BiLight：一种用于估计大型图中六顶点二部模体计数的快速方法

¨ ¨Mustafa C. elik、Ozg¨ur Okumus. 和
Lale Ozkahya

摘要——模体（即小型子图）是复杂网络的基本构成单元。分析其频率分布，能够深入了解图的局部结构，并应用于社会网络分析、生物信息学和网络挖掘。然而，随着模体和网络规模的增大，对模体进行精确枚举的计算成本也会急剧上升。对于二部模体而言，这一挑战尤为突出，因为它们广泛用于建模用户-物品、作者-论文以及演员-电影交互等关系。在本文中，我们提出了 **BiLight**（二部轻量模体估计），这是一个可扩展的框架，用于估计全部 17 种非同构六顶点二部模体的频率。具体而言，我们着重估计这些模体的非诱导计数，这在许多图挖掘任务中是常见需求。我们的方法基于对四顶点路径进行随机路径采样，并采用高效的加权策略，统计每条采样路径可扩展为 17 种模体之一的方式。该框架能够对全局模体计数进行无偏估计。我们在各种规模的网络上进行了大量实验。实验结果表明，即使是在无法进行精确计数的超大型网络上，BiLight 也能快速收敛至真实值，并保持较高精度。源代码可在 github.com/celikmustaa/BiLight 获取。

索引词—图元、模体计数、二部图、路径采样、随机化算法、BiLig ht.

arXiv:2611.30249v1 [cs.SI] 23 Nov 2026

1 引言

模体 OTIFS 是网络中的小型子图，统计它们出现的频率是网络分析的核心内容，在生物学、社会科学以及大型在线网络研究中都有应用 [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13]。真实网络中的模体频率是与随机网络进行比较时的一项重要特征 [1],[4], [14]。模体的频率分布通常称为模体度分布或模体谱，能够有力地表征网络的拓扑结构。模体谱也被用作图特征，以评估表示网络的图模型的性能，并对顶点角色进行分类 [1], [5], [6], [7], [8], [9], [15], [16], [17]。理解这些网络的局部结构，对于社区检测、链接预测、异常检测以及用于图分类的图相似性等任务至关重要；这些方法已广泛用于各类图挖掘和学习任务。

Milo 等人 [4] 将网络模体定义为：与随机网络相比，在真实世界网络中出现频率更高的子图。网络模体是转录调控网络研究 [18]，、在线社交网络分析 [17], [19], [20], [21] 以及互联网流量分类和异常检测 [22], [23]中的重要结构。

图元（作为诱导子图的模体）计数可作为图分类和图相似性比较中的重要图特征。两种广泛使用的度量方法是图元核 [24]，，它通过计算归一化图元计数向量的内积来评估一对图；以及相对图元频率距离 [6],，它对经对数缩放的相对频率之差的绝对值求和。这类比较在计-算生物学、化学信息学和软件分析等领域中都十分必要。例如，蛋白质功能预测问题要求判断给定蛋白质是否为酶。蛋白质生物网络的拓扑结构可用于确定该蛋白质的这一特征 [25], [26]。同样，表示化合物的网络可用于预测候选药物的活性或不良反应 [27],[28]。使用图模型计算网络相似性对于社区检测和聚类问题也很有用。例如，Yanardag 和 Vishwanathan [29]计算图相似性，以确定在线社交网络中某个社区的类型。图相似性比较在恶意软件检测中的应用还包括：使用图模型表示软件中的子程序、类及其关系 [30], [31],[32] ，从而通过比较这些图的拓扑结构来识别恶意软件。

模体分析广泛应用于图挖掘问题 [33], [34], [35]。在这些应用中，模体计数的计算过程开销很大，即使对于中等规模的图也是如此—问题的组合爆炸使其无法应用于中等规模的图 [36], [37], [38]。尽管四顶点模体 [39], [40] 和五顶点模体 [41], [42],的计数研究已有进展，但六顶点模体的计数却鲜少受到关注：现有工作只针对单一模式，例如诱导六环 [43], [44], 或一般的非二部图元 [45]。三顶点或四顶点模体的计数（如单部图中的三角形或二部图中的蝴蝶结构）已得到广泛研究，而较大模体（例如包含 6 个顶点的模体）的计数则带来了显著的计算挑战。这类子图的数量会随着图的规模呈组合式增长，使得在大规模网络中进行精确枚举变得不可能。因此，研究者转而采用基于采样的近似方法 [42], [45]。在这项工作中，我们提出 BiLight（二部轻量级模体估计），一种可扩展的

• *M. C.elik* 和 *¨O. Okumus* 任职于土耳其安卡拉哈哲特佩大学计算机工程系，*L. ¨Ozkahya* 任职于该校人工智能与数据工程系。

框架，用于估计图1所示全部 17 种非同构六顶点二部模体的非诱导计数。尽管这些模体是二部的，输入图 G 却不一定是二部图；BiLight 能够在任意一般图中正确计数这些二部子结构。

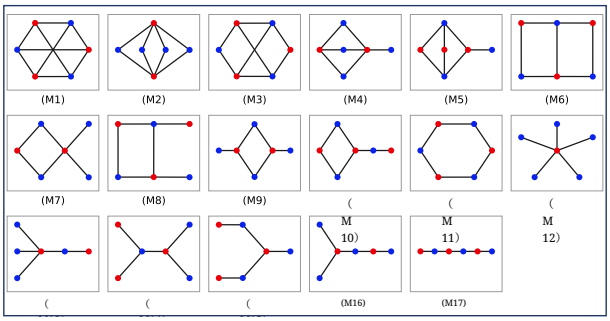

图 1：17 种非同构的六顶点二部模体（M1– M17）。

我们的贡献可概括如下：

- BiLight 的核心贡献是一种方法：均匀随机地采样包含四个顶点的路径，并将其扩展以估计六顶点二部模体的计数。我们的方法具有良好的可扩展性，且计算效率高。

- 为验证我们的方法，我们进行了大量实验，以展示其准确性。我们还提供了理论分析，证明我们的框架能够对六顶点二部模体计数进行无偏估计。为此，我们通过对估计量的方差进行界定，进而界定估计计数的误差。

- 我们在不同规模的真实网络上开展了实验，包括亚马逊评分（图书）、亚马逊评分（电子产品）、IMDB 和 Flickr。实验结果表明，BiLight 能够快速收敛到真实值，并具有较高的准确性（真实值通过高样本量运行进行近似），即使在无法进行精确计数的海量图上，也能实现较低的相对误差和较高的稳定性。我们的实证比较表明，BiLight 的性能优于当前最先进的方法。为便于独立复现实验，我们发布了 BiLight 的源代码，以及用于将局部计数模式与蛮力预言机进行校验的测试套件。

本文其余部分的组织结构如下。第2节介绍必要的符号表示和术语。第3节总结相关工作。第4节介绍我们框架的预处理步骤。第5节介绍我们的 6-顶点二部模体采样方法 BiLight。第6节给出性能评估和测试结果。

2 预备知识

设 $G=(V,E)$ 为一个含有 $n=|V|$个顶点和 $m=|E|$ 条边的无向图。对于顶点 $v\in V$，令 $\Gamma(v)$表示其邻居集合，$d(v)=|\Gamma(v)|$ 表示其度。若一个图 G 的顶点集可以划分为两个二部，不相交集合 U和 W，使得每条边都连接 U中的一个顶点与 W中的一个顶点，则称该图为二部图。模体 M是用作模式的小型连通图。诱导子图 $G'=(V',E')$ 是 G 的一个子图，其中 E' 恰好是 E 中两个端点都位于 V'中的边集。例如，环 C_3包含三个作为子图的 P_3 副本，但它们都不是诱导子图，因为这些路径的端点在 C_3.中由一条边相连。非诱导计数 $N(M)$是 G 中与 M 同构的子图数量（所选顶点间允许存在额外的边），而诱导计数则是诱导子图 $G[S]$ 与 M 同构的顶点子集 S 的数量（不允许存在额外的边）。以诱导方式计数的模体通常称为图元。含 n 个顶点的路径（环）分别记作 P_n （C_n）。表1总结了全文所用的符号表示。

表1：全文使用的符号表示。

符号	说明
$G=(V,E)$	顶点集为 V、边集为 E 的无向图
n,m	顶点数 $\|V\|$ 和边数 $\|E\|$
$\Gamma(v)$	顶点 v的邻居集合
$d(v)$	顶点的度 v，即 $\|\Gamma(v)\|$
Δ	G中的最大度
α	G的树深度：能够划分成的森林的最小数量其 E 可以被划分
P_n	包含 n 个顶点（$n-1$ 条边）的路径
C_n	n环（ n 顶点上的环）
$N(P_4)$	P_4 在 G中的非诱导计数
M_i	第 i个二分6顶点模体（$i=1,\ldots\ldots,17$）
$N(M)$	模体 M 在 G中的非诱导计数
$\widehat{N}(M)$	模体 M 在 G中的估计的非诱导计数
c_M	单个实例内部 P_4 的不同副本数
$A,\ A^{-1}$	模体 M的诱导计数与非诱导计数之间的转换矩阵计数
k	采样预算：采样的数量 P_4
w_e	$e=(u,\ v)$ 加权 P_4 采样中的边权重：$(d(u)-1)(d(v)-1)-\|\Gamma(u)\cap\Gamma(v)\|$
$\mathrm{ext}(M,\ P)$	一个采样的 P_4 扩展为 M 模体的方式数量，在 G中

3 相关工作

本节将回顾网络中模体计数的相关工作。子图和模体计数是一个广泛且研究较为深入的领域；关于全面的综述，请参阅 Ribeiro 等人 [46]的工作。

精确计数。 三角形计数受到的关注最多，既包括精确计数 [47] 也包括近似方法 [34],[48]。Gonen 与 Shavitt [49] 给出了所有四顶点模体非诱导计数（包括逐顶点计数）的近似算法，而 Marcus 与 Shavitt [50] 则精确计数全部四顶点模体。Ahmed 等人 [35], [39] 利用三顶点与四顶点模体计数之间的关系，在四顶点模体计数方面取得了更多进展，并开发了更快的算法。Pınar 等人 [41] 计算了至多五个顶点的所有子图的精确计数。超过五个顶点后，精确计数被证明是困难的：Bera 等人 [51]表明，在有界退化图（涵盖大多数现实世界网络）中，除非标准细粒度复杂性猜想不成立，否则不存在近线性时间算法能够计数包含长度为六或更长的诱导环的模式。这一出现在规模六处的障碍，是估计 而非精确计数六顶点模体的核心动机。

