度分布垂直距离 *ks*；(b) 聚类系数；(c) 同配性；(d) 传递性

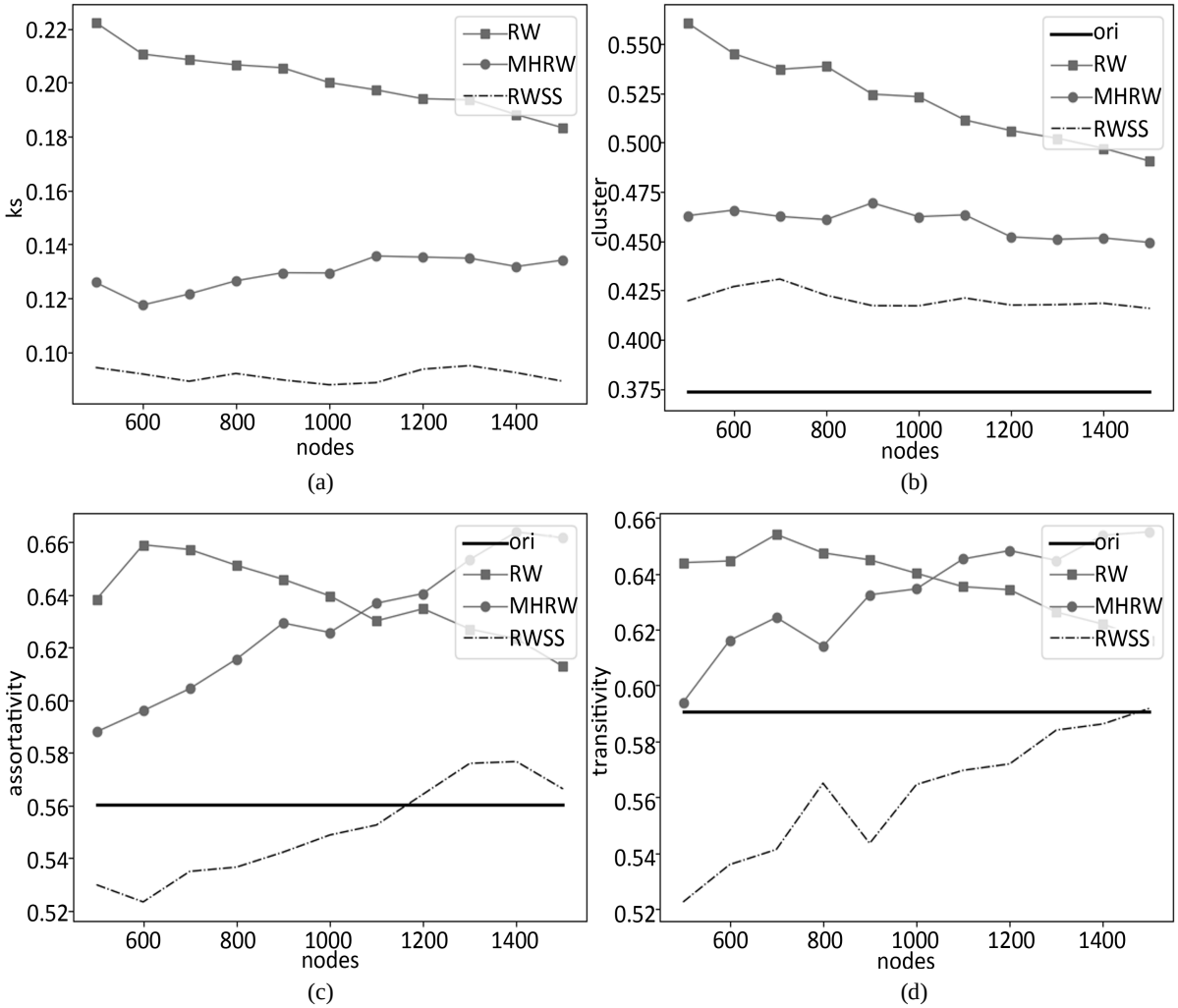


Figure 8. Tvshow: (a) Degree distribution vertical distance *ks*; (b) Cluster; (c) Assortativity; (d) Transitivity
图 8. Tvshow：(a) 度分布垂直距离 *ks*；(b) 聚类系数；(c) 同配性；(d) 传递性

