计 算 机 研 究 与 发 展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330628
61(7): 1643−1669, 2024

基于本地差分隐私的分布式图统计采集算法

傅培旺　丁红发　刘　海　蒋合领　唐明丽　于莹莹
（贵州省高等学校区块链与金融科技重点实验室（贵州财经大学）　贵阳　550025）
（贵州财经大学信息学院　贵阳　550025）
（peiwangfu@foxmail.com）

Statistics Collecting Algorithms of Distributed Graph via Local Differential Privacy

Fu Peiwang, Ding Hongfa, Liu Hai, Jiang Heling, Tang Mingli, and Yu Yingying
（*Key Laboratory of Blockchain and FinTech of Guizhou Provincial Colleges and Universities* (*Guizhou University of Finance and Economics*), *Guiyang* 550025）
（*College of Information*, *Guizhou University of Finance and Economics*, *Guiyang* 550025）

Abstract　Mass distributed graph data, generated by social network, social IoT (Internet of things) and other scenarios, are collected by the service providers and used to provide diverse data oriented services, and they arise serious privacy conserns. In this context, how to achieve secure and effective collecting of distributed graph-structure data with strong correlation has become a bottleneck in large-scale graph-structure data application services. The Node/Edge-LDP (local differential privacy) model of distributed graph-structure data cannot effectively handle the conflict between effectiveness of privacy preserving and the utility of data. To this end, a statistics collecting algorithm named GS-LDP of distributed graph via LDP is proposed. This algorithm can collect three different statistics, i.e., degree distribution, triangle count sequence and clustering coefficient, simultaneously. And it can adapt to different effectiveness and privacy preserving requirements. First, a degree distribution collecting algorithm with Node-LDP is designed. By using the grouping mechanism and symmetric unary coding mechanism, this algorithm can achieve high strength privacy protection. Second, obtaining the threshold value with the algorithm mentioned above, then alleviating the problem of excessive noisy edges in random noise through introducing a prune algorithm, the triangle count sequence collecting algorithms with Node-LDP and Edge-LDP are proposed respectively. Third, based on the aforementioned algorithms, clustering coefficient collecting algorithms with Node-LDP and Edge-LDP are proposed respectively, by introducing the Laplace mechanism. And these algorithms achieve multi-metrics collecting of distributed graph-structure for different privacy and utility requirements. At last, experimental and comparative results show that, the proposed algorithms can strengthen both privacy and data utility, and perform significantly better than existing single or multiple statistics collecting algorithms.

Key words　graph-structure data; local differential privacy; privacy-preserving techniques; data utility; graph statistics; data collection

摘　要　社交网络、社交物联网等应用场景产生的海量分布式图结构数据，被应用服务商采集并以此提供各类以数据为驱动的服务，或将引发严重的隐私风险. 在此背景下，如何针对具备强关联性的分布式图结

收稿日期：2023-07-31；**修回日期**：2023-12-04
基金项目：国家自然科学基金项目 (62002080, 62062017)；贵州省教育厅自然科学研究项目 (黔教技 [2023]065, 黔教技 [2023]014)
This work was supported by the National Natural Science Foundation of China (62002080, 62062017) and the Natural Science Researching Program of D.o.E. of Guizhou (Qian Education and Technology[2023] 065, Qian Education and Technology[2023]014).
通信作者：丁红发（hongfa.ding@foxmail.com）

构数据实现安全高效的采集，成为大规模图结构数据应用服务的瓶颈. 面向分布式图结构数据隐私保护的节点或边本地差分隐私模型无法有效处理隐私保护效果和数据有效性之间的冲突关系. 针对该问题，提出基于本地差分隐私的分布式图统计采集算法，同时实现度分布、三角计数序列和聚类系数3个不同统计指标采集，并适应不同有效性和隐私保护的需求. 首先，采用分组机制及对称一元编码机制，设计具备高强度隐私保护的基于Node-LDP的度分布采集算法；其次，基于所提度分布采集算法获取阈值，引入剪枝算法缓解随机加噪的噪声边过多问题，并分别提出基于Node-LDP和Edge-LDP的三角计数序列采集算法；再次，在前述三角计数序列采集算法基础上引入拉普拉斯机制，从而分别提出基于Node-LDP和Edge-LDP的聚类系数采集算法，进而实现不同保护强度及数据效用需求下的分布式图结构多指标采集；最后，实验和对比结果表明，所提算法能同时提高隐私保护强度和数据效用，比现有单一或多统计指标采集算法更具优势.