模体计数的近似。已有研究致力于设计用于计算模体浓度的采样方法 [38], [52], [53], [54], [55], [56], [57]。BiLight 属于基于采样的估计器这一类。楔形采样 [34], [58], [59] 抽取随机二路径并检验其是否闭合，以估计三角形计数和聚类系数。Alon 等人 [60]提出彩色编码方法，通过随机为顶点着色并仅枚举多彩模体来降低计算成本。在 [40],中，据报告路径采样在约一百万条边的图上比颜色编码快数百倍，且精度相当。在 [61], [62], [63], [64], 中，采样方法被用于估计静态图和动态图中的三角形数量。Jha 等人 [40] 将其推广为路径采样，选取随机三路径（P_4），以估计所有四顶点子图的计数，并给出可证明的误差界。Wang 等人 [42] 设计了一种采样算法来估计五顶点模体的计数。对于更大的模式，Bressan 等人 [45] 使用颜色编码来估计一般图中 5 至 8 个顶点图元的诱导计数；其工具 Motivo [65], [66], [67]借助简洁数据结构和自适应采样加快了这一过程。BiLight 将路径采样范式再推进一步，对 P_4 子图进行采样并逐一扩展，从而同时估计所有六顶点二部模体。

4 预处理

BiLight 分两个阶段估计每种 6 顶点二部模体的非诱导计数: 预处理 （算法 1）构建别名表，以便均匀 P_4 抽样；估计（算法 2）抽取 k 随机 P_4 子图，并累积局部扩展计数。种子 结构是一个 P_4，它是 17 种模体中 16 种（除星形模体 M_{12}外）的共有结构。

命题 1. 对于任意一条包含 *4* 个顶点的路径，算法 *2* 选取 P 的概率为 $1/N(P_4)$。

证明。设 P是一条由不同顶点构成的路径(v_0, v_1, v_2, v_3)，其中心边 为 $e = \{v_1, v_2\}$，扩展顶点 为 $\{v_0, v_3\}$，并设

$$w_e := (d(v_1) - 1)(d(v_2) - 1) - |\Gamma(v_1) \cap \Gamma(v_2)|,$$

使得 $N(P_4) = \sum_{e \in E} w_e$。事实上，该乘积统计的是有序对 $(v_0,\ v_3) \in (\Gamma(v_1) \setminus \{v_2\}) \times (\Gamma(v_2) \setminus \{v_1\})$，而减去的项剔除了满足$v_0 = v_3$的退化对：这些顶点恰好是 v_1 和 v_2的公共邻居，它们闭合成一个三角形，而非构成一条 P_4。在无三角形宿主图中（尤其是二部图中），这一项为零，因此 $w_e = (d(v_1) - 1)(d(v_2) - 1)$。

要在所有 P_4 中均匀随机地抽样 P，我们执行以下步骤：（*i*）以概率 w_e/N（P_4）选择一条中心边 e；（*ii*）从 $\Gamma(v_1) \setminus \{v_2\}$ 中均匀随机地选择 v_0，并从 $\Gamma(v_2) \setminus \{v_1\}$ 中均匀随机地选择 v_3；以及 （*iii*）如果 $v_0 = v_3$，则拒绝该顶点对，并重复步骤 （*ii*）。步骤 （*ii*）以相等概率抽取每一个 （$d(v_1) - 1$） （$d(v_2) - 1$） 有序对，因此在给定接受条件下，w_e 个有效顶点对中的每一个都具有相同的可能性，即概率为 $1/w_e$。这些步骤得到

$$\Pr(P) = \frac{w_e}{N(P_4)} \cdot \frac{1}{w_e} = \frac{1}{N(P_4)}.$$

算法 1： BiLight：预处理

输入： 图 $G = (V,\ E)$ 及其边 $E = \{e_1, \ldots, e_m\}$
输出： 数组 概率[i]、别名[i]，对应 $i = 1$、.....、 m; 总 P_4 计数 $N(P_4)$
对于每条边 $e_i = (u,\ v) \in E$，**执行**
  $w_{e_i} \leftarrow (d(u) - 1)(d(v) - 1) - |\Gamma(u) \cap \Gamma(v)|$
  // 以 e_i为中心的数量
$N(P_4) \leftarrow \sum_{i=1}^{m} w_{e_i}$
对于 $i \leftarrow 1$ **到** m，**执行**
  $q_i \leftarrow m\, w_{e_i}/N(P_4)$ // 按 m 缩放，因此均值为1
小 $\leftarrow \{i : q_i < 1\}$;　　大 $\leftarrow \{i :\ q_i \geq 1\}$
当 小 **和** 大 **都非空时**
  从小 中移除一个索引 s，并从中移除一个索引 ℓ 从 大
  概率[s] $\leftarrow q_s$;　别名 [s] $\leftarrow \ell$
  $q_\ell \leftarrow q_\ell - (1 - q_s)$　　// 桶 ℓ 涵盖桶 s的不足
  如果 $q_\ell < 1$**则**
    移动 ℓ至 小
  else
    移动 ℓ至 大
对索引 i剩余的 小 或 大 **执行**
  概率[i] $\leftarrow 1$
返回概率[$\cdot$]，别名[$\cdot$]，　$N(P_4)$

拒绝整个顶点对（而不是重新抽取单个端点）正是保持所接受路径均匀性的关键。该步骤的开销也很低。记 $a = d(v_1) - 1$、$b = d(v_2) - 1$ 和 $c = |\Gamma(v_1) \cap \Gamma(v_2)|$，则 $w_e = ab - c$，接受概率为 $1 - c/(ab)$。如果 $\max(a, b) \geq 2$，则 $c \leq \min(a, b)$ 给出 $c/(ab) \leq 1/\max(a, b) \leq 1/2$；而如果 $a = b = 1$，则 $w_e > 0$ 迫使 $c = 0$，因此抽样总会被接受。因此，每条满足 $w_e > 0$ 的边都至少以 $1/2$ 的概率被接受，抽样次数的期望最多为两次。实际上，该步骤的开销远低于这一上界：在表3的十二个非二部数据集中，拒绝最多只在 3% 的抽样中触发（其中 com-amazon 的三角形密度最高），而在二分图宿主中从不触发，因为对于每条边，$\Gamma(v_1) \cap \Gamma(v_2) = \emptyset$。

□

接下来需要高效实现步骤 *(i)* 中的加权边抽样。教科书式的做法是存储 $w_e/N(P_4)$ 的前缀和，并对每次抽样的均匀随机值进行二分查找；对于含有 m 条边的图，每次抽样的开销为 $O(\log m)$。由于 BiLight 最多要抽取数百万条路径，我们改用 ***Walker*** 别名法 [68]，如算法 1 所述，将每次抽样的开销降至 $O(1)$。设 $p : E \to [0, 1]$ 为一个概率分布。若要从 $E = \{e_1, \ldots, e_m\}$ 中抽取一条边，我们均匀随机地选取 $i \in \{1, \ldots, m\}$ 和一个值 $r \in [0, 1)$。如果 $r < prob[i]$，则返回边 e_i；否则返回它的别名边$e_{alias[i]}$。因此，只需进行一次索引抽样并比较一次，就能以预期概率 $p(e)$返回一条边。

模体 ext(M, P)	
M1	$\mathbb{I}_{c_4} \cdot \sum_{u \in I_L} \lvert\Gamma(u) \cap I_R\rvert$
M2	$\mathbb{I}_{c_4} \cdot \left(\binom{\lvert I_L\rvert}{2} + \binom{\lvert I_R\rvert}{2}\right)$
M3	$E(I_L, I_R) + \mathbb{I}_{c_4} \cdot \left(\sum_{u \in I_L} \left(\lvert\Gamma(u) \cap \Gamma'(v_0)\rvert + \lvert\Gamma(u) \cap \Gamma'(v_2)\rvert\right) + \sum_{u \in I_R} \left(\lvert\Gamma(u) \cap \Gamma'(v_1)\rvert + \lvert\Gamma(u) \cap \Gamma'(v_3)\rvert\right) + \lvert I_L\rvert\lvert I_R\rvert - \lvert I_L \cap I_R\rvert\right)$
M4	$\binom{\lvert I_L\rvert}{2} + \binom{\lvert I_R\rvert}{2} + \mathbb{I}_{c_4} \cdot (\lvert I_L\rvert(d'(v_1) + d'(v_3) - 2) + \lvert I_R\rvert(d'(v_0) + d'(v_2) - 2))$
M5	$\sum_{u \in I_L} \lvert\Gamma'(u) \cap \Gamma'(v_2)\rvert + \sum_{u \in I_R} \lvert\Gamma'(u) \cap \Gamma'(v_1)\rvert + \mathbb{I}_{c_4} \cdot \left(\sum_{u \in I_L} d'(u) + \sum_{u \in I_R} d'(u) + \sum_{u \in I_L} (d'(v_0) + d'(v_2) - \mathbb{I}(u \in \Gamma(v_0)) - \mathbb{I}(u \in \Gamma(v_2))) + \sum_{u \in I_R} (d'(v_1) + d'(v_3) - \mathbb{I}(u \in \Gamma(v_1)) - \mathbb{I}(u \in \Gamma(v_3)))\right)$
M6	$E(I_R, \Gamma'(v_3)) + E(I_L, \Gamma'(v_0)) + \lvert I_L\rvert\lvert I_R\rvert - \lvert I_L \cap I_R\rvert + \mathbb{I}_{c_4} \cdot (E(\Gamma'(v_1)\ \Gamma'(v_0)) + E(\Gamma'(v_1)\ \Gamma'(v_2)) + E(\Gamma'(v_3)\ \Gamma'(v_0)) + E(\Gamma'(v_3)\ \Gamma'(v_2)))$
M7	$\lvert I_L\rvert(d'(v_1) - 1) + \lvert I_R\rvert(d'(v_2) - 1) + \mathbb{I}_{c_4} \cdot \sum_{i=0}^{3} \binom{d'(v_i)}{2}$
M8	$\sum_{u \in \Gamma'(v_1)} \lvert\Gamma\ u\ \cap \Gamma'\ v_2\ \rvert + \sum_{u \in I_L} d'\ u\ + \sum_{u \in I_R} d'\ u\ + \sum_{u \in I_L} (d'\ v_2\ - \mathbb{I}\ u \in \Gamma\ v_2\ ) + \sum_{u \in I_R} (d'\ v_1\ - \mathbb{I}\ u \in \Gamma\ v_1\ ))$
M9	$+ \mathbb{I}_{c_4} \cdot (d'\ v_0\ d'\ v_1\ + d'\ v_1\ d'\ v_2\ + d'\ v_2\ d'\ v_3\ + d'\ v_3\ d'\ v_0\ - \lvert\Gamma'\ v_0\ \cap \Gamma'\ v_1\ \rvert - \lvert\Gamma'\ v_1\ \cap \Gamma'\ v_2\ \rvert - \lvert\Gamma'\ v_2\ \cap \Gamma'\ v_3\ \rvert - \lvert\Gamma'(v_3) \cap \Gamma'(v_0)\rvert)$
M10	$\lvert I_L\rvert(d'(v_3) - 1) + \lvert I_R\rvert(d'(v_0) - 1) + \mathbb{I}_{c_4} \cdot (d'(v_1)d'(v_3) - \lvert I_L\rvert + d'(v_0)d'(v_2) - \lvert I_R\rvert)$
M11	$\sum_{u \in I_L} (d'(v_0) - \mathbb{I}(u \in \Gamma(v_0))) + \sum_{u \in I_R} (d'(v_3) - \mathbb{I}(u \in \Gamma(v_3))) + E(\Gamma'(v_0)\ \Gamma'(v_1)) + E(\Gamma'(v_3)\ \Gamma'(v_2)) + \mathbb{I}_{c_4} \cdot \sum_{i=0}^{3} \sum_{u \in \Gamma'(v_i)} d'(u)$
M12	$\sum_{u \in \Gamma'(v_0)} \lvert\Gamma(u) \cap \Gamma'(v_3)\rvert$
M13	$\binom{d'(v_1)}{2} + \binom{d'(v_2)}{2}$
M14	$d'(v_1)d'(v_2) - \lvert\Gamma'(v_1) \cap \Gamma'(v_2)\rvert$
M15	$\sum_{u \in \Gamma'(v_1)} d'(u) + \sum_{u \in \Gamma'(v_2)} d'(u) + (d'(v_1)d'(v_0) - \lvert\Gamma'(v_1) \cap \Gamma'(v_0)\rvert) + (d'(v_2)d'(v_3) - \lvert\Gamma'(v_2) \cap \Gamma'(v_3)\rvert)$
M16	$\binom{d'(v_0)}{2} + \binom{d'(v_3)}{2} + d'(v_1)d'(v_3) - \lvert I_L\rvert + d'(v_0)d'(v_2) - \lvert I_R\rvert$
M17	$(d'(v_0)d'(v_3) - \lvert\Gamma'(v_0) \cap \Gamma'(v_3)\rvert) + \sum_{u \in \Gamma'(v_0)} d'(u) + \sum_{u \in \Gamma'(v_3)} d'(u)$