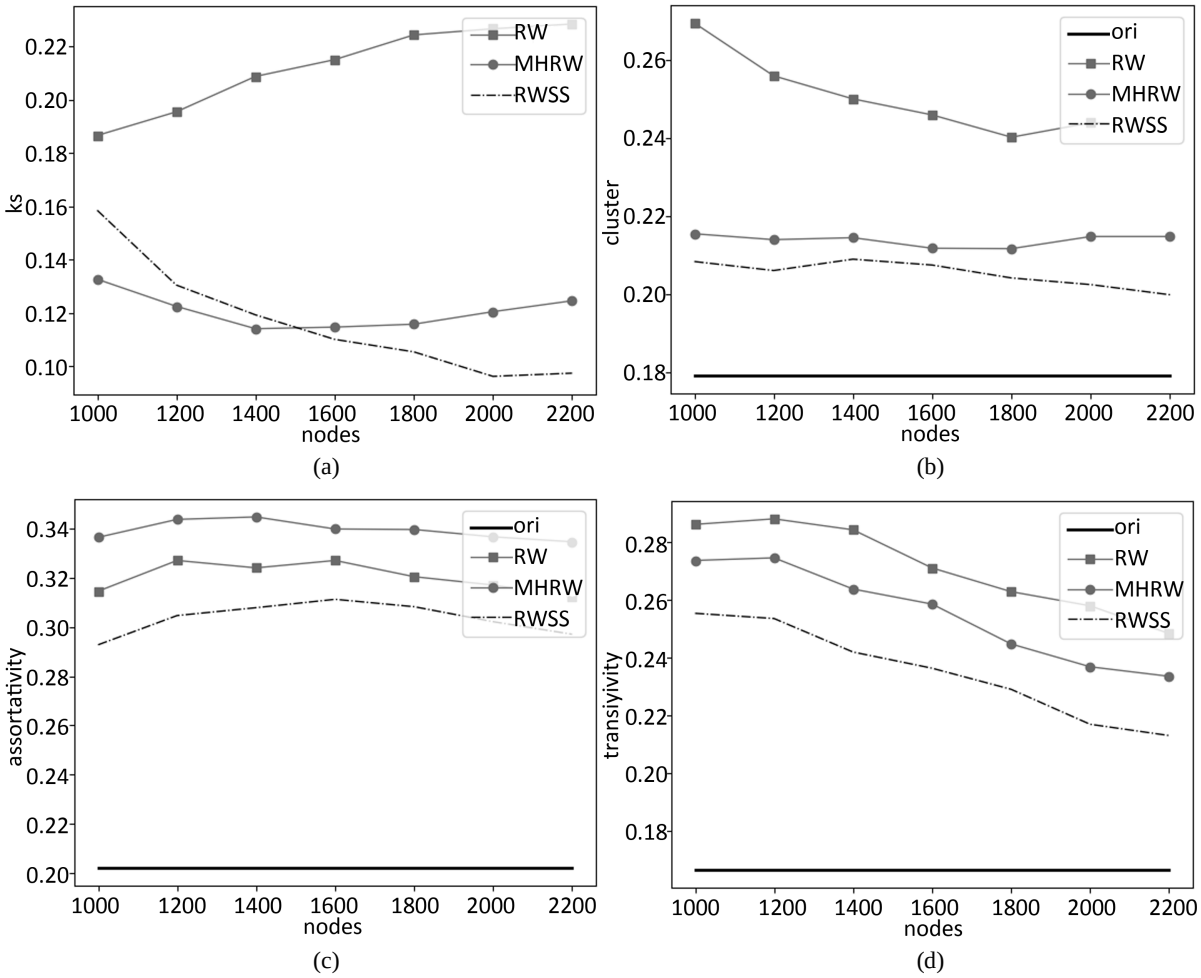


Figure 9. Public_Figure: (a) Degree distribution vertical distance *ks*; (b) Cluster; (c) Assortativity; (d) Transitivity
图 9. Public_Figure：(a) 度分布垂直距离 *ks*；(b) 聚类系数；(c) 同配性；(d) 传递性

