基于两种子结构感知的社交网络 Graphlets 采样估计算法

赵倩倩 吕 敏 许胤龙

(中国科学技术大学计算机科学与技术学院高性能计算安徽省重点实验室 合肥 230026)

摘 要 graphlets是指大规模网络中节点数目较少的连通诱导子图,在社交网络和生物信息学领域有着广泛的应用。由于精确计数的计算成本较高,目前大多采用随机游走采样算法来近似估计graphlets的频率。随着节点数目的增多,graphlets的种类数增长迅速且结构变化复杂,快速估计大规模网络中所有种类的graphlets的频率是一项挑战。文中提出了基于两种子结构的随机游走采样算法CSRW2来估计graphlets频率,即给定graphlets节点数k($k=4,5$),通过采样k-graphlets的子结构($k-1$)-path和3-star得到两种样本,之后用比例放大法综合,以高效估计graphlets并适应graphlets结构的复杂变化。实验结果表明,CSRW2能以统一的框架估计所有k-graphlets类型的频率,其估计精度优于现有代表性算法,更适用于频率较低且结构较稠密的graphlets。例如,用CSRW2估计真实网络sofb-Penn94中的5-graphlets,当样本数为2万时,标准均方根误差的平均值由WRW算法的0.8降低至CSRW2算法的0.22左右。

关键词 社交网络,Graphlet,Graphlet频率,随机游走,采样算法,无偏估计

中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.03.046

Estimating Graphlets via Two Common Substructures Aware Sampling in Social Networks

ZHAO Qian-qian LV Min XU Yin-long

(Anhui Key Laboratory on High Performance Computing, School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract Graphlets refer to the connected induced subgraphs with small amount of nodes in large-scale network, and have extensive applications in social networks and bioinformatics. Due to extremely high computational costs of exactly counting graphlets, approximately estimating graphlets concentrations via random walk sampling algorithms already becomes the mainstream approach. As node size k increases, the number of k-graphlets increases rapidly and their structures change dramatically, so it is a challenge to quickly estimate the relative frequency of all types of graphlets (graphlet concentrations) in a large-scale network. Aiming at this problem, this paper proposed a novel sampling algorithm, namely common substructures path and 3-star based graphlets sampling via random walk (CSRW2), to efficiently estimate graphlets concentrations. Given k ($k=4,5$), apart from sampling path via random walk, CSRW2 also samples another substructure 3-star, and then derives graphlets concentrations by proportional amplification to find the dense graphlets with less appearance more efficiently and adapt to the complex structural changes. Experimental evaluations on real networks demonstrate that CSRW2 can estimate k-graphlets in a uniform framework. CSRW2 outperforms the representative methods in terms of accuracy and is more accurate for the k-graphlets with more edges and less appearances in graphs. For example, when 5-graphlets in sofb-Penn94 is estimated, the average *NRMSE* of all 5-graphlets is decreased to 0.22 via CSRW2 in contrast to 0.8 obtained by WRW.

Keywords Social network, Graphlet, Graphlet concentration, Random walk, Sampling algorithm, Unbiased estimation

1 引言

graphlets[1]是指大规模网络中节点数较少的连通诱导子图(见图1),不同类型的graphlets在网络中的频率可以表征网络的局部拓扑结构,在网络分析和生物信息学方面有着广泛的应用,如社交网络分析与比较[2]、网络垃圾邮件及异常检测[3]、生物蛋白质网络分析和疾病基因识别等[4]。例如:一个3节点的graphlet g_2^3描述了网络中节点之间的紧密程度。它的频率称为聚类系数[5],是网络中的一个重要指标,常被用于社交网络挖掘、精准广告投放和朋友兴趣推荐等[6]。

随着节点数的增长,graphlets种类的数量迅速增加,结构愈趋复杂,graphlets频率的计算开销呈指数级增长。精确计算graphlets频率的一种方法是穷举,即枚举给定节点数的所有k-连通子图,并计算与每种graphlet同构的子图数量。

到稿日期:2018-01-31 返修日期:2018-05-15 本文受国家自然科学基金面上项目(61672486)资助。

赵倩倩(1993—),女,硕士生,主要研究方向为社交网络、图计算;**吕 敏**(1977—),女,博士,讲师,CCF会员,主要研究方向为图计算、数据隐私与安全,E-mail:lvmin05@ustc.edu.cn(通信作者);**许胤龙**(1963—),男,博士,教授,博士生导师,CCF会员,主要研究方向为存储系统、图计算。

即使用当下最快的精确计数算法 ESCAPE[7] 来计算图 Flickr（含 244K 个节点）中的 5-graphlets 数也需超过 11 天。有些研究利用分布式系统的并行性或 graphlets 间的组合关系[8]来加速计数，但它们并未大幅降低计算成本。