表 2：一般 宿主（二部或非二分）中的局部计数模式：对于每个模体，ext(M, P)的闭式表达式。

5 采样算法

在算法 2 中，ext(M, P) 表示一个采样得到的 P_4 可扩展为模体 M在 G中的一个出现的方式数，而 c_M 表示单个 M 实例内部 P_4 的不同副本数。表 2 中的许多表达式包含两个加性分量：始终计算的路径项，以及乘以$\mathbb{I}_{c_4}$的环项；仅当种子路径闭合成一个 C_4时，后者才有贡献。模体 M_{12}是六顶点星形图 $K_{1,\,5}$；其计数可直接根据度序列计算为 $\sum_v \binom{d(v)}{5}$，因此不纳入采样。

我们按照路径顺序，将算法 2 中采样得到的 P_4 的顶点重新标记为 $P = (v_0, v_1, v_2, v_3)$，其中 v_0 和 v_3 是端点，并为每个模体 M计算 ext(M, P)。每个P_4 都可能通过从路径顶点的邻域中添加另外两个顶点，扩展为一个六顶点模体。我们为每个模体推导了一个闭式表达式；表 2 列出了所有表达式。这些表达式使用邻域交集、顶点集之间的边数以及顶点度。

我们在表 2 中使用以下辅助定义。对于顶点 u，令 $\Gamma'(u) = \Gamma(u) \setminus \{v_0, v_1, v_2, v_3\}$ 表示位于路径之外、与 u 相邻的顶点，$d'(u) = |\Gamma'(u)|$。我们定义两个过滤后的交集：

$$I_L = \Gamma(v_1) \cap \Gamma(v_3) \setminus \{v_0, v_2\},$$
$$I_R = \Gamma(v_0) \cap \Gamma(v_2) \setminus \{v_1, v_3\}.$$

当 v_0 与 v_3相邻时，指示函数 $\mathbb{I}_{c_4}$ 等于1，否则为0。类似地，$\mathbb{I}(Q)$表示命题 Q. 的指示变量。我们还使用符号表示 $E(U, W) := |\{(u, w) \in E\colon u \in U, w \in W\}|$。我们通过遍历较小的集合，并查询以哈希集合形式存储的较大集合的邻接关系，从而高效地完成计算。在局部计数过程中，所有交集和边数计算都采用相同的非对称性感知策略：比较两个大小不同的集合时，我们总是遍历较小的集合。

尽管我们的模体是二部的，但采样器和计数模式都不假设宿主图是二分图。算法 1 的边权重 w_e 以及命题 1 证明中的拒绝步骤，已经扣除了每条中心边所经过的三角形，因此无论输入图为何种图，抽样得到的都是均匀随机的 P_4。表 2 中的模式同样适用于一般情形：邻接指示函数 $\mathbb{I}(u \in \Gamma(v_i))$以及诸如 $-|I_L \cap I_R|$ 和 $-|\Gamma'(v_i) \cap \Gamma'(v_j)|$ 这样的交集修正项，排除了二分图宿主在构造上就不允许出现的同侧邻接和三角形。在二分图宿主中，这些项全都为零，每个表达式都会化简为更简单的二部图形式。

算法 2： BiLight：估计

输入： G；数组 *prob*[i]，别名[i] 对于 $i = 1, \ldots, m$ 以及 $N(P_4)$ 来自算法 1；采样预算 k

输出： 估计计数 $\widehat{N}(M_1) \ldots, \widehat{N}(M_{17})$

$s_j \leftarrow 0$ 对于 $j \neq 12$；$\widehat{N}(M_{12}) \leftarrow \sum_v \binom{d(v)}{5}$

对于 $t \leftarrow 1$ **到** k，**执行**

- 采样边 $e = (u, v)$，概率为 $w_e/N(P_4)$
- **重复**
  - 均匀随机地选取 $u' \in \Gamma(u) \setminus \{v\}$
  - 均匀随机地选取 $v' \in \Gamma(v) \setminus \{u\}$
- **直到** $u' \neq v'$
- // $u' = v'$ 闭合了一个三角形，而不是 P_4
- $P \leftarrow (u', u, v, v')$
- **遍历**模体 $M_j, j \neq 12$**执行**
  - $s_j \leftarrow s_j +$ ext(M_j, P) / c_{M_j}

// 表 2 给出了 ext(M_j, P)

对于 $j \leftarrow 1$**至** 17, $j \neq 12$**执行**

- $\widehat{N}(M_j) \leftarrow \frac{N(P_4)}{k} \cdot s_j$

返回 $\widehat{N}(M_1)$ ....., $\widehat{N}(M_{17})$

数据集	顶点	边	P_4 计数	运行时间（k=5K）	来源
二部					
亚马逊（电子产品）	4.5M	7.7M	25.8B	12 s	[69], [70], [71]
亚马逊（图书）	10.0M	22.3M	352.0B	72 s	[69], [70], [71]
IMDB	1.2M	3.8M	4.6B	2 s	[72]
Flickr	470.3K	8.5M	2.6T	232 秒	[73]
额外的非二分					
词语关联	10.6K	63.8K	79.2M	1 s	[74], [75], [76]
Facebook 墙	44.0K	182.4K	126.3M	1 s	[77], [78]
Facebook 链接	63.4K	816.9K	7.5B	9 s	[77], [78]
com-amazon.ungraph	334.9K	925.9K	81.0M	<1 s	[79]
com-dblp.ungraph	317.1K	1.0M	675.6M	4 s	[79]
web-BerkStan	654.8K	6.6M	1.7T	340 秒	[79]
as-skitter	1.7M	11.1M	1.4T	350 秒	[79]
cit-Patents	3.8M	16.5M	9.1B	3 s	[79]
LiveJournal	5.4M	49.5M	2.6T	106 秒	[74], [75], [76]
好莱坞	1.1M	56.3M	60.2T	3,326 秒	[74], [75], [76]
Soc-Orkut	3.0M	106.3M	16.3T	298 秒	[80]
Orkut-Links	3.1M	117.2M	22.2T	564 秒	[81], [82]

表 3：实验中使用的数据集。顶点数和边数在适用时指最大连通分量（LCC）中的数量。P_4 计数是图（采样空间）中的 P_4 子图数量。运行时间（k=5K）是 BiLight 在采样预算为 $k = 5$K 时的估计时间，单位为秒，不包括一次性的图初始化；二部图数值为多次独立运行的平均值。Flickr 文件是一个混合了用户–用户边和用户–群组边的图；我们只保留用户–群组边，因此该图成为二分图。

引理1。对于任意具有 $i \neq 12$ 的模体 $M = M_i$，该估计量是^无偏的，$\mathbb{E}[N(M)] = N(M)$，并且

$$\mathbb{V}ar[\widehat{N}(M)] \leq \frac{1}{k}\frac{N(P_4)}{c_M}\big(N(M)+p\big),$$

其中，p 表示共享一个 P_4 种子的 M实例的有序对数量。因此，通过取 $O(\log(1/\delta))$ 个估计量的中位数，我们的估计量可得到一个 (ϵ,δ)-估计量。

证明。令 $\mathcal{P}$ 表示 G 中所有 P_4 子图的集合，因此 $|\mathcal{P}| = N(P_4)$。模体 M 的非诱导计数估计量，将局部计数的平均值按路径总数$N(P_4)$ 进行缩放：

$$\widehat{N}(M) = \frac{N(P_4)}{k}\sum_{t=1}^{k}\frac{\mathrm{ext}(M,P_t)}{c_M}$$