关键词 图结构数据；本地差分隐私；隐私保护技术；数据效用；图统计；数据采集

中图法分类号 TP309.2

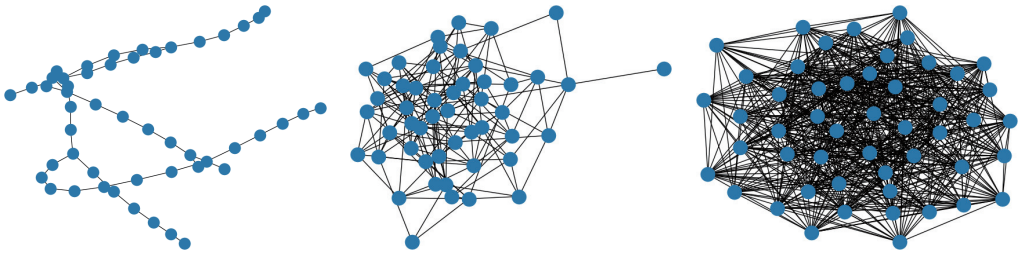
近年来，物联网、大数据及云计算等[1-2]新兴技术的快速发展使得海量数据采集、存储和处理变得便捷，促使社交网络、商业推荐、智慧交通等以数据为驱动的信息服务快速普及，深刻影响人们的生活与工作. 特别是生物医药研发、社交物联网[3-4]、社交网络、金融保险等领域，产生了海量数据并兴起了AI for Science、联邦学习[5]、图神经网络[6]等新兴研究方向.

然而，各类以数据为驱动的信息服务系统均以不同形式采集和保留用户的个人数据. 这些数据包含用户的身份、职业及爱好等隐私信息，还可能包含不同用户间的交互信息. 如果在采集和共享过程中，不能对这些数据进行有效隐私保护，将引发信息泄露及财产损失等隐患[4]. 因此，将隐私数据共享应用于信息服务之前，有必要落实合规的隐私保护措施，以实现隐私数据的安全采集和共享. 在不同类型的数据采集和共享中，社交网络、社交物联网等产生的图结构数据可以表示不同实体间的复杂关系，较关系、文本和图像等类型的数据在分析主体间复杂关联关系方面更具优势. 然而，图结构数据的节点和边表示的身份隐私、关系隐私和属性隐私相互关联，尤其在采集或共享分布式用户持有的图结构数据子集时，具有更大的隐私泄露风险和保护难度. 因此，迫切需要有效的隐私保护手段，在图结构数据采集和共享过程实现高效的隐私保护效果[7].

在分布式图结构数据应用场景中(如社交物联网、社交网络)，往往由单个节点(用户或设备)持有其自身节点信息及其与邻居的关联信息，即一阶邻居子图[7-8]. 为了更好地分析数据或提供服务，数据采集者(服务提供商)会从分布式用户端采集整个图结构数据，以供应用服务商分析和应用，进而提高服务便捷性或精准营销. 具体的数据分析过程会采用统计学、图论等技术对整个图结构数据的统计指标(如度分布、三角计数序列和聚类系数等)进行处理，从而有效地挖掘其中的潜在高价值信息[7-9]. 但这些统计指标与各用户拥有的子图结构数据高度关联，且包含大量用户轨迹、关联关系等隐私信息，在采集、共享或发布时极易泄露隐私信息. 因此，对高度关联的分布式图结构数据的各种统计指标进行隐私保护采集共享，且不泄露用户身份隐私以及用户间的关系隐私，成为以图结构数据为驱动的信息服务领域中尤为迫切的挑战之一.

为了有效保护分布式场景下的图结构数据隐私，本地差分隐私(local differential privacy, LDP)[10-11]从关系型数据集、Key-Value型数据应用场景中被扩展到了图结构数据应用场景. LDP具有可证明强度的隐私保护能力，且无需可信第三方，故而广受学术界和产业界的关注[8-9,12-17]. 面向图结构数据的LDP方案，其基本假设是各用户持有包含自身节点信息的一阶邻居子图，即各用户仅知道其自身信息以及与其相邻的关系信息. 在数据采集者采集数据时，首先，各用户对自身一阶邻居子图的邻接向量进行随机响应(randomized response, RR)[18]扰动加噪；然后，数据收集方根据各用户发送的噪声邻接向量进行聚合统计，获取分布式图结构数据整体统计指标的估计数据. 根据保护强度不同，LDP可应用于对图结构数据的节点或边进行加噪，产生了节点LDP(Node-LDP)和边LDP(Edge-LDP)[8]两种变型. 其中，Node-LDP能同时保护多条边的隐私，保证敌手无法获取图结构数据上任意节点的邻居数量[8,14,17]，从而提供更强的隐私保护效果，但会大幅度降低图结构数据采集结果