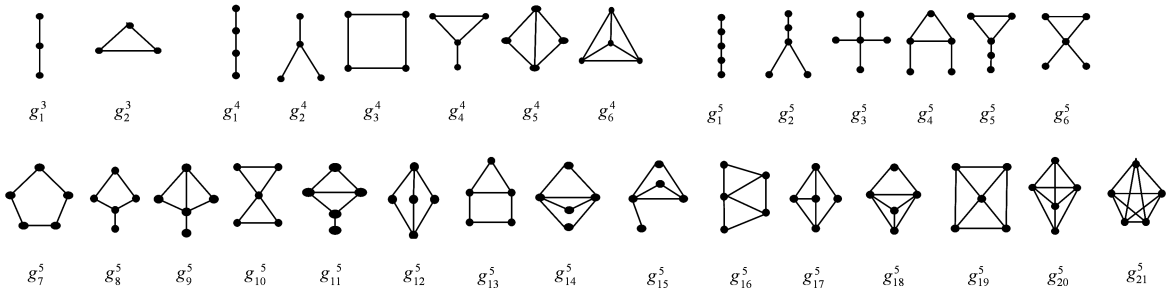


图 1　所有 3，4，5-graphlets 示意图

Fig. 1　All 3，4，5-graphlets diagrams

采样是一种常见的近似估计 graphlets 频率的方法。Wedge sampling[9] 从图中生成随机楔形来估计三角形 graphlet g_2^3 的频率，path sampling[10] 将其扩展到 4-graphlets，但是精确性有待提升。近年来，随机游走采样是估计 graphlets 的主流方法，因为它具有简单适用性，在网络中通过连续 k 步的随机游走来抽样 k-graphlets 并计算其频率。但是简单的 k 步随机游走并不能采样到某些特定类型的 k-graphlets，并且现有采样算法对不同类型 graphlets 的估计精度参差不齐，对于频率较低、结构较稠密的 graphlets 来说，使用随机游走采样估计其频率的精度比估计频率较高的 graphlets 的精度更低。WRW[11] 为无法用 k 步采样到的 k-graphlets 设计了一种特定的摇摆游走方式，但它需要分别为每个 $k(k=4,5)$ 设计不同的摇摆策略，这较为复杂且没有通用做法。

另一类抽样方法考虑在超图上执行随机游走，超图中的节点是原图中的一个连通诱导子图，超图中的边则代表着两端的子图只有一个顶点不同。GUISE[12] 在节点为 3，4，5-graphlets 的超图上导出了一个马尔可夫-蒙特卡洛采样，并基于 Metropolis-Hastings 来获得均匀采样样本，可以同时估计 3，4，5-graphlets 信息。一些工作，如 SRW，PSRW 和 MSS[13]，进一步改进了 GUISE，在一个超级节点为 $(l-1)$-graphlets 的超图上做随机游走来估计 l-graphlets $(l\leqslant k)$。然而，当 $l\geqslant 3$ 时，建立 l-超图并扩展邻居的计算开销很大，使得随机游走难以快速收敛；当 $l\leqslant 2$ 时，在 l-超图上进行随机游走来采样较大的 k-graphlets 效率也不高。

若给定 k，算法目标是计算所有类型的 k-graphlets 的频率。文中提出了一种基于两种子结构采样再扩展的算法 CSRW2（Common Substructures path and 3-star based graphlet sampling via Random Walk）来高效估计大规模网络中的所有 k-graphlets 频率（$k=4,5$）。CSRW2 除了采样子结构 $(k-1)$-path 外，同时采样了另一个子结构 3-star，然后从多个已访问节点的邻居中选择扩展点来生成所有的 k-graphlets，两种方式采样后用比例放大法综合两种样本来高效估计 graphlets，适应 graphlets 结构的复杂变化。CSRW2 用简单统一的方法计算出所有类型的 graphlets 的频率。在真实网络数据集中的实验表明，CSRW2 的估计精度和时间皆优于当前代表性的 graphlets 采样算法，尤其适用于频率较少、更稠密的 graphlets。

本文第 2 节给出了 graphlets 和随机游走的背景知识；第 3 节详细介绍了 CSRW2 算法的细节；第 4 节通过实验验证了 CSRW2 算法的优越性；最后总结全文并对未来进行展望。

2　背景

2.1　图与 graphlets

本文将社交网络抽象为图模型，记为 $G=(V,E)$，其中 V 是此图的节点集，E 是图中的边集，N_v 是节点 v 的邻居。

2.1.1　诱导子图与同构

若有图 $G=(V,E)$ 和 $G'=(V',E')$，$V'\subseteq V$，$E'=\{(u,v)\mid u,v\in V',(u,v)\in E\}$，则 G' 为 G 的诱导子图。对于 V' 中的两个顶点，只要它们在 G 中有边，则它们在 G' 中同样有边。

若两个图 G 和 G' 的节点间存在一个双射 $\varphi:V\to V'$，使得对于 G 中的任意一对节点 $u,v\in V$，均有 $(u,v)\in E$ 当且仅当 $(\varphi(u),\varphi(v))\in E'$，则称图 G 和 G' 同构。

2.1.2　graphlets

graphlets 是指大图中节点数目相对较少的连通诱导子图（见图 1）。各种 graphlets 在图中出现的频率可以表征图的局部拓扑结构。含 k 个节点的 graphlet 记为 k-graphlets。k-graphlets的种类数 T^k 随着节点数目的增加而急剧增加，如 3，4，5，6，7-graphlets 分别有 2，6，21，121，853 种。

给定 k，图中第 i 种 k-graphlet 的数目记为 C_i^k。算法的目标是估算每种 k-graphlet 占所有 k-graphlets 的比例，即频率 c_i^k，其计算式为：

$$c_i^k=\frac{C_i^k}{\sum_{i=1}^{T^k}C_i^k}$$

2.2　随机游走

随机游走是一种适用于社交网络限制访问特性的策略，实际中可通过爬虫技术实现，过程如下：从连通图中任意一个节点出发，并从它的邻居中随机地选取一个节点，然后访问此节点，重复这一过程，其状态转移矩阵 $P=(p(i,j))_{n\times n}$ 如下：

$$p(i,j)=\begin{cases}\dfrac{1}{d(i)}, & \text{如果}(i,j)\in E\\ 0, & \text{其他}\end{cases}$$

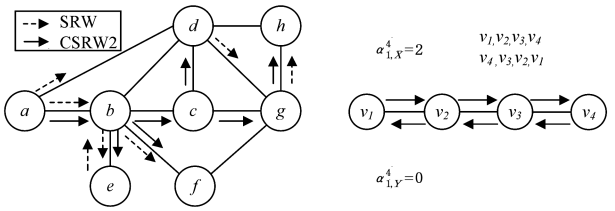
重复这一过程若干步后，节点 v 被访问的概率只与它的度 $d(v)$ 相关，与初始点的选取无关。当随机游走达到静态分布时，某个状态出现的概率为 $P(v)=d(v)/D$。