根据命题 1：

$$\begin{aligned}\mathbb{E}[\widehat{N}(M)] &= \frac{N(P_4)}{k}\sum_{t=1}^{k}\mathbb{E}\Big[\frac{\mathrm{ext}(M,P_t)}{c_M}\Big]\\ &= \frac{N(P_4)}{c_M}\sum_{P\in\mathcal{P}}\Pr(P)\cdot\mathrm{ext}(M,P)\\ &= \frac{1}{c_M}\sum_{P\in\mathcal{P}}\mathrm{ext}(M,P) = N(M).\end{aligned}$$

最后一个等式源于以下事实：G 中模体 M 的每个实例都恰好包含 c_M 个 P_4的不同副本，因此得到 $\sum_{P\in\mathcal{P}}\mathrm{ext}(M,P) = c_M\cdot N(M)$。

对于方差，考虑一次抽样 $Y = N(P_4)\,\mathrm{ext}(M,P)/c_M$；由于 k路径是独立^抽样的（有放回），$N(M)$是 k个独立同分布副本 Y的平均值。令 X_i表示 i个副本 M，$1 \leq i \leq N(M)$中的第 i个副本是否包含所抽样的路径。

$$\begin{aligned}\mathbb{V}ar[Y] &= \mathbb{V}ar\Big[\frac{N(P_4)}{c_M}\sum_{i=1}^{N(M)}X_i\Big]\\ &= \frac{N(P_4)^2}{c_M^2}\,\mathbb{V}ar\Big[\sum_{i=1}^{N(M)}X_i\Big]\\ &= \frac{N(P_4)^2}{c_M^2}\Big[\sum_{i=1}^{N(M)}\mathbb{V}ar[X_i] + \sum_{i\neq j}\mathbb{C}ov(X_i,X_j)\Big]\\ &= \frac{N(P_4)^2}{c_M^2}\Big[N(M)\Big(\frac{c_M}{N(P_4)} - \frac{c_M^2}{N(P_4)^2}\Big)\\ &\quad + \sum_{i\neq j}\big(\mathbb{E}[X_iX_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j]\big)\Big]. \qquad (1)\end{aligned}$$

我们只考虑至少共享一条种子路径的配对 $\{X_i, X_j\}$，因为否则 $\mathbb{E}[X_iX_j] = 0$ 和 $\mathbb{C}ov(X_i,X_j) = 0$。若实例 i,j 共享 a_{ij} 条种子路径，则 $\mathbb{E}[X_iX_j] = \Pr[X_i{=}1]\Pr[X_j{=}1|X_i{=}1] = \frac{c_M}{N(P_4)}\frac{a_{ij}}{N(P_4)} = \frac{a_{ij}}{c_M}$，因此

$$\mathbb{C}ov(X_i,X_j) = \frac{a_{ij}}{N(P_4)} - \frac{c_M^2}{N(P_4)^2}.$$

利用 $a_{ij} \leq c_M$ 并忽略负贡献，我们得到

$$\mathbb{V}ar[Y] \leq \frac{N(P_4)}{c_M}\big(N(M)+p\big),$$

^和 $\mathbb{V}ar[N(M)] = \mathbb{V}ar[Y]/k$，由独立性可得，这就是所述界。令 $k = \frac{8N(P_4)(N(M)+p)c_M}{\epsilon^2N(M)^2}$。由切比雪夫不等式可得

$$\begin{aligned}\Pr[|\widehat{N}(M) - N(M)| \geq \epsilon N(M)] &\leq \frac{\mathbb{V}ar[\widehat{N}(M)]}{\epsilon^2N(M)^2}\\ &\leq \frac{N(P_4)(N(M)+p)}{c_Mk\epsilon^2N(M)^2} = \frac{1}{8}.\end{aligned}$$

因此，我们可以通过取 $O(\log(1/\delta))$ 个估计量的中位数，将此估计量转化为一个(ϵ,δ)-估计量。

□

命题 2. 对于一个最大度为 Δ、树深度为 α、含有n 个顶点和 m 条边的图，其运行时间和总存储空间

表4：二部数据集上的模体诱导计数：BiLight（$k = 100$K）与 Motivo（5000 万个颜色编码样本）的比较。缺失项表示 Motivo 未报告该预算下的模体。

数据集	方法	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M13	M14	M15	M16	M17	运行时间
亚马逊电气	BiLight	3.62e6	1.07e12	2.45e8	5.36e13	3.76e10	2.98e9	8.70e14	3.57e12	5.92e14	8.42e11	1.24e10	2.22e17	1.21e15	5.94e14	6.01e16	1.04e14	4 分钟
	Motivo	—	—	—	2.72e12	—	—	3.72e14	1.54e12	1.55e14	2.21e11	—	2.17e17	1.24e15	6.09e14	5.35e16	9.97e13	5 小时 3 分钟
亚马逊图书	BiLight	3.52e9	1.25e13	1.10e11	5.45e14	1.21e13	8.78e11	1.96e16	7.93e14	5.10e15	1.36e14	1.32e12	2.12e19	6.72e17	8.77e16	4.34e17	5.23e15	23 分钟
	Motivo	—	1.15e13	—	4.66e14	1.78e13	—	1.96e16	9.71e14	4.97e15	9.17e13	—	2.41e19	8.40e17	1.17e17	5.09e17	7.12e15	12小时 57分钟
IMDB	BiLight	1.06e8	5.84e9	1.13e9	1.27e11	1.93e10	5.44e9	1.56e12	4.58e11	9.68e11	2.84e11	1.34e10	7.60e13	1.24e13	1.95e13	4.14e13	9.73e12	45 s
	Motivo	1.16e8	5.07e9	1.14e9	1.13e11	1.86e10	5.34e9	1.51e12	4.51e11	9.22e11	2.79e11	1.30e10	7.50e13	1.22e13	1.94e13	4.05e13	9.67e12	16 分钟
Flickr	BiLight	2.46e13	1.55e16	5.87e14	3.34e17	1.22e16	2.96e15	3.26e18	1.85e17	1.16e18	6.21e16	2.32e15	7.17e19	1.45e18	1.66e18	1.36e19	4.41e17	1 小时 17 分钟
	Motivo	1.98e13	1.69e16	5.86e14	3.47e17	1.22e16	3.00e15	3.30e18	1.87e17	1.18e18	6.26e16	2.31e15	7.16e19	1.46e18	1.65e18	1.36e19	4.40e17	5 小时 34 分钟

（包括所有预处理）分别为 *BiLight* 的 $O(\alpha m + k\Delta^2)$ 和$O(n + m)$。

证明。预处理（算法 1）会构建边权重和别名表。计算每条边的权重需要 $|\Gamma(u) \cap \Gamma(v)|$，因此预处理成本为 $O\left(\sum_{(u,v)\in E} \min(d_u, d_v)\right)$，这也是枚举三角形的通常界限；对于树深度为 α [47] 的图，该界限为 $O(\alpha m)$。我们的实现首先在 $O(n+m)$中测试二分性，并在二分图宿主上跳过交集计算，因为这些交集全部为空，从而将此步骤降至 $O(m)$。这是一项一次性成本，与采样预算 k 无关。每个样本的抽样成本为 $O(1)$，评分成本为 $O(\Delta^2)$：表 2 中的模式可归约为邻接表交集，其中每个交集使用哈希集合在 $O(\Delta)$ 时间内完成；而对于需要在邻居集合上对交集求和的模式（例如 M11），则需要 $O(\Delta)$ 个交集。因此，k 个样本的总时间为 $O(\alpha m+k\Delta^2)$，与采样预算呈线性关系；相比之下，对 6 顶点元组进行精确枚举的复杂度为 $O(\Delta^4)$ 或更高，在大型图上难以承受。附录 A 中的图 7 通过实验证实了与 k 的线性关系。

存储预算与 k 无关。 G 的邻接表表示需要 $O(n + m)$ 空间。在构建每条边的边权重 w_e 时，我们的实现还会额外保留一个显式的 m边列表e。两个别名数组 $prob[\cdot]$ 和 $alias[\cdot]$ 额外占用 $O(m)$ 个字。每次采样之间，BiLight 只更新一组固定的模体累加器（每个模体一个）；而局部计数期间使用的临时结构（例如邻居集的交集）在最坏情况下为 $O(\Delta)$ ，并在各次 k迭代之间重复使用，而不是不断累积。□

注 1。 在采样预算较小时，运行时间主要由一次性预处理决定。

例如，在 IMDB 上，设置大约需要 10 秒，而在 $k = 10$ 上的估计耗时不到 30 毫秒。图 7（附录 A）仅绘制了估计时间；从最小预算开始，该时间呈线性增长。其轻微的不规则性反映出，对采样 P_4 进行评分的成本会随其顶点的度而变化，并非偏离 $O(\alpha m + k\Delta^2)$ 界。

6 实验评估

所有实验（BiLight 和 Motivo [65], [66], [67]）均在一台机器上运行：搭载 Intel Core i7-10750H CPU（2.60GHz，6 核）、32 GB 内存并运行 Windows 11 企业版。用于生成这些结果的 BiLight 实现已作为开源代码发布，地址为 github.com/ celikmustaa/BiLight。

BiLight 可用于任何无向图，并估算该图中作为子图的 17 种二部 六顶点模体的数量。如表3所列，我们使用了四个大型二部数据集：亚马逊评分（图书和电子产品）、IMDB（演员–电影）、Flickr（用户–群组），以及中型到大型的非二部图。表3列出了所有数据集；顶部区块为二部数据集，并限制在最大连通分量（LCC）内。

6.1 诱导与非诱导模体

使用 BiLight，我们能够获得二部网络以及非二部网络中图1所示模体的非诱导副本计数。如 [41], 所观察到的，可以将网络中模体的诱导计数与非诱导计数形式化为线性关系。我们将这一思想推广到我们的框架中，令$\mathbf{C}$ 和$\mathbf{N}$ 表示具有17 个条目的向量，其中对于每个 i、 C_i 和 N_i 分别表示模体 M_i 的诱导副本数和非诱导副本数。我们在附录B中给出了 17×17 矩阵A，使得对于任意给定的二部网络，$\mathbf{N} = A\mathbf{C}$，其中$A_{i,j}$ 表示在模体 M_j的一个实例中出现的模体M_i 的（非诱导）副本数。反之，给定 N 后，可以通过关系$\mathbf{C} = A^{-1}\mathbf{N}$获得 C。对于适当的模体排序， A 是下三角矩阵。通过这一转换，我们能够获得二部网络中图1所示每个模体的诱导副本计数，并将其用于把我们的结果与最先进的算法进行比较。