传统的采样算法，其次对于传递性的表现亦是较为优异。图 8 研究结果展示了平均度数、聚类系数、同配性、传递性均偏高的 Tvshow 数据集所构建的社交网络图采样结果，其中，参数 α 取值 0.12，RWSS 采样算法在各个方面的表现均优于传统的采样算法，且较为接近原图。最后，图 9 研究结果展示了节点数较少、平均度数与聚类系数较低的 Public_Figure 数据集所构建的社交网络图采样结果，其中，参数 α 取值 0.05，RWSS 算法在度分布、聚类系数、同配性、传递性等方面均优于传统的采样算法，尤其是随着采样节点数量的增多，在度分布方面的误差逐渐减少。综上所述，本文所提的采样算法在大部分情况下，相对于原有算法，均进行了改善优化，其表现效果更加接近原图。

6. 总结

本文在传统的随机游走算法基础上，对采集样本构造诱导子图的偏差问题提出了新的算法。该算法在大部分社交网络图采样所构造的诱导子图，相比较原先的随机游走算法，均进行了优化改善。无论是平均度数、度分布，还是聚类系数、同配性、传递性校验，通过对数十种社交网络图进行研究，每种图按照一定比例进行多次采样，每次样本构造的诱导子图与原图对比度分布、聚类系数、同配性、传递性。结果表明，本文所提算法的综合表现均优于原先游走算法，所生成的诱导子图更加精确接近于原图，并且节点的采样效率相比较 MHRW 算法，也得到大幅度的提高。

最后，通过对大量实验数据研究证明，RWSS 算法在面对不同类型的 OSN 进行采样时，对于各个方面的指标，均有着十分优异的表现。

参考文献

[1] 吴信东, 李毅, 李磊. 在线社交网络影响力分析[J]. 计算机学报, 2014, 37(4): 735-752.

[2] 李嘉兴, 王晰巍, 常颖, 王薇. 社交网络用户行为国内外研究动态及发展趋势[J]. 现代情报, 2020, 40(4): 167-177.

[3] 李立耀, 孙鲁敬, 杨家海. 社交网络研究综述[J]. 计算机科学, 2015, 42(11): 8-21+42.

[4] Viswanath, B., Mislove, A., Cha, M. and Gummadi, K.P. (2009) On the Evolution of User Interaction in Facebook. *Proceedings of the* 2*nd ACM Workshop on Online Social Networks*, Barcelona, 17 August 2009, 37-42. https://doi.org/10.1145/1592665.1592675

[5] 许进, 杨扬, 蒋飞, 金舒原. 社交网络结构特性分析及建模研究进展[J]. 中国科学院院刊, 2015, 30(2): 216-228. https://doi.org/10.16418/j.issn.1000-3045.2015.02.009

[6] 方滨兴, 贾焰, 韩毅. 社交网络分析核心科学问题、研究现状及未来展望[J]. 中国科学院院刊, 2015, 30(2): 187-199. https://doi.org/10.16418/j.issn.1000-3045.2015.02.007

[7] 魏涛. 社交网络结构特性研究[D]: [硕士学位论文]. 北京: 北京邮电大学, 2015.