3 基于两种子结构的 graphlets 估计算法

3.1 引例及算法思想

实际上,给定 k 后不同类型的 graphlets 结构变化繁多,各自在图中的频率差别很大。由 graphlets 频率统计信息(由 ESCAPE[7] 得到)可知,图 sofb-Penn94 中仅 g_1^4 和 g_2^4 的频率之和(0.442+0.379=0.821)就占所有 4-graphlets 的 82%,而图 socfb-B-anon 中 g_1^5 和 g_3^5 的频率之和(0.18+0.405=0.585)占所有 5-graphlets 的 58%。相反,一些 graphlets 的频率极小,如图 com-Amazon 中的 g_6^4,g_{20}^5 和 g_{21}^5 的频率皆为 10^{-6} 左右,这些 graphlets 的结构往往更稠密,大多共享了一个子结构 3-star:g_2^4。先前的研究很少关注各 graphlets 之间的结构特征及频率差异。

传统的简单随机游走算法 SRW(Simple Random Walk)为了收集 k-graphlets 样本,随机选择一个节点 v_1 作为起始点,然后连续随机访问 $k-1$ 个节点得到 $v_2,\cdots,v_k$,其中每个节点皆是上个被访问节点的任意一个邻居节点。若 $v_1,\cdots,v_k$ 中有重复节点,则继续访问下一个节点直至找到 k 个不同的节点或退出。以图 2(a)中的 4-graphlets 采样为例,SRW 以节点 a 为起始点,从其邻居节点 $N_a=\{b,d\}$ 中随机选取一个节点,例如 d,访问 d 节点;再随机选择 d 的邻居节点,例如 g,访问 g 节点;然后随机选择 g 的邻居节点,例如 h,访问 h 节点。通过连续 4 步访问得到随机游走路径 $s^{(4)}=adgh$,由路径 $s^{(4)}$ 上的 4 个节点$\{a,d,g,h\}$诱导产生的 graphlet 是有尾三角形 g_4^4。显然我们可以通过连续 4 步的简单随机游走获得除 g_2^4 外的所有 4-graphlets,这是因为它们中有一条 Hamilton 路径。但对于 g_2^4(如节点集$\{a,b,e,f\}$),SRW 需要 5 步,即沿着路径 $s^{(5)}=abebf$ 或者 $s^{(5)}=abfbe$,需要从 e 到 b 或 f 到 b 回溯来采样。同样地,SRW 取样 g_3^5 需要 7 步而不是 5 步。即使对于相同的顶点数 k,采样不同 graphlets 也需要不同的步数。



(a)SRW 和 CSRW2 (b)CSRW2 的 α_i^k

图 2 SRW 和 CSRW2 以及 CSRW2 对应的重复系数示例图

Fig. 2 Example of SRW, CSRW2 and repeat coefficient of CSRW2

综上所述,各 graphlet 结构和真实频率的差异、抽样过程中是否需要回溯和采样步数的差异,都使得 SRW 对某些 graphlets 的估计准确度不高。因此本文设计了一个新的算法 CSRW2,该算法通过随机游走同时采样 $(k-1)$-path 和 3-star 来适应 graphlets 结构的复杂变化,提高 graphlets 频率估计值的精确度。

CSRW2 的主要思想如下:为了统计图 G 中各 graphlet 的频率,算法采样两种子结构 $(k-1)$-path 和 3-star,再将样本综合。为此把 k-graphlets 按结构分为 2 类:1)含 Hamilton 路径的 graphlets 类型记为集合 X,如当 $k=4$ 时,$X=\{g_1^4,g_3^4,g_4^4,g_5^4,g_6^4\}$;2)含 3-star 的 graphlets 类型记为集合 Y,如当 $k=4$ 时,$Y=\{g_2^4,g_4^4,g_5^4,g_6^4\}$。

采样过程分为以下 3 步。

1)采样 $(k-1)$-path 子结构:按 SRW 运行 k 步得到一个含 k 个节点的随机游走路径 $s_X^{(k)}=v_1v_2\cdots v_k$,直接得到由 $s_X^{(k)}$ 所诱导出的 k-graphlet 样本,若它同构于 g_i^k,则 g_i^k 的频率计数 x_i^k 乘以系数 $\frac{L(s_X^{(k)})}{\alpha_{(i,X)}^k}$(原理见 3.4 节)后累加,即 $x_i^k+=\frac{L(s_X^{(k)})}{\alpha_{i,X}^k}$。

2)采样 3-star 子结构再扩展:从步骤 1)中的邻居中随机选择一个额外节点 v_{k+1} 与 v_1,v_2,v_3 一起得到 3-star,再以 3-star 为基础,从这 4 个节点的邻居中随机选择 $k-4$ 个节点进行扩展,得到 k 个节点的随机游走路径 $s_Y^{(k)}=v_1v_2v_3v_{k+1}\cdots v_{2k-3}$,从而得到 $s_Y^{(k)}$ 诱导出的 k-graphlet 样本,若它同构于 g_i^k,则 g_i^k 的频率计数 y_i^k 乘以系数 $\frac{L(s_Y^{(k)})}{\alpha_{i,Y}^k}$(原理见 3.4 节)后累加,即 $y_i^k+=\frac{L(s_Y^{(k)})}{\alpha_{i,Y}^k}$。需要指出的是,步骤 2)是为了更精确地估计更稠密、出现频率较小的 graphlets。注意步骤 1)—步骤 2)中若有重复节点,则需继续访问下一个节点直至找到 k 个不同的节点或退出。算法 1 给出了 CSRW2 采样 k-graphlets 样本的过程。

3)比例放大综合:在图 G 中多次运行算法 1 可得到许多 $s_X^{(k)}$ 和 $s_Y^{(k)}$ 诱导出的样本,后续需做比例放大来整合 X 和 Y,并消除估计偏差(见 3.2 节、3.3 节)以得到最终的 graphlets 频率估计值 $\hat{c}_i^k$(见算法 2)。

算法 1 CSRW2 采样 k-graphlets 样本

输入:起始点 v_i

输出:k 节点路径 $s_X^{(k)}=v_1v_2\cdots v_k$,$s_Y^{(k)}=v_1v_2v_3v_{k+1}\cdots v_{2k-3}$