表5：非二部数据集上的平均距离误差（越低越好）。所有条目均为比例（例如， 0.10 =10%）。参考值为 BiLight 在 $k = 100$K 下的自身估计值。

数据集	$k = 1$K	$k = 5$K	$k = 20$K
词语关联	0.16	0.05	0.01
facebook-wall	0.15	0.05	0.05
facebook-links	0.10	0.03	0.02
com-amazon	0.06	0.03	0.01
com-dblp	0.06	0.03	0.00
web-BerkStan	0.10	0.08	0.02
as-skitter	0.05	0.03	0.02
cit-Patents	0.04	0.10	0.02
LiveJournal	0.07	0.05	0.03
好莱坞	0.05	0.00	0.01
Soc-Orkut	0.22	0.22	0.08
Orkut-Links	0.11	0.12	0.06

6.2 性能与先前工作的比较

由于精确计数不可行，我们将*groundtruth*定义为 BiLight 在 $k = 100$K 时的估计值。对于二部数据集，

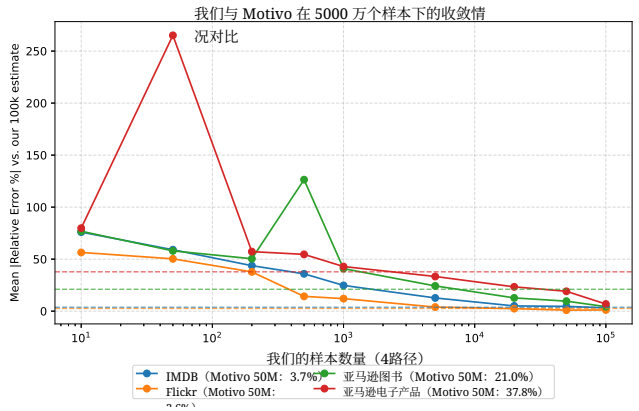

图 2：BiLight 的收敛性：在预算为 k 和 $k = 100$K 时，BiLight 的诱导计数之间的平均相对差异（%），在采样的 16 个模体上取平均。水平虚线：在适用情况下，在 Motivo 报告的模体上取平均，BiLight 在 $k = 100$K 时与 Motivo 使用 50M 个样本时之间的平均相对差异。

每个预算下均进行了五次独立运行；在计算误差前，我们先分别对目标预算和参考预算下各次运行的估计值取平均。对于非二部数据集的单次运行实验，则直接将 $k = 100$K 时的值作为参考值。我们使用平均距离误差衡量准确度：对 16 个非平凡模体（不包括 M12），计算各模体相对误差的平均值 $|\widehat{N}(M_i) - N_{\text{ref}}(M_i)|/N_{\text{ref}}(M_i)$。表5报告了非二部数据集在 $k = 1$K、5K 和 20K 时的误差。12 个网络中有 9 个在 $k = 5$ K 时的误差收敛至 8% 以内，另有 9 个在 $k = 20$K 时达到 4% 或更低；Orkut 的两张图按边数计是其中最大的，也是收敛最慢的。对于二部数据集，四张图在 $k = 50$K 时的平均误差均低于 10%，其中 Flickr 达到 0.3%。重复性与图的密度相关：在 $k = 50$K 时，各模体估计值在不同运行间的变异系数，Flickr 平均为 0.01，IMDB 为 0.06；对于四者中最稀疏的亚马逊图书和亚马逊电子产品，则分别升至 0.12 和0.18。

Bressan 等人 [45] 使用彩色编码解决了一般图中超过五个顶点的模体计数问题，随后发布了 Motivo [65], [66], [67],，通过简洁数据结构和自适应采样扩展了该方法。Motivo 在单一框架中估计一般图（包括非二部图和二部图）中具有 5 至 8 个顶点的诱导 模体的频率。

我们在四个二部数据集（IMDB、亚马逊图书和电子产品评分数据集、Flickr）上运行了这两种方法：BiLight 的采样预算从 $k = 1$K 到 $k = 100$K，Motivo [65], [66], [67] 使用 10K、100K、1M、10M 以及（在可行的情况下）50M 个颜色编码样本。

表4详细列出了每种模体的诱导计数：BiLight 在 $k = 100$ K 时的计数（对独立运行取平均值，并通过 A^{-1} 换算）与 Motivo 在 50M 个样本时的计数。由于 BiLight 通过解析方法而非采样获得 M12，因此未列出 M12。在 IMDB 和 Flickr 上，16 种可比较模体的计数都相当接近，IMDB 上只有两种模体的差异超过 10%，Flickr 上则只有一种；而在亚马逊数据集上，即使有 50M 个样本，Motivo 对若干模体的观测仍几乎为零（在图书数据集上为 M1、M3、M6 和 M11，在电子产品数据集上还包括 M2 和 M5；在电子产品数据集上，它对 M3 返回 74，而 BiLight 的结果为 2.45×10^8），而 BiLight 则报告了所有这些模体。图8（附录 C）展示了同一比较在不同预算下的变化：在 IMDB 和 Flickr 上，BiLight 与 Motivo 颜色最深的柱形“在中间相遇”，直观地证实随着预算增加，两种方法的诱导计数逐渐趋于相同；而在亚马逊数据集上，对于 Motivo 几乎未能命中的模体，两者之间的差距依然存在。

图2将这些按模体划分的差异汇总为每个数据集的一条曲线。对于给定的预算 k，我们通过 A^{-1}将 BiLight 的非诱导估计值转换为诱导计数，并绘制 16 个采样模体上估计值在 k 与$k = 100$ K 时的相对差异均值；每条虚线表示一个数据集中，Motivo 所报告模体在 $k = 100$K 时的 $|C_i^{\text{BiLight}} - C_i^{\text{Motivo}}|/C_i^{\text{BiLight}}$ 均值。下文提到的百分比即为第二个量的取值。随着 k 增加，曲线逐渐下降；在 IMDB 和 Flickr 上，Motivo 报告了全部 16 种非平凡模体，BiLight 在 $k = 100$K 时的结果与其使用 50M 个样本运行所得的结果相差不超过 4%（3.7% 和 2.6%）。在亚马逊数据集上，即使使用 50M 个样本，Motivo 的覆盖范围仍不完整（图书为 12/16，电子产品为 10/16），且对于其确实报告的模体，两种方法的结果一致性不一。在图书数据集上，平均差异约为 21%。在四个数据集中最稀疏的电子产品数据集上，对于数量丰富的树状模体（M_{13}–M_{15} 和 M_{17}），两种方法的结果相差不超过4%；但对于较罕见的含环模体（M_4、M_7–M_{10}），结果相差两倍或更多，而 Motivo 使用 50M 个样本仅获得少量命中；BiLight 对这些模体的估计值在 $k = 50$K 和$k = 100$K 之间保持稳定。这反映出通用着色编码方法难以处理这些稀疏的二部结构。

6.2.1 收敛

我们通过将采样预算从 $k = 10$增加到 $k = 500$K，分析了 BiLight 的收敛性。图3展示了四个二部数据集上全部16种采样模体的估计计数：随着 k 增大，曲线逐渐重叠，方差不断减小，表明其具有很强的收敛性。图 4 根据表3中的边数，在四个最大的非二分网络上重复了该实验。相同的 P_4 采样和局部计数模式无需任何修改即可适用，估计结果也以相同方式趋于稳定，证实 BiLight 并不局限于二部输入。

6.2.2 运行效率：

表4的运行时间列总结了运行时间差距：BiLight 在 $k = 100$K 时，在 IMDB 上用时 45秒，在 Flickr 上最多用时 1小时 17分钟；相比之下，Motivo [65], [66],[67] 在相同输入上使用 50M 个样本时，运行时间从 16 分钟到近 13小时不等。就我们所知，BiLight 是首个能够在大型网络上高效估计全部十七种六顶点二部模体的框架。

模体计数趋势：

图5和图6总结了最大预算（分别为$k = 500$K和 $k = 100$K）下估算的模体分布。不同数据集的分布趋势相似（仅尺度不同），说明尽管领域不同，模体的相对重要性仍具有可比性；非二部网络则高出

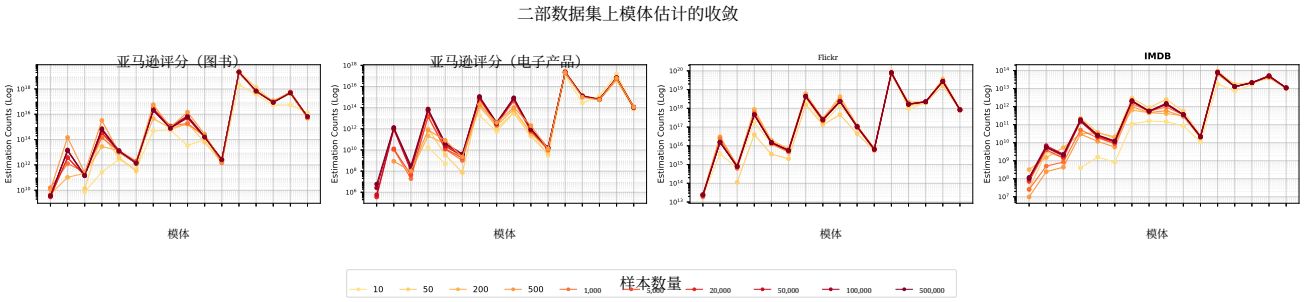

图 3：四个二部数据集上的模体计数估计收敛情况（每个数据集对应一个子图）。 x-轴为 = 模体索引（不包括 M12）， y-轴为 = 估计计数（对数刻度），线条颜色表示 = 采样预算 k （颜色越浅表示 = 越小 k）。