[8] Ahn, Y. and Moon, S. (2008) Analysis of Topological Characteristics of Huge Online Social Networking Services. *Proceedings of the* 16*th International Conference on World Wide Web*, Banff, 8-12 May 2007, 835-844. https://doi.org/10.1145/1242572.1242685

[9] Ferrara, E. (2012) A Large-Scale Community Structure Analysis in Facebook. *EPJ Data Science*, **1**, Article No. 9. https://doi.org/10.1140/epjds9

[10] Chau, P., Pandit, S., Wang, S. and Faloutsos, C. (2007) Parallel Crawling for Online Social Networks. *Proceedings of the* 16*th International Conference on World Wide Web*, Banff, 8-12 May 2007, 1283-1284. https://doi.org/10.1145/1242572.1242809

[11] Gjoka, M., Kurant, M., Butts, C.T. and Markopoulou, A. (2011) Practical Recommendations on Crawling Online Social Networks. *IEEE Journal on Selected Areas in Communications*, **29**, 1872-1892. https://doi.org/10.1109/JSAC.2011.111011

[12] Wang, T., Yang, C., Zhang, Z., *et al*. (2011) Understanding Graph Sampling Algorithms for Social Network Analysis. 2011 31*st International Conference on Distributed Computing Systems Workshops*, Minneapolis, 20-24 June 2011, 123-128. https://doi.org/10.1109/ICDCSW.2011.34

[13] 尤枫, 曹天亮, 卢罡. 在线社交网络的自适应 UNI 采样方法[J]. 计算机工程, 2017, 43(4): 200-206.

[14] 许南山, 李浩, 卢罡. 在线社交网络的 UNI64 采样方法[J]. 计算机系统应用, 2014, 23(12): 206-212.

[15] Guilotin-Plantard, N. and Schott, R. (2006) Dynamic Random Walks. Theory and Applications. Elsevier, Amsterdam.

[16] Gjoka, M., Kurant, M., Butts, C.T. and Markopoulou, A. (2010) Walking in Facebook: A Case Study of Unbiased Sampling of OSNs. 2010 *Proceedings IEEE INFOCOM*, San Diego, 14-19 March 2010, 1-9. https://doi.org/10.1109/INFCOM.2010.5462078

[17] Salehi, M., Rabiee, H.R. and Rajabi, A. (2012) Sampling from Complex Networks with High Community Structures. *Chaos*, **22**, 2202-2229. https://doi.org/10.1063/1.4712602

[18] 蔡君, 余顺争. 基于随机聚类采样算法的复杂网络社团探测[J]. 计算机应用研究, 2013, 30(12): 3560-3563.

[19] 贺云天. 面向大规模社交网络的采样技术研究及其应用[D]: [硕士学位论文]. 合肥: 中国科学技术大学, 2019.

[20] Dong, W., Li, Z. and Xie, G. (2011) Towards Unbiased Sampling of Online Social Networks. *IEEE International Conference on Communications*, Kyoto, 5-9 June 2011, 1-5.

[21] Jin, L., Chen, Y., Hui, P., *et al*. (2011) Albatross Sampling: Robust and Effective Hybrid Vertex Sampling for Social Graphs. *Proceedings of the* 3*rd ACM International Workshop on MobiArch*, Bethesda, June 2011, 11-16. https://doi.org/10.1145/2000172.2000178

[22] Chen, B., Liu, L., Jia, H. and Zhang, Y. (2017) Reducing Repetition Rate: Unbiased Delay Sampling in Online Social Networks. *Recent Patents on Computer Science*, **10**, 308-314. https://doi.org/10.2174/2213275911666180403110851

[23] Rezvanian, A. and Meybodi, M.R. (2015) Sampling Social Networks Using Shortest Paths. *Physica A*: *Statistical Mechanics & Its Applications*, **424**, 254-268. https://doi.org/10.1016/j.physa.2015.01.030

[24] 崔颖安，李雪，王志晓，等. 在线社交媒体数据抽样方法的比较研究[J]. 计算机学报, 2014, 37(8): 18.

[25] Costa, L. Da F., Rodrigues, F.A., Travieso, G. and Villas Boas, P.R. (2005) Characterization of Complex Networks: A Survey of Measurements. *Advances in Physics*, **56**, 167-242.