1. FOR i=2 TO k DO
2. v_i←从 $N_{v_{i-1}}$ 中随机选取一点
3. ENDFOR
4. v_{k+1}← 从 N_{v_2} 中随机选取一点
5. IF k>4 THEN
6. FOR j=1TO k−4 DO
7. v_{k+1+j}←从 N_{v_1},N_{v_2},N_{v_3},$N_{v_{k+1}}$ 中随机选取一点
8. ENDFOR
9. ENDIF
10. RETURN k 节点路径 $s_X^{(k)}$ 及 $s_Y^{(k)}$

算法 2 CSRW2 估计各 k-graphlet 的频率 $\hat{c}_i^k$

输入:G,样本数 n,graphlets 节点数 k,$\alpha_{i,X}^k$,$\alpha_{i,Y}^k$

输出:各 graphlet 的频率估计值 $\hat{c}_i^k$

1. $\hat{C}_i^k$←0,在随机游走收敛后选择起始点 s
2. FOR l=1 TO n DO
3. v_1←起始点 s,用算法 1 产生 k 路径 $s_X^{(k)}=v_1v_2\cdots v_k$ 及 $s_Y^{(k)}=v_1v_2v_3v_{k+1}\cdots v_{2k-3}$
4. IF $s_X^{(k)}$ 含有 k 个不同的节点 THEN
5. i←$s_X^{(k)}$ 诱导的 graphlet 标号
6. $x_i^k+=\frac{L(s_X^{(k)})}{\alpha_{i,X}^k}$
7. ENDIF
8. IF $s_Y^{(k)}$ 含有 k 个不同的节点 THEN
9. i←$s_Y^{(k)}$ 诱导的 graphlet 标号

10. $y_i^k += \frac{L(s_Y^{(k)})}{\alpha_{i,Y}^k}$

11. ENDIF

12. s←从 N_s 中随机选取一点，l ++

13. ENDFOR

14. FOR i=1 TO T^k DO

15. 用比例放大法（见 3.3 节）处理得到 $\hat{C}_i^k$

16. ENDFOR

17. RETURN $\hat{C}_i^k = \frac{\hat{C}_i^k}{\sum_{i=1}^{T^k} \hat{C}_i^k}$

3.2　重复系数计算

在统计各 graphlet 的频率时，有两个原因可能导致统计误差：1）每种 g_i^k 都可能沿着多个采样路径采样到，每种 g_i^k 的不同采样路径数称为它的重复 α_i^k 系数，不同 g_i^k 的 α_i^k 不同。2）给定一个起始点，每次随机游走访问各节点而得到样本的概率不同，这取决于游走路径的拓扑结构和采样算法。为了消除误差以保证无偏估计，需除去重复系数和每种 graphlet 样本被选取的概率带来的影响。

给定采样算法，重复系数是可以预先计算出的固定常数。如图 2(b)所示，对于 g_1^4，CSRW2 按步骤 1）采样 3-path 时可由 v_1 至 v_4 采样得到 g_1^4，或由 v_4 至 v_1 采样得到 g_1^4，即 $\alpha_{1,X}^4 = 2$；但按步骤 2）采样 3-star 时没有可行的采样路径，这是因为 g_1^4 不含 3-star 结构，即 $\alpha_{1,X}^4 = 0$。表 1 和表 2 列出了 CSRW2 估计 4，5-graphlets 的重复系数 $\alpha_{1,X}^k$ 及 $\alpha_{1,Y}^k$。

表 1　按 CSRW2 采样的重复系数 α_i^k 及 $\hat{c}_i^k$（$k=4$）

Table 1　α_i^k and $\hat{c}_i^k$ of CSRW2 ($k=4$)

Graph-let	g_1^4	 g_2^4	 g_3^4	 g_4^4	 g_5^4	 g_6^4
$\alpha_{i,X}^4$	2	0	8	4	12	24
$\alpha_{i,Y}^4$	0	6	0	6	12	24
$\hat{C}_i^4$	$x_1 \times \frac{(\hat{C}_X^4 + \hat{C}_Y^4)}{\hat{C}_X^4}$	$y_2 \times \frac{(\hat{C}_X^4 + \hat{C}_Y^4)}{\hat{C}_Y^4}$	$x_3 \times \frac{(\hat{C}_X^4 + \hat{C}_Y^4)}{\hat{C}_X^4}$	$x_4 + y_4$	$x_5 + y_5$	$x_6 + y_6$
$\hat{c}_i^4$	$\frac{\hat{C}_1^4}{\sum \hat{C}_i^4}$	$\frac{\hat{C}_2^4}{\sum \hat{C}_i^4}$	$\frac{\hat{C}_3^4}{\sum \hat{C}_i^4}$	$\frac{\hat{C}_4^4}{\sum \hat{C}_i^4}$	$\frac{\hat{C}_5^4}{\sum \hat{C}_i^4}$	$\frac{\hat{C}_6^4}{\sum \hat{C}_i^4}$

表 2　按 CSRW2 采样的重复系数 α_i^k 及 $\hat{c}_i^k$（$k=5$）

Table 2　α_i^k and $\hat{c}_i^k$ of CSRW2 ($k=5$)

Graphlet	g_1^5	 g_2^5	g_3^5	 g_4^5	 g_5^5	g_6^5	g_7^5	g_8^5	g_9^5	 g_{10}^5	g_{11}^5
$\alpha_{i,X}^5$	2	0	0	2	4	0	10	4	4	8	8
$\alpha_{i,Y}^5$	0	6	24	12	6	24	0	6	30	24	18
$\hat{C}_i^5$	$x_1 \times \frac{(\hat{C}_X^5 + \hat{C}_Y^5)}{\hat{C}_X^5}$	$y_2 \times \frac{(\hat{C}_X^5 + \hat{C}_Y^5)}{\hat{C}_Y^5}$	$y_3 \times \frac{(\hat{C}_X^5 + \hat{C}_Y^5)}{\hat{C}_Y^5}$	$x_4 + y_4$	$x_5 + y_5$	$y_6 \times \frac{(\hat{C}_X^5 + \hat{C}_Y^5)}{\hat{C}_Y^5}$	$x_7 \times \frac{(\hat{C}_X^5 + \hat{C}_Y^5)}{\hat{C}_Y^5}$	$x_8 + y_8$	$x_9 + y_9$	$x_{10} + y_{10}$	$x_{11} + y_{11}$
$\hat{c}_i^5$	$\frac{\hat{C}_1^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_2^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_3^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_4^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_5^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_6^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_7^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_8^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_9^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{10}^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{11}^5}{\sum \hat{C}_i^5}$