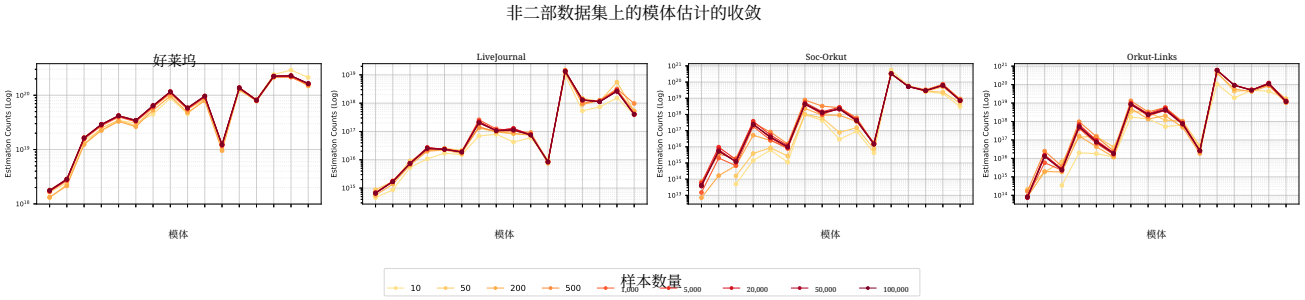

图4：四个最大的非二部网络上模体计数估计的收敛情况；布局同图3。

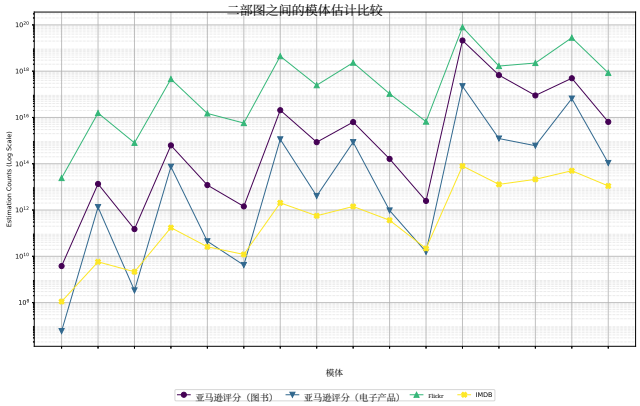

图5：四个二部数据集在 $k = 500$K 处的估计模体谱（非诱导）（每个数据集一条线）。

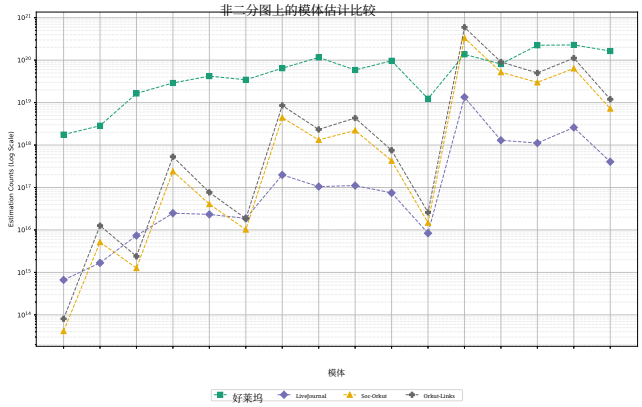

图6：四个最大的非二部网络在 $k = 100$K 处的估计模体谱（非诱导）。

好几个数量级，这归因于其更高的边密度和 P_4 计数。在图6中，好莱坞网络似乎各种模体的分布较为均衡，这或许也与该网络的 P_4数量最多有关，正如表3所示。图5表明，M1、M3、M6和M11这四种包含六个顶点环的模体，在全部四个数据集中始终属于出现频率最低的模体。正如预期，模体边数越多，计数越低。五种树模体（M_{13}–M_{17}）边数最少，在每个数据集中排名都处于前半；其中 M_{13} 始终最常见；$M_1 = K_{3,\ 3}$是唯一一种有九条边的模体，在所有数据集中都最罕见。

P_4 采样，从而得到无偏且低方差的估计。我们在大规模网络上的评估表明，BiLight 准确且具备良好的可扩展性。与 Motivo 的实证比较 [65], [66], [67] 表明，我们的框架速度显著更快；在 Motivo 的采样覆盖全部 16 种模体的数据集上，两者的结果也高度一致。BiLight 专为 17 种二部六顶点模体设计，能够快速且稳定地给出所有估计值。我们提供了源代码，也可用于将我们的框架推广至非二部模体。

7 结论

我们提出了 BiLight，这是一个轻量级框架，通过加权方式估计

6 顶点二部模体频率

附录 A 运行时间结果

图 7 绘出了 BiLight 的估计时间；在二部图和非二部图上，该时间都随着采样预算 k 近似线性增长。

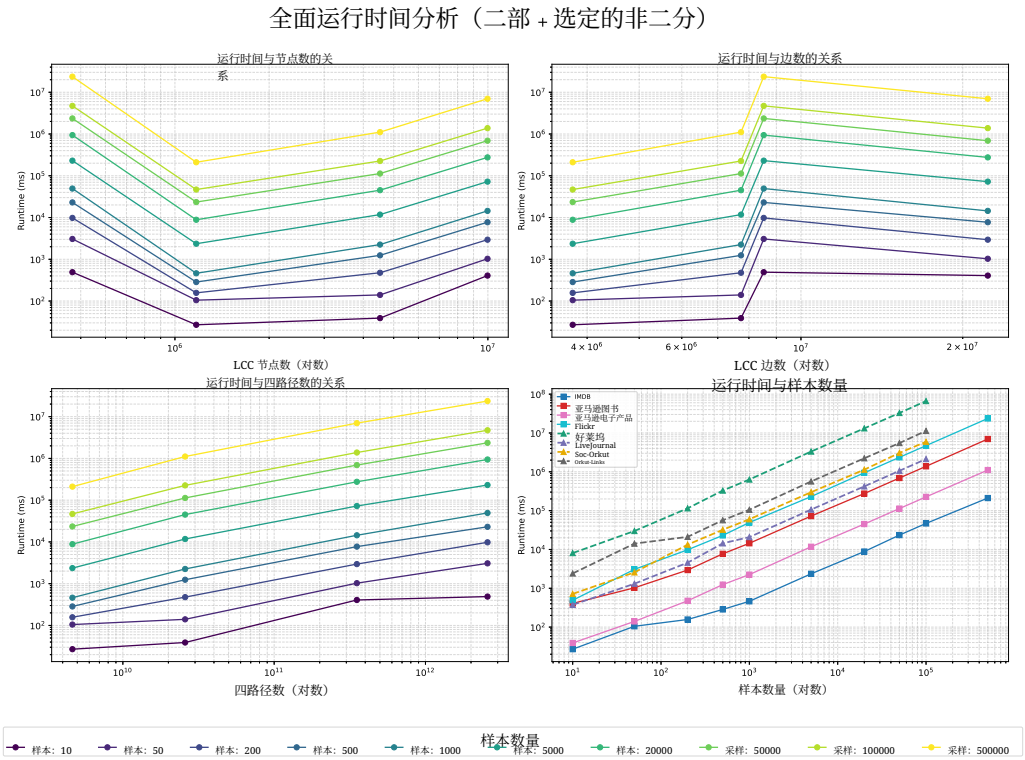

图 7：BiLight 运行时间分析；四个面板分别绘制估计时间（不包括一次性设置时间）与不同量的关系。上排（对数刻度）：运行时间与最大连通分量顶点数的关系（左），以及运行时间与边数的关系（右）。下排：运行时间与 P_4 数的关系（左），以及运行时间与采样预算 k 的关系（右）。顶点数、边数和 P_4数面板仅使用二部数据集；在采样预算面板中，实线表示二部数据集，虚线表示图 4 中的四个大型非二分图。

附录 B 转换矩阵

表 6 和 7 给出了 17 种模体的 A 和 A^{-1}（行/列为 M1–M17；空白单元格为零）。

表6：直接矩阵 A：$\mathbf{N} = A\mathbf{C}$。条目 $A_{i,j}$ 表示模体 M_i 在模体 M_j中所包含的副本数。

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17
M1	1																
M2		1															
M3	9		1														
M4		8		1													
M5	18		4		1												
M6	18		4			1											
M7		12		3			1										
M8	36		12		4	2		1									
M9		12		3					1								
M10	36		12		2	4				1							
M11	6		2			1					1						
M12												1					
M13		8		3			2						1				
M14	9		4		2	1		1						1			
M15	36		16		6	6		2		2					1		
M16		24		9			2		4							1	
M17	36		16		4	8		1		2	6						1

表7：逆矩阵 A^{-1}：$\mathbf{C} = A^{-1}\mathbf{N}$。行 i，列 j 分别对应模体 M_i、M_j。

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17
M1	1																
M2		1															
M3			1														
M4		−8		1													
M5	18		−4		1												
M6	18		−4			1											
M7		12		−3			1										
M8			12		−4	−2		1									
M9		12		−3					1								
M10			12		−2	−4				1							
M11	−6		2			−1					1						
M12												1					
M13		−8		3			−2						1				
M14	9		−4		2	1		−1						1			
M15	36		−16		6	6		−2		−2					1		
M16		−24		9			−2		−4							1	
M17	36		−16		4	8		−1		−2	−6						1