的数据效用[8,14]. 因此, 为了保持更好的数据效用, 现有基于 LDP 的图结构数据隐私保护方案[8-9,12-14] 通常采用 Edge-LDP 模式对隐私保护效果进行折中以提高数据效用. 如 Zhan 等人[8] 提出一种 Edge-LDP 下基于度序列进行聚类的图生成算法. Ye 等人[12] 引入 Edge-LDP 提出一种可同时采集聚类系数和模块化信息的框架. 随着图结构数据规模的增长以及图结构数据应用的扩展, 获取一些图结构数据信息时所面临的隐私需求将更加严格, 特别是图结构数据多条边的隐私需要同时保护的需求愈发迫切[17]. 具体而言, 需要在实施隐私保护的过程中进一步加强图结构数据的节点隐私(如度值隐私)和关系隐私保护, 或者实现节点的个性化隐私保护. 然而, Edge-LDP 无法同时保护多条边的隐私, 且无法实现对节点度值隐私的保护[8,14,17]. 为进一步强化图结构数据的隐私保护效果, Fu 等人[17] 和 Liu 等人[19] 分别基于 Node-LDP 设计了分布式图结构数据的隐私保护聚类及度分布采集算法. 真实图结构应用中, 往往需要同时采集多种图结构数据统计指标, 且各用户的隐私偏好不一[14,17], 需要平衡分布式图结构数据上各节点的隐私需求以及数据效用间的冲突, 同时优化隐私保护机制, 尽可能降低扰动加噪对数据效用的影响. 此外, 直接应用现有的算法来组合完成多统计指标的隐私保护采集, 一方面会增加分布式用户与数据采集者间的通信开销, 另一方面会增加存储和计算开销, 再一方面各个算法的隐私保护强度不一, 难以实现统一的隐私保护效果, 且无法满足各节点的个性化隐私保护偏好需求. 同时, 真实世界中的图结构数据多数为稀疏的, 使得无论 Edge-LDP 或 Node-LDP 都会在随机化过程中产生大量的虚假边[14,17], 进而严重降低最终采集的图结构数据的精度和效用. 如图 1 所示, 当采用 2-Edge-LDP 和 2-Node-LDP($\varepsilon = 2$)对图结构数据加噪时, 图结构数据的隐私效用会受到不同程度的影响(边密度大幅度提高).



(a) 原始图 (边密度: 0.041) (b) 满足2-Edge-LDP (边密度: 0.147) (c) 满足2-Node-LDP (边密度: 0.493)

Fig. 1 Impact of perturbing graph-structure data on edge density utility under Edge-LDP and Node-LDP

图 1 Edge-LDP 和 Node-LDP 下扰动图结构数据对边密度效用的影响

为了满足分布式图结构数据应用中的多样化统计指标采集需求, 实现高强度和个性化的隐私保护, 降低扰动加噪对数据效用的影响, 本文通过引入 Node-LDP 和 Edge-LDP 提出基于本地差分隐私的图统计指标采集算法(graph statistics collecting algorithms via local differential privacy, GS-LDP), 将度值采集作为基础统计指标来采集, 在采集过程中通过限制加噪的扰动域来降低噪声输入, 并在加噪过程中对用户子图进行剪枝来大幅度降低无效噪声量来提升数据效用, 通过自适应隐私设置来调节满足 Node-LDP 或 Edge-LDP 以适应用户的个性化隐私偏好. 具体地: 首先, 引入分组机制和对称一元编码(symmetric unary encoding, SUE)[11], 提出一种满足 Node-LDP 的图结构数据度分布采集算法, 通过限制 SUE 扰动域上限的方式降低采集过程中的噪声输入, 实现 Node-LDP 保护的同时实施高效用的分布式图结构数据度分布采集; 其次, 基于所提出的度分布采集算法, 利用剪枝(Prune)算法对用户持有的子图进行剪枝, 通过设置度值阈值来对一阶邻居子图的邻居位向量进行剪切投影, 从而限制随机过程中所产生的噪声边的量, 通过大幅度降低无效噪声提高数据效用; 再次, 引入拉普拉斯机制、随机响应机制和 2 轮交互模型, 提出满足 Node-LDP 和 Edge-LDP 的三角计数序列采集算法及满足 Node-LDP 和 Edge-LDP 的聚类系数采集算法, 通过自适应隐私设置来调节满足 Node-LDP 或 Edge-LDP 以适应用户的个性化隐私偏好, 实现不同隐私保护强度和数据效用需求的分布式图结构数据多统计指标采集. 本文的主要贡献有 4 个方面:

1)基于分组机制及对称一元编码机制, 提出一种满足 Node-LDP 的分布式图结构数据度分布采集算法, 提高度分布采集的隐私保护强度和数据效用;

2)采用剪枝算法控制随机加噪过程中虚假边的上限, 降低无效噪声输入, 并进一步利用 2 轮交互模型分别提出满足 Node-LDP 和 Edge-LDP 的三角计数

序列采集算法;

3)在2)中三角计数序列采集算法的基础上,引入拉普拉斯机制分别提出满足Node-LDP和Edge-LDP的聚类系数采集算法;

4)进行理论分析和多个数据集实验对比,结果表明,所提出的算法能在不同约束下实现多个统计指标的有效采集,与现有的统计指标采集算法相比有显著的效用优势.

1 相关工作

差分隐私(differential privacy, DP)[20-24]由Dwork[20]提出,主要应用于关系型数据库的统计信息隐私保护,后来被逐步应用于图结构数据的隐私保护[21-24].DP在应用过程中,通常需要一个可信第三方对数据进行统筹处理,其仅适用于中心化的数据共享采集场景.然而,在诸如位置服务网络、社交物联网等真实图结构数据应用场景中,往往不存在可信第三方[12],且数据被分布式持有在不同的用户或设备上.在此类场景中,各用户相互不信任,均不愿把自己的子图数据直接共享给彼此或数据采集者.

为了适应分布式场景的隐私保护,LDP[10-11]被提出并应用于均值估计[10]及频繁模式挖掘[10]等.直至2017年,Zhan等人[8]提出将LDP应用于图结构数据隐私保护,通过在分布式端迭代收集用户节点度值并采用*k*-means算法聚合获取生成图,但由于该方案仅通过度值进行聚类,所得到的生成图缺少大量原始图中的结构特性.为了提高数据效用,Ye等人[12]通过引入RR机制[18]扰动用户的子图邻接向量,并采用拉普拉斯机制采集用户度值,提出一种基于LDP的图聚类系数和模块化采集框架.而后,Imola等人[9]引入2轮交互模型,将加噪后的子图邻接向量合成噪声图后发送给用户,由各用户分别估计三角计数后将整体三角计数结果发布,并最终由数据采集者统一采集全局计数信息.然而,Ye等人[12]和Imola等人[9]的方法都会产生大量噪声边,从而影响最终采集结果的精度.为了解决该问题,Hou等人[14]通过设计一种基于最优长度的度向量编码模型对图结构数据进行分散聚类,但该方法仅适用于图结构数据聚类.为了获得更好的数据效用,Sun等人[25]和Liu等人[26]通过减弱分布式图结构数据的隐私保护假设,允许各用户和邻居交换各自的邻居信息(即各用户持有的二阶子图信息),分别设计了隐私保护的分布式图结构数据子图计数算法.此外,针对二元属性图结构数据的隐私保护,Wei等人[16]提出一种随机跳转(random jump, RJ)的分布式属性图结构数据度分布采集算法,并将RR机制应用于采集用户的二元属性,在此基础上,通过抽样方法构建整体的二元属性图结构数据用以数据分析.

前述图结构数据的LDP模型均基于Edge-LDP.然而,随着数据规模的增长及隐私保护需求的增强,Edge-LDP的隐私保护强度难以满足同时保护多条边的隐私需求[14,17].因此,为了提高对图结构数据的隐私保护水平,Fu等人[17]提出基于Node-LDP的2阶段图结构数据聚类框架,通过剪影系数测量模型进行节点聚合,并利用自适应加噪来提高聚类结果的效用.Liu等人[19]结合Node-LDP和密码投影方法提出一种最优图投影参数选择方法,并基于此提出Node-LDP下的度分布采集算法,但度值的敏感度过大,使得度分布采集结果效用不高.

由此可见,现有的分布式图结构数据隐私保护方案大多基于Edge-LDP设计,其通过放宽对边的数量保护的隐私约束,实现高效用的数据采集[14,17].也正因如此,此类方法无法实现同时保护多条边的隐私信息[14,17].而Node- LDP的方案能够保证无法从算法输出中获取任何用户的朋友数(边数)[14],可有效避免此类缺陷.因而,Node-LDP较Edge-LDP的方案具有更强的隐私保护效果,但严格的隐私约束及敏感度会大幅度削弱其实际效用[8,14].同时,2种类型的隐私保护方法在作用于分布式图结构数据时,均直接采用随机响应扰动各用户的子图邻接向量,从而会产生大量噪声边(如图1所示),尚未发现有效的机制来识别无效噪声并降低这些噪声的添加量,这使得无论Edge-LDP还是Node-LDP采集的图结构数据效用均无法达到理想状态.此外,常规LDP模型[9,12,25-26]多数仅针对单个统计指标的采集需求,无法同时适应图结构数据不同统计指标的采集.若通过简单组合此类算法来实现多统计指标采集,会造成通信复杂度、存储复杂度、计算复杂度提高等问题,且难以使不同算法的隐私保护水平一致.