Graphlet	g_{12}^5	 g_{13}^5	g_{14}^5	 g_{15}^5	 g_{16}^5	g_{17}^5	g_{18}^5	g_{19}^5	 g_{20}^5	 g_{21}^5
$\alpha_{i,X}^5$	12	14	12	12	20	28	36	48	72	120
$\alpha_{i,Y}^5$	12	12	48	42	36	24	60	48	84	120
$\hat{C}_i^5$	$x_{12} + y_{12}$	$x_{13} + y_{13}$	$x_{14} + y_{14}$	$x_{15} + y_{15}$	$x_{16} + y_{16}$	$x_{17} + y_{17}$	$x_{18} + y_{18}$	$x_{19} + y_{19}$	$x_{20} + y_{20}$	$x_{21} + y_{21}$
$\hat{c}_i^5$	$\frac{\hat{C}_{12}^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{13}^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{14}^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{15}^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{16}^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{17}^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{18}^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{19}^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{20}^5}{\sum \hat{C}_i^5}$	$\frac{\hat{C}_{21}^5}{\sum \hat{C}_i^5}$

3.3　比例放大法

对于两种方式都能采样到的 graphlets：$g_i^k \in X \cap Y$（既含 Hamilton 路径也含 3-star），把按步骤 1）采样得到的样本数之和记为 $\hat{C}_X^k = \sum_{g_i^k \in X \cap Y} x_i$，其中 $x_i^k = \frac{D}{n} \sum_1^n G(s_X^{(k)}, k, i) \times \frac{L(s_X^{(k)})}{\alpha_{i,X}^k}$，$x_i^k$ 是对 C_i^k 的无偏估计（证明见 3.4 节），$g_i^k \in X$；把按步骤 2）采样得到的样本数之和记为 $\hat{C}_Y^k = \sum_{g_i^k \in X \cap Y} y_i$，其中 $y_i^k = \frac{D}{n} \sum_1^n G(s_Y^{(k)}, k, i) \times \frac{L(s_Y^{(k)})}{\alpha_{i,Y}^k}$，$y_i^k$ 是对 C_i^k 的无偏估计（证明见 3.4 节），$g_i^k \in Y$。例如，当 $k=4$ 时，$g_4^4, g_5^4, g_6^4 \in X \cap Y$，它们既在 X 中也在 Y 中，则 $\hat{C}_X^4 = x_4 + x_5 + x_6$，$\hat{C}_Y^4 = y_4 + y_5 + y_6$。

由于给定 k，有些 graphlets 只在 X 中，有些只在 Y 中，还有的既在 X 中也在 Y 中，因此本文用两种方式把得到的 graphlets 样本数按比例放大，以求得无偏估计值 $\hat{C}_i^k$。

1）若 g_i^k 可在 X 和 Y 中同时出现，则 $\hat{C}_i^k = x_i + y_i$；

2）若 g_i^k 只在 X 或 Y 中出现，即 y_i 或 x_i 为 0，则 $\hat{C}_i^k =$

$x_i \times \frac{(\hat{C}_X^k + \hat{C}_Y^k)}{\hat{C}_X^k}$ 或 $\hat{C}_i^k = y_i \times \frac{(\hat{C}_X^k + \hat{C}_Y^k)}{\hat{C}_Y^k}$。

表 1 给出了 $k=4,5$ 的比例放大法示例。根据此种方式，$\hat{C}_i^k (1 \leqslant i \leqslant T^k)$ 按同样的比例放大，且 $\hat{C}_1^k : \hat{C}_2^k : \cdots : \hat{C}_{T^k}^k$ 的连比比值也未变化。因此求得的估计值 $\hat{C}_i^k$ 及 $\hat{c}_i^k$ 如表 1 所列。

3.4 无偏估计

当随机游走收敛达到静态分布时，访问节点 v_j 的概率为 $P(v_j) = \frac{d(v_j)}{D} = \frac{d(v_j)}{2|E|}$。考虑按 CSRW2 采样 $(k-1)$-path 及 3-star 两种子结构得到的两条随机游走路径 $s_X^{(k)} = v_1 v_2 \cdots v_k$ 及 $s_Y^{(k)} = v_1 v_2 v_3 v_{k+1} \cdots v_{2k-3}$，其被选取到的概率为：

$$P(s_X^{(k)}) = \frac{d(v_1)}{D} \frac{1}{d(v_1)} \frac{1}{d(v_2)-1} \frac{1}{d(v_3)-1} \cdots \frac{1}{d(v_{k-1})}$$

$$P(s_Y^{(k)}) = \frac{d(v_1)}{D} \frac{1}{d(v_1)} \frac{1}{d(v_2)-1} \frac{1}{d(v_2)-2} \times \left(\frac{1}{\sum_{u=1}^{3} d(v_u) + d(v_{k+1})} \right)^{k-4} \tag{1}$$

记：

$$L(s_X^{(k)}) = (d(v_2)-1) \times (d(v_3)-1) \times \cdots \times d(v_{k-1})$$

$$L(s_Y^{(k)}) = (d(v_2)-1) \times (d(v_2)-2) \times \left(\sum_{u=1}^{3} d(v_u) + d(v_{k+1}) \right)^{k-4} \tag{2}$$

则有：

$$P(s_X^{(k)}) = \frac{1}{D} \frac{1}{L(s_X^{(k)})}, P(s_Y^{(k)}) = \frac{1}{D} \frac{1}{L(s_Y^{(k)})} \tag{3}$$