附录 C 逐模体比较

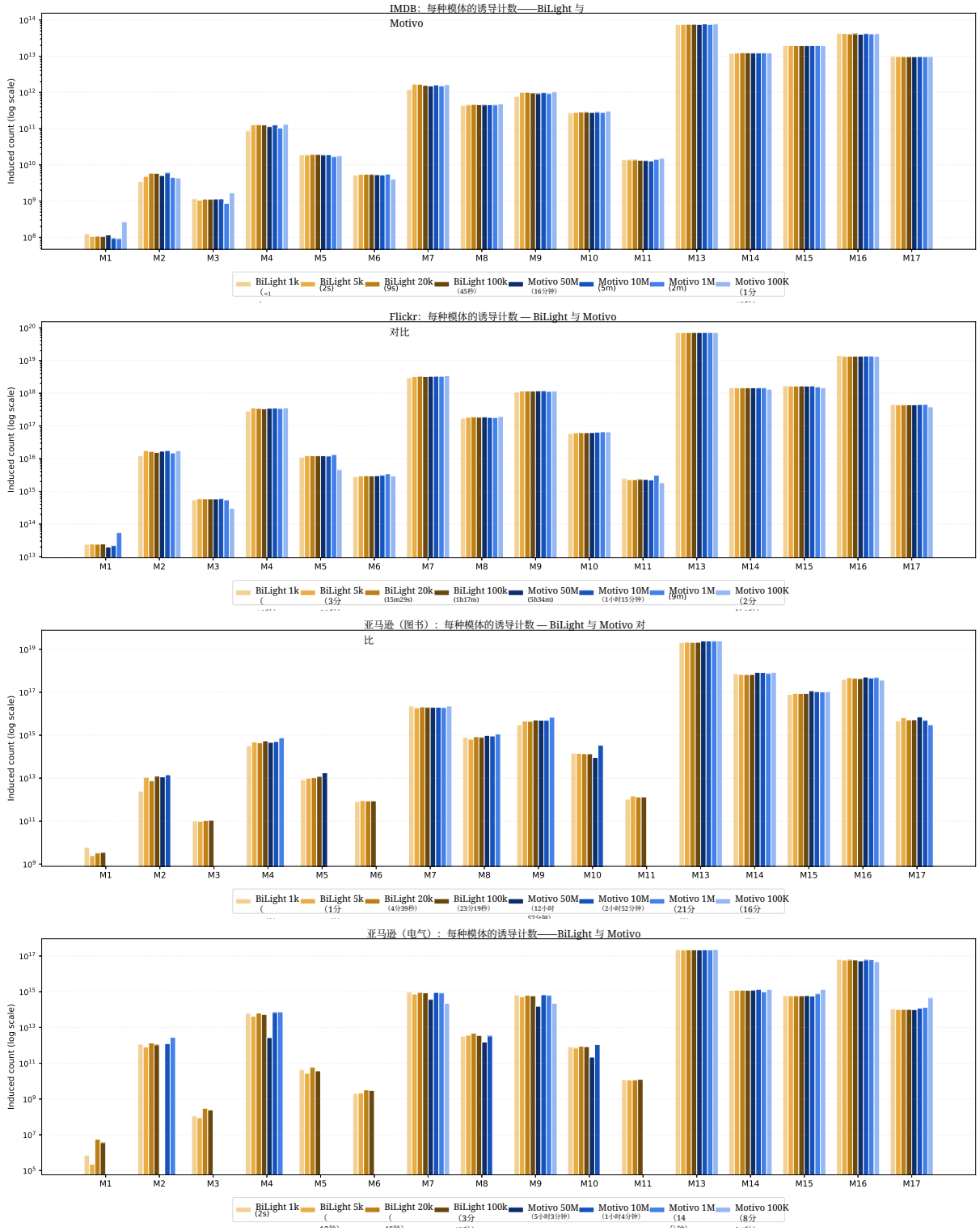

图 8：随着样本预算增加，各模体的诱导计数。面板（从上到下）：IMDB；Flickr；亚马逊评分（图书）；亚马逊评分（电子产品）。青绿色柱 = BiLight（由浅到深：k = 1K、5K、20K、100K）；橙色柱 = Motivo（由深到浅：50M、10M、1M、100K 个颜色编码样本）。实际运行时间显示在括号中。缺失的柱表示 Motivo 未报告该预算下的模体计数。BiLight 柱为五次独立运行的平均值。最稀有的模体 $M_1(K_{3,\,3})$，在亚马逊（电子产品）数据集中仅在五次运行中的一次（预算为 k=1K 时）和五次中的三次（预算为 k=5K 时）被发现，因此这些预算下的柱仅基于少数几次命中；从 k=20K 起，它在每次运行中都能被发现。

致谢

作者在准备本研究时使用了 Anthropic 的 Claude。算法、局部计数模式、理论分析、转换矩阵和实验均由作者完成。Claude 的贡献主要体现在第 4 和第 5 节：建议使用 Walker 别名法代替 $O(\log m)$ 前缀和抽样，将预处理成本修正为 $O(\alpha m)$，通过暴力验证检查局部计数模式，并协助高效地呈现实验结果。作者审阅并修订了 Claude 对文本和代码所作的贡献，并对本文承担全部责任。

参考文献

[1] P. Holland 和 S. Leinhardt，《一种检测社会计量数据中结构的方法》，《美国社会学杂志》，第 76 卷，第 492–513 页，1970 年。[2] J. Coleman，《社会资本在创造人力资本中的作用》，《美国社会学杂志》，第 94 卷，第 S95–S120 页，1988 年。[3] A. Portes，《社会资本：其起源及其在现代社会学中的应用》，《社会学年度评论》，第 24 卷，第 1 期，第 1–24 页，1998 年。[4] R. Milo、S. Shen-Orr、S. Itzkovitz、N. Kashtan、D. Chklovskii 和 U. Alon，《网络模体：复杂网络的简单构建模块》，《科学》，第 298 卷，第 5594 期，第 824–827 页，2002 年。[5] R. Burt，《结构洞与好点子》，《美国社会学杂志》，第 110 卷，第 2 期，第 349–399 页，2004年。[6] N. Pržulj、D. G. Corneil 和 I. Jurisica，《互作组建模：无标度还是几何？》，《生物信息学》，第 20 卷，第 18 期，第 3508–3515 页，2004年。[7] G. Fagiolo，《复杂有向网络中的聚类》，《物理评论E》，第 76 卷，第 026107 页，2007 年。[8] F. Hormozdiari、P. Berenbrink、N. Pržulj 和 S. C. Sahinalp，《并非所有无标度网络生来平等：种子图在蛋白质相互作用网络演化中的作用》，《公共科学图书馆计算生物学》，第 3 卷，第 7 期，第 e118 页，2007 年。[9] L. Becchetti、P. Boldi、C. Castillo 和 A. Gionis，《海量图中局部三角形计数的高效半流式算法》，载于 KDD '08 论文集，2008 年，第 16–24 页。[10] K. Faust，《关于社交网络中三元组的一个难题：图约束与三元组普查》，《社交网络》，第 32 卷，第 3 期，第 221–233 页，2010 年。[11] M. Szell 和 S. Thurner，《测量大型多人在线游戏中的社会动态》，《社交网络》，第 32 卷，329，2010 年，第 313– .页。[12] B. Welles、A. Van Devender 和 N. Contractor，《朋友是否真是朋友？：考察虚拟世界中友谊网络的结构》，载于 CHI-EA '10 论文集，2010 年，第 4027–4032 页。[13] S. Son、A. Kang、H. Kim、T. Kwon、J. Park 和 H. Kim，《大型多人在线角色扮演游戏中社会交互网络情境依赖性的分析》，《公共科学图书馆·综合》，第 7 卷，第 4 期，第 e33918 页，2012 年 4 月。[14] D. J. Watts 和 S. H. Strogatz，《"小世界"网络的集体动态》，《自然》，第 393 卷，第 6684 期，第 440–442 页，1998 年。[15] C. Seshadhri、T. G. Kolda 和 A. Pinar，《Erd˝os-R´enyi 图的群落结构与无标度集合》，《物理评论E》，第 85 卷，第 5 期，第 056109 页，2012 年 5 月。[16] N. Durak、A. Pinar、T. G. Kolda 和 C. Seshadhri，《现实世界网络与图模型中三角形的度关系》，载于第 21 届 ACM 信息与知识管理国际会议（CIKM '12）论文集，2012 年，第 1712–1716 页。[17] J. Ugander、L. Backstrom 和 J. M. Kleinberg，《子图频率：描绘大型图集合的经验与极值地理分布》，载于万维网大会（WWW）论文集，2013 年，第 1307–1318 页。[18] S. S. Shen-Orr、R. Milo、S. Mangan 和 U. Alon，《大肠杆菌转录调控网络中的网络模体》，《自然遗传学》，第 31 卷，第 1 期，第 64–68 页，2002 年。[19] H. Chun、H. Kwak、Y.-H. Eom、Y.-Y. Ahn、S. Moon 和 H. Jeong，《从数量与交互角度比较在线社会关系：以 Cyworld 为例》，载于第 8 届 ACM SIGCOMM互联网测量会议论文集，2008 年，第 57–70 页。[20] J. Kunegis、A. Lommatzsch 和 C. Bauckhage，《Slashdot 动物园：利用负边挖掘社交网络》，载于《第18届国际万维网大会论文集》。ACM，2009年，第741–750页。[21] J. Zhao、J. C. S. Lui、D. Towsley、X. Guan 和 Y. Zhou，《关注者网络演化的实证分析：以豆瓣为例》，载于《2011年 IEEE 计算机通信会议研讨会（INFOCOM WKSHPS）》，2011年，第924–929页。[22] Y. Jin、E. Sharafuddin 和 Z.-L. Zhang，《通过应用流量活动图分解揭示核心网络范围内的通信模式》，载于《第11届计算机系统测量与建模国际联合会议论文集》，2009年，第49–60页。[23] M. Iliofotou、M. Faloutsos 和 M. Mitzenmacher，《利用基于图的流量分析中的动态性：技术与应用》，载于《第5届新兴网络实验与技术国际会议论文集》，2009年，第241–252页。[24] N. Shervashidze、S. V. N. Vishwanathan、T. Petri、K. Mehlhorn 和 K. Borgwardt，《用于大型图比较的高效图元核》，载于《第十二届人工智能与统计学国际会议论文集》，机器学习研究论文集，第5卷。PMLR，2009年，第488–495页。[25] P. D. Dobson 和 A. J. Doig，《无需比对即可区分酶结构与非酶结构》，《分子生物学杂志》，第330卷，第4期，第771–783页，2003年。[26] S. V. N. Vishwanathan、N. N. Schraudolph、R. Kondor 和 K. Borgwardt，《图核》，《机器学习研究杂志》，第11卷，第1201–1242页，2010年。[27] N. Wale 和 G. Karypis，《化合物检索与分类的描述符空间比较》，载于《2006年 IEEE 国际数据挖掘会议》，2006年，第678–689页。[28] P. Yanardag 和 S. V. N. Vishwanathan，《用于鲁棒图比较的结构平滑框架》，载于《神经信息处理系统进展28（NIPS 2015）》，2015年。[29]——，《深度图核》，载于《第21届 ACM SIGKDD国际知识发现与数据挖掘会议论文集》。ACM，2015年，第1365–1374页。[30] K. Chen、P. Wang、Y. Lee、X. Wang、N. Zhang、H. Huang、W. Zou 和 P. Liu，《10秒内发现未知恶意行为：面向 Google Play 规模的新威胁大规模审查》，载于《第24届 USENIX 安全研讨会（USENIX Security 15）》。USENIX协会，2015年，第659–674页。[31] T. Shen、Y. Zhongyang、Z. Xin、B. Mao 和 H. Huang，《利用基于组件的拓扑图检测 Android 恶意软件变种》，载于《2014年 IEEE 第13届计算与通信中的信任、安全与隐私国际会议》，2014年，第406–413页。[32] H. Gascon、F. Yamaguchi、D. Arp 和 K. Rieck，《利用嵌入式调用图对 Android 恶意软件进行结构检测》，载于《2013年 ACM 人工智能与安全研讨会论文集》，2013年，第45–54页。[33] A. E. Sariy¨uce 和 A. Pinar，《通过剥离二部网络发现稠密子图》，载于《第11届 ACM 网络搜索与数据挖掘国际会议论文集》，2018年，第504–512页。[34] C. Seshadhri、A. Pinar 和 T. G. Kolda，《图上的三元组测度：楔形采样的效力》，载于《SIAM 国际数据挖掘会议论文集》。SIAM，2013年，第10–18页。[35] N. K. Ahmed、J. Neville、R. A. Rossi、N. G. Duffield 和 T. L. Willke，《图元分解：框架、算法与应用》，《知识与信息系统》，第50卷，第3期，第689–722页，2017年。[36] N. Betzler、R. van Bevern、M. R. Fellows、C. Komusiewicz 和 R. Niedermeier，《在生物网络中寻找连通模体的参数化算法》，《IEEE/ACM计算生物学与生物信息学汇刊》，第8卷，第5期，第1296–1308页，2011年。[37] E. Wong、B. Baur、S. Quader 和 C.-H. Huang，《生物网络模体检测：原理与实践》，《生物信息学简报》，第13卷，第2期，第202–215页，2012年。[38] M. Rahman、M. A. Bhuiyan 和 M. Al Hasan，《GRAFT：一种用于大图分析的高效图元计数方法》，《IEEE知识与数据工程汇刊》，第26卷，第10期，第2466–2478页，2014年。[39] N. K. Ahmed、J. Neville、R. A. Rossi 和 N. Duffield，《大型网络的高效图元计数》，载于《第15届 IEEE 国际数据挖掘会议（ICDM）论文集》，2015年，第1–10页。