记 $g(s^{(k)})$ 为 $s^{(k)}$ 诱导的 graphlet，对于每个序列 $s^{(k)}$，定义

$$G(s^{(k)}, k, i) = \begin{cases} 1, & \text{若 } g(s^{(k)}) \text{同构于 } g_i^k \\ 0, & \text{其他} \end{cases}$$

因为每个同构于 g_i^k 的 $g(s^{(k)})$ 在算法 CSRW2 中都有 α_i^k 种方式被采样到，即真实值 C_i^k 乘以了 α_i^k 次，故有：

$$\sum_{s^{(k)} \in S^{(k)}} G(s^{(k)}, k, i) = \alpha_i^k C_i^k$$

定理 1 第一种方式采样 n 个样本得到的 x_i^k 是对 C_i^k 的无偏估计，其中 $g_i^k \in X$；第二种方式采样 n 个样本得到的 y_i^k 是对 C_i^k 的无偏估计，其中 $g_i^k \in Y$。

证明：对于 $x_i^k = \frac{D}{n} \sum_{1}^{n} G(s_X^{(k)}, k, i) \times \frac{L(s_X^{(k)})}{\alpha_{i,X}^k}$ 的期望：

$$\begin{aligned} E(x_i^k) &= \frac{D}{n} \sum_{1}^{n} E\left(G(s_X^{(k)}, k, i) \times \frac{L(s_X^{(k)})}{\alpha_{i,X}^k} \right) \\ &= \frac{D}{n} \sum_{1}^{n} \sum_{s_X^{(k)} \in S_X^{(k)}} P(s_X^{(k)}) G(s_X^{(k)}, k, i) \frac{L(s_X^{(k)})}{\alpha_{i,X}^k} \\ &= \frac{D}{n} \sum_{1}^{n} \sum_{s_X^{(k)} \in S_X^{(k)}} \frac{1}{DL(s_X^{(k)})} G(s_X^{(k)}, k, i) \frac{L(s_X^{(k)})}{\alpha_{i,X}^k} \\ &= \frac{1}{n} \sum_{1}^{n} \sum_{s^{(k)} \in S^{(k)}} G(s_X^{(k)}, k, i) \frac{1}{\alpha_{i,X}^k} \\ &= \frac{1}{n} \sum_{1}^{n} \alpha_{i,X}^k C_i^k \frac{1}{\alpha_{i,X}^k} \\ &= C_i^k \end{aligned}$$

同理，y_i^k 可证。

根据比例放大规则，$\hat{C}_i^k (1 \leqslant i \leqslant T^k)$ 按同样的比例放大，且 $\hat{C}_1^k : \hat{C}_2^k : \cdots : \hat{C}_{T^k}^k$ 的连比比值未变化，故由大数定理得：

$$\hat{c}_i^k = \frac{\sum_{1}^{n} G(s^{(k)}, k, i) \times \frac{L(s^{(k)})}{\alpha_i^k}}{\sum_{i=1}^{T^k} \sum_{1}^{n} G(s^{(k)}, k, i) \times \frac{L(s^{(k)})}{\alpha_i^k}} \to c_i^k, \text{当 } n \to \infty$$

即各 k-graphlet 频率的估计值 $\hat{c}_i^k$ 是无偏的。

3.5 CSRW2 算法框架

CSRW2 的整体过程如下：先在图 G 上用算法 1 采样 n 个 k-graphlets 样本，然后消除误差，用比例放大法综合样本得到各 k-graphlet 频率的无偏估计值 $\hat{c}_i^k$（见算法 2）。

4 实验结果及分析

本节用实验来评估当前代表性算法 SRW2CSS[13]、WRW[11] 和 CSRW2 的估计精确性。SRW2CSS 在超图上随机游走，每个节点对应原图 G 中的边；WRW 通过连续 k 步随机游走并向左或右摇摆得到全部 graphlets 类型。

4.1 实验数据集及评价标准

文中选用斯坦福网络分析项目[14] 中的真实网络数据集（见表 3[7]），用 C++ 编程实现各 graphlets 频率估计算法，并在图的最大连通分支上运行各算法，去除方向、自环和重边。我们使用标准均方根误差 $NRMSE$：

$$NRMSE(\hat{c}_i^k) = \frac{\sqrt{E[(\hat{c}_i^k - c_i^k)^2]}}{c_i^k} = \frac{\sqrt{Var[\hat{c}_i^k] + (c_i^k - E[\hat{c}_i^k])^2}}{c_i^k}$$

来衡量 graphlets 频率估计值 $\hat{c}_i^k$ 的精确度。$NRMSE$ 是误差和方差的组合，能全面反映估计精确性。对于各算法都采样相同的样本数，样本是指采样得到的一个 k 节点子图。

表 3 实验数据集及 graphlets 频率统计信息

Table 3 Graph datasets and graphlets concentrations

Graph	$\|V\|$	$\|E\|$/M	$c_1^4 \times 10^{-1}$	$c_2^4 \times 10^{-1}$	$c_6^4 \times 10^{-2}$	$c_1^5 \times 10^{-1}$	$c_2^5 \times 10^{-1}$	$c_{21}^5 \times 10^{-6}$
Ep10K	10 K	0.4	4.22	5.27	0.105	1.08	3.36	3.76
sofb-Penn94	41 K	1.4	4.42	3.79	0.036	0.53	2.75	7.63
Googleplus	64 K	2.9	3.04	6.52	0.771	1.67	3.97	716
com-Amazon	335 K	0.9	2.10	6.99	0.16	0.30	1.55	7.24
com-Youtube	1.13 M	3.0	0.16	9.80	8.55×10^{-5}	0.00381	0.154	21700
soc-Pokec	1.63 M	22.3	1.45	8.47	0.00353	0.55	0.0639	174
socfb-B-anon	2.93 M	20.9	4.28	5.28	0.0144	1.80	4.05	76.4

4.2 实验结果

4.2.1 精确度分析

我们首先通过实验验证比例放大综合法的有效性。图 3 用 CSRW2 估计了图 com-Youtube 中 g_{10}^5 的频率,这是因为按两种子结构都能采样到 g_{10}^5。可见按$(k-1)$-path 或 3-star 采样的偏差较高,而二者经比例放大后,精确性有了显著提升。

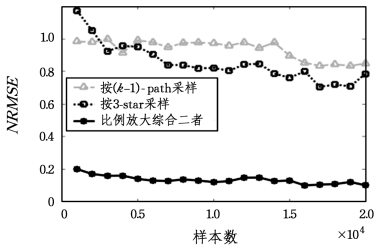
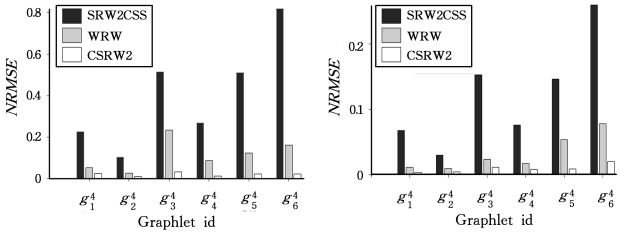


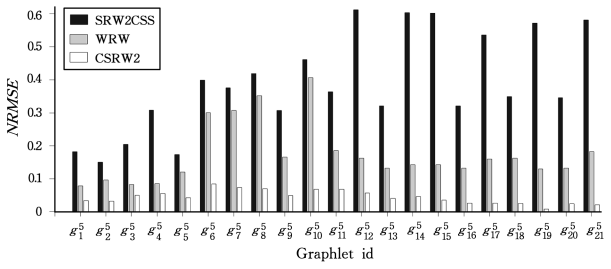
图 3　验证比例放大的有效性(估计 g_{10}^5,com-Youtube)

Fig. 3　Proportional amplification verification (Estimate g_{10}^5, com-Youtube)

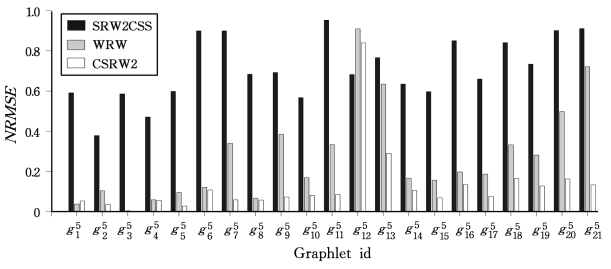
图 4 比较了 CSRW2,WRW 和 SRW2CSS 估计 4,5-graphlets 频率的精确度。



(a)com-Amazon (335 K/926 K)　　(b)socfb-B-anon (3 M/21 M)



(c)soc-Pokec(1.63 M/22 M)



(d)com-Youtube(1.13 M/3 M)

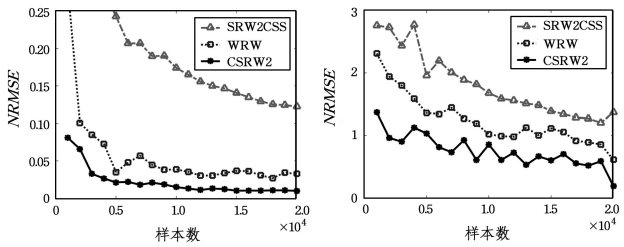
图 4　4,5-graphlets 频率估计值的标准均方根误差(样本数量 20 K)

Fig. 4　*NRMSE* of graphlet concentration estimations for $k=4,5$ (number of sample is 20 K)

样本大小为 2 万,远小于图的节点大小,如只访问图 socfb-B-anon 中 2.67%的节点。*NRMSE* 是 1 000 次独立采样结果的平均值。对于 4-graphlets,CSRW2 与 SRW2CSS 相比,精度至少提高了 3 倍,最多能提高 16 倍(如 com-Amazon 的 g_6^4),与 WRW 相比精度提高了 4 倍(如 com-Amazon 的 g_3^4)。对于 5-graphlets,CSRW2 的精度是 SRW2CSS 的 12 倍(如 soc-Pokec 的 g_{21}^5),是 WRW 的 4 倍(如 com-Youtube 的 g_{21}^5)。

横向比较图 4(c)和图 4(d)可以看出,在 21 种 5-graphlets 中,CSRW2 对 g_9^5 以后的估计精确性更高。因此 CSRW2 明显优于 WRW 和 SRW2CSS,且对很少出现及结构密集的 graphlets 的精确度提升幅度更大。

图 5 给出了不同样本数下所有类型 graphlets 估计值 *NRMSE* 的平均值。随着样本数的增长,各算法性能皆渐趋于稳定,CSRW2 优于 WRW 且大幅优于 SRW2CSS。例如图 5(b)中,在样本数为 20 K 时,采样 5-graphlets 的 *NRMSE* 平均值由 WRW 的 0.8 降低至 CSRW2 的 0.22 左右。进一步考察,当样本数增长至 40 K 时,CSRW2 的 *NRMSE* 平均值继续由 0.22 降低至 0.12 左右。

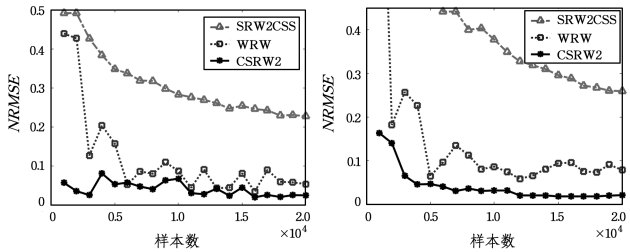


(a)sofb-Penn94 (41.5 K/1.4 M), $k=4$　　(b)sofb-Penn94 (41.5 K/1.4 M), $k=5$

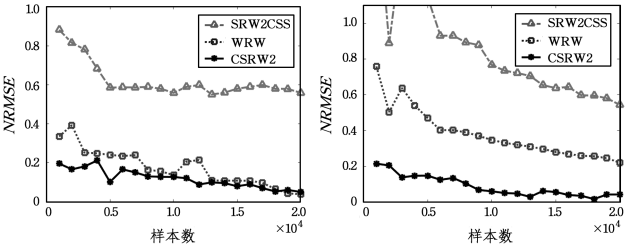
图 5　不同样本数下所有 graphlets 类型的平均 *NRMSE*($k=4,5$)