[40] M. Jha、C. Seshadhri 和 A. Pinar，《路径采样：一种快速且可证明有效的4顶点子图计数估计方法》，载于《第24届国际万维网大会论文集》，2015年，第495–505页。[41] A. Pinar、C. Seshadhri 和 V. Vishal，《ESCAPE：高效计算所有5顶点子图》，载于《第26届国际万维网大会（WWW'17）论文集》。ACM，2017年，第1431–1440页。[42] P. Wang、J. Zhao、X. Zhang、Z. Li、J. Cheng、J. C. Lui、D. Towsley、J. Tao 和 X. Guan，《MOSS-5：一种在大型图中近似计算5节点图元计数的快速方法》，《IEEE知识与数据工程汇刊》，第30卷，第1期，第73–86页，2018年。[43] J. Niu、J. Zola 和 A. E. Sariyuce，《二部图中诱导6-环的计数》，载于《第51届国际并行处理会议论文集》，2022年，第1–10页。[44] J. Niu、J. Zola 和 A. E. Sarıy¨uce，《二部网络中诱导6-环的快速计数与利用》，《IEEE知识与数据工程汇刊》，2025年。[45] M. Bressan、F. Chierichetti、R. Kumar、S. Leucci 和 A. Panconesi，《五节点以上的模体计数》，《ACM数据知识发现汇刊》，第12卷，第4期，第48:1–48:25页，2018年。[46] P. Ribeiro、P. Paredes、M. E. P. Silva、D. Aparicio 和 F. Silva，《子图计数综述：概念、算法及其在网络模体和图元中的应用》，《ACM计算调查》，第54卷，第2期，第28:1–28:36页，2021年。[47] N. Chiba 和 T. Nishizeki，《树深度与子图枚举算法》，《SIAM计算杂志》，第14卷，223，1985 1，210第。–页。[48] C. Tsourakakis、M. N. Kolountzakis 和 G. Miller，《三角形稀疏化器》，《图算法与应用杂志》，第15卷，第703–726页，2011年。[49] M. Gonen 和 Y. Shavitt，《网络模体数量的近似计算》，《互联网数学》，第6卷，第3期，第349–372页，2009年。[50] D. Marcus 和 Y. Shavitt，《网络模体的高效计数》，载于 ICDCS研讨会。IEEE计算机学会，2010年，第92–98页。[51] S. K. Bera、L. Gishboliner、Y. Levanzov、C. Seshadhri 和 A. S hapira，《退化图中的子图计数》，《ACM杂志》，第69卷，第3期，第23:1–23:21页，2022年。[52] M. A. Bhuiyan、M. Rahman、M. Rahman 和 M. Al Hasan，《GUISE：用于大图分析的图元均匀采样》，载于《第12届 IEEE 国际数据挖掘会议（ICDM）论文集》，2012年，第91–100页。[53] P. Wang、J. C. S. Lui、B. Ribeiro、D. Towsley、J. Zhao 和 X. Guan，《大型网络模体统计量的高效估计》，《ACM数据知识发现汇刊》，第9卷，第2期，第8:1–8:27页，2014年。[54] N. Kashtan、S. Itzkovitz、R. Milo 和 U. Alon，《用于估计子图浓度和检测网络模体的高效采样算法》，《生物信息学》，第20卷，1758，2004 11，1746第。–页。[55] S. Wernicke，《网络模体的高效检测》，《IEEE/ACM计算生物学与生物信息学汇刊》，第3卷，第4期，第347–359页，2006年。[56] S. Omidi、F. Schreiber 和 A. Masoudi-Nejad，《MODA：一种用于生物网络网络模体发现的高效算法》，《基因与遗传系统》，第84卷，第5期，第385–395页，2009年。[57] X. Chen、Y. Li、P. Wang 和 J. C. S. Lui，《通过随机游走估计图元统计量的通用框架》，《VLDB基金会会刊》，第10卷，第3期，第253–264页，2016年。[58] T. Schank 和 D. Wagner，《聚类系数与传递性的近似计算》，《图算法与应用杂志》，第9卷，第265–275页，2005年。[59] T. G. Kolda、A. Pinar、T. Plantenga、C. Seshadhri 和 C. Task，《利用 MapReduce 在海量图中计数三角形》，《SIAM科学计算杂志》，第36卷，第5期，第S48–S77页，2014年。[60] N. Alon、R. Yuster 和 U. Zwick，《彩色编码：一种在大型图中寻找简单路径、环及其他小型子图的新方法》，载于《第26届 ACM计算理论年度研讨会（STOC）论文集》，1994年，第326–335页。[61] C. E. Tsourakakis、U. Kang、G. L. Miller 和 C. Faloutsos，《DOULION：用硬币在海量图中计数三角形》，载于《第15届 ACM SIGKDD国际知识发现与数据挖掘会议论文集》，2009年，第837–846页。[62] A. Pavan、K. Tangwongsan、S. Tirthapura 和 K. Wu，《从图流中计数和采样三角形》，《VLDB基金会会刊》，第6卷，第14期，第1870–1881页，2013年。[63] M. Jha、C. Seshadhri 和 A. Pinar，“利用生日悖论进行三角形计数的空间高效流式算法”，载于第19届 ACM SIGKDD国际知识发现与数据挖掘会议论文集。ACM，2013年，第589–597页。[64] N. K. Ahmed、N. Duffield、J. Neville 和 R. Kompella，“图采样与保持：大图分析框架”，载于第20届 ACM SIGKDD国际知识发现与数据挖掘会议论文集。ACM，2014年，第1446–1455页。[65] M. Bressan、S. Leucci 和 A. Panconesi，“Motivo：通过简洁色彩编码和自适应采样快速进行模体计数”，arXiv:1906.01599，2019年。[66] ——，“Motivo：通过简洁色彩编码和自适应采样快速进行模体计数”，VLDB基金会会刊，第12卷，第11期，第1651–1663页，2019年。[67] ——，“通过简洁色彩编码和自适应采样加速模体计数”，ACM数据知识发现汇刊，第15卷，第6期，第96:1–96:27页，2021年。[68] A. J. Walker，“一种生成具有一般分布的离散随机变量的高效方法”，ACM数学软件汇刊，第3卷，第3期，第253–256页，1977年。[69] 加州大学圣迭戈分校，“亚马逊产品评论和元数据”，2023年。 [在线]。网址：https://cseweb.ucsd.edu/~jmcauley/datasets.html# amazon reviews[70] J. McAuley、C. Targett、Q. Shi 和 A. van den Hengel，“基于图像的风格与替代品推荐”，载于第38届国际 ACM SIGIR信息检索研究与发展会议论文集，2015年，第43–52页。[71] R. He 和 J. McAuley，“潮起潮落：利用单类协同过滤对时尚趋势的视觉演变进行建模”，载于第25届国际万维网大会论文集，2016年，第507–517页。[72] “IMDB二部网络”，2006年，Network Repository。 [在线]。网址：https://networkrepository.com/IMDB.php [73] “Flickr用户—群组网络”，Network Repository。 [在线]。网址：https://networkrepository.com/flickr.php[74] “LAW：大规模网络数据”，LAW，米兰大学。 [在线]。网址：http://law.di.unimi.it/ [75] P. Boldi 和 S. Vigna，“WebGraph框架 I：压缩技术”，载于第13届国际万维网大会（WWW）论文集，2004年，第595–602页。[76] P. Boldi、M. Rosa、M. Santini 和 S. Vigna，“分层标签传播：一种用于压缩社交网络的多分辨率无坐标排序方法”，载于第20届国际万维网大会（WWW）论文集，2011年，第587–596页。[77] “社交网络数据（WOSN、Facebook等）”，MPI-SWS。 [在线]。网址：http://socialnetworks.mpi-sws.org/ data-wosn2009.html[78] B. Viswanath、A. Mislove、M. Cha 和 K. P. Gummadi，“Facebook中用户交互的演变”，载于第2届 ACM在线社交网络研讨会（WOSN）论文集，2009年，第37–42页。[79] J. Leskovec 和 A. Krevl，“SNAP数据集：斯坦福大学大型网络数据集集合”，2014年6月。 [在线]。网址：https://snap.stanford.edu/data [80] R. A. Rossi 和 N. K. Ahmed，“带有交互式图分析与可视化功能的网络数据存储库”，载于第二十九届 AAAI人工智能会议论文集，2015年，第4292–4293页。 [在线]。网址：https://networkrepository.com [81] “Orkut链接（IMC 2007）”，MPI-SWS。 [在线]。网址：https://socialnetworks.mpi-sws.org/data-imc2007.html [82] A. Mislove、M. Marcon、K. P. Gummadi、P. Druschel 和 B. Bhattacharjee，“在线社交网络的测量与分析”，载于第7届 ACM SIGCOMM互联网测量会议论文集，2007年，第29–42页。