Fig. 5　Average *NRMSE* over all k-graphlets under different sample number($k=4,5$)

图 6 还比较了在不同样本数下,各算法在不同稀疏图 com-Amazon、com-Youtube,以及稠密图 Ep10K 和 Google-plus 上采样 g_1^4,g_6^4,g_1^5,g_{21}^5 的收敛情况,可知 CSRW2 的估计精度优于 SRW2CSS 和 WRW。g_6^4 的精度提升幅度大于 g_1^4,g_{21}^5 的精度提升幅度大于 g_1^5,g_{21}^5 的精度提升幅度比 g_6^4 稍高,即可得出结论:CSRW2 对含 3-star 结构的 graphlets 更友好。



(a)g_1^4,com-Amazon (335 K/926 K)　　(b)g_6^4,Ep10K (10 K/0.4 M)



(c)g_1^5,com-Youtube (1.13 M/3 M)　　(d)g_{21}^5,Googleplus (64 K/2.9 M)

图 6　4,5-graphlets 频率估计值的收敛性(样本数量 20 K)

Fig. 6　Convergence of 4,5-graphlet concentrations estimates (number of sample is 20 K)

总体而言，SRW2CSS 表现最差，这可能是因为它在超图上采样会导致更大的状态空间和计算代价，而 WRW 和 CSRW2 在原始图上采样。CSRW2 优于 WRW，因为它是基于两种子结构再扩展的方法，更好地适应了 graphlets 种类繁多、频率各异、结构变化复杂的特征。

4.2.2 时间分析

我们采用 3.2GHz Intel Core i5 处理器和 4GB 内存的配置进行实验，来比较不同算法在达到一定精确度时的时间开销。表 4 给出了估计图 com-Amazon 中 c_1^5 的 *NRMSE* 达到 0.1 以下时，估计算法 WRW，CSRW2 及当前最优精确计数算法 ESCAPE 的平均运行时间。由于 SRW2CSS 的时间开销比 WRW 和 CSRW2 高很多，故未把它纳入比较范围。采样算法 WRW 和 CSRW2 都比精确算法 ESCAPE 快两个数量级，这也是精确算法应用受限的原因。与 WRW 相比，采样算法 CSRW2 的计算速度较快。

表 4 不同算法的平均运行时间(估计 c_1^5，$NRMSE<0.1$)

Table 4 Comparison of runtime of different algorithms (Estimate c_1^5, $NRMSE<0.1$)

(单位：s)

Graph	WRW	CSRW2	ESCAPE
com-Youtube	0.5	0.39	1750
soc-Pokec	0.18	0.12	1900
socfb-B-anon	0.23	0.16	2530
com-Amazon	0.36	0.21	20

结束语 为了提高估计大规模社交网络中 graphlets 频率的精确性，尤其是频率更少、结构更稠密的 graphlets，本文提出了一种基于随机游走采样子结构($k-1$)-path 与 3-star 的算法 CSRW2，来有效地估计社交网络中各 graphlets 的频率。实验表明，算法 CSRW2 能高效、快速地估计 4，5-graphlets，其精确性优于现有的代表性算法 SRW2CSS 及 WRW。未来考虑将算法扩展到估计 6，7-graphlets 在大图中的频率并通过实验验证其估计精确性。

参 考 文 献

[1] MILO R，SHEN-ORR S，ITZKOVITZ S，et al. Network motifs：simple building blocks of complex networks[J]. Science，2002，298(5594)：824-827.

[2] UGANDER J，BACKSTROM L，KLEINBERG J. Subgraph frequencies：mapping the empirical and extremal geography of large graph collections[C]// Proceedings of the 22nd international conference on World Wide Web. ACM，2013：1307-1318.

[3] AKOGLU L，TONG H，KOUTRA D. Graph based anomaly detection and description：a survey[J]. Data Mining and Knowledge Discovery，2015，29(3)：626-688.

[4] MILENKOVIĆ T，PRŽULJ N. Uncovering Biological Network Function via Graphlet Degree Signatures[J]. Cancer Informatics，2008，6：257-273.

[5] HARDIMAN S J，KATZIR L. Estimating clustering coefficients and size of social networks via random walk[C]// Proceedings of the 22nd International Conference on World Wide Web. ACM，2013：539-550.

[6] WANG J C，CHANG C H. How online social ties and product-related risks influence purchase intentions：A Facebook experiment[J]. Electronic Commerce Research and Applications，2013，12(5)：337-346.

[7] PINAR A，SESHADHRI C，VISHAL V. Escape：Efficiently counting all 5-vertex subgraphs[C]// Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee，2017：1431-1440.

[8] HOČEVAR T，DEMŠAR J. A combinatorial approach to graphletcounting[J]. Bioinformatics，2014，30(4)：559-565.

[9] SESHADHRI C，PINAR A，KOLDA T G. Triadic measures on graphs：The power of wedge sampling[C]// Proceedings of the 2013 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics，2013：10-18.

[10] JHA M，SESHADHRI C，PINAR A. Path sampling：A fast and provable method for estimating 4-vertex subgraph counts[C]// Proceedings of the 24th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee，2015：495-505.

[11] HAN G，SETHU H. Waddling random walk：Fast and accurate mining of motif statistics in large graphs[C]// Proceedings of the 2016 IEEE 16th International Conference on Data Mining. IEEE，2016：181-190.

[12] BHUIYAN M A，RAHMAN M，AL HASAN M. GUISE：Uniform Sampling of Graphlets for Large Graph Analysis[C]// Proceedings of the 2012 IEEE 12th International Conference on Data Mining. IEEE Computer Society，2012：91-100.

[13] CHEN X，LI Y，WANG P，et al. A general framework for estimating graphlet statistics via random walk[J]. Proceedings of the VLDB Endowment，2016，10(3)：253-264.

[14] LESKOVEC J，KREVL A. SNAP Datasets：Stanford Large Network Dataset Collection[OL]. http://snap.stanford.edu/data.