为此,本文结合Node-LDP和Edge-LDP面向分布式图统计采集隐私保护需求,引入分组SUE、RR及拉普拉斯机制,并提出一种优化图结构噪声量的剪枝算法,进而设计具备强隐私保护能力和高数据效用的多样化图统计指标采集算法.首先,针对性质单一的度分布指标,通过分组报告的方式限制SUE扰动域上限,降低采集过程的噪声输入,从而实现满足Node-LDP的高效度分布采集;其次,针对三角计

数序列，提出剪枝算法，用于优化随机化过程中边缘的噪声量；再次，引入 2 轮交互模型、RR 及拉普拉斯机制，通过设置不同隐私参数提出满足 Node-LDP 和 Edge-LDP 的三角计数序列和聚类系数采集算法.

2 基本定义及预备知识

2.1 符号定义

本文所述算法均基于简单无向图$G=(V,E)$，其中顶点集由V表示，边集由E表示，顶点集中的顶点数$n=|V|$代表当前图G中的总用户数量. 本文假设图G是分布式的，各用户在本地持有以自身为中心的星图，可称为一跳邻居子图. 用户可将其抽象为一个n维的邻接向量$\boldsymbol{B}_i=(b_1,b_2,\cdots,b_n)$，其中任意$b_j\in\{0,1\}$为$\boldsymbol{B}_i$中的第$j$位，代表用户$i$与用户$j$之间的关联关系，$b_j=1$表示用户$i$与用户$j$之间存在边，$b_j=0$表示用户$i$与用户$j$之间不存在边. $\boldsymbol{M}=(\boldsymbol{B}_1,\boldsymbol{B}_2,\ldots,\boldsymbol{B}_n)=(B_{ij})_{n\times n}$为图$G$的邻接矩阵，由$G$中所有用户的邻接向量组成，代表全局用户的邻居信息. $d_i=\sum_{j=1}^{n}B_{ij}$（或$d_i=\sum_{j=1}^{n}b_j$）为用户i的度值，代表用户i的邻居数量. T_i代表用户i所关联的三角计数，即用户i的局部三角计数. CC_i代表用户i的局部聚类系数，可通过用户局部三角计数和度值进行计算，即

$$CC_i=\frac{2T_i}{d_i(d_i-1)},$$

其中$T=(T_1,T_2,\cdots,T_n)$和$CC=(CC_1,CC_2,\cdots,CC_n)$分别代表图$G$的三角计数序列及全局聚类系数. ε为差分隐私的隐私预算，用于控制噪声输入及保护强度. ε越小，保护强度越高，噪声量越大.

为了便于阅读，表 1 中列出本文常用符号及其描述.

Table 1 Symbols and Its Discription

表 1 符号及其描述

符号	描述
ε	隐私预算
G	简单无向图
d_i	用户度值
$\boldsymbol{B}_i$	用户一阶邻接向量
T_i	用户局部三角计数
CC_i	用户局部聚类系数
$n=\|V\|$	图G中总用户数量
T	图G中的三角计数序列
CC	图G的全局聚类系数
$\boldsymbol{M}$	图G的邻接矩阵

2.2 本地差分隐私及加噪机制

定义 1. ε-LDP[10]. 给定n个用户和随机算法$\mathcal{A}$，若算法$\mathcal{A}$作用在任意 2 条记录t和t'上得到相同的输出结果t^*，有

$$\frac{Pr[\mathcal{A}(t)=t^*]}{Pr[\mathcal{A}(t')=t^*]}\leqslant e^{\varepsilon},$$

则算法$\mathcal{A}$满足ε-LDP. 其中，$t^*\in range(\mathcal{A})$为算法$\mathcal{A}$的输出.

定义 2. ε-Edge-LDP[8]. 当随机算法$\mathcal{A}$作用在任意 2 个仅在 1 位上不同的邻接向量$\boldsymbol{B}_i=(0,1)^n$和$\boldsymbol{B}_i'=(0,1)^n$上时，若算法$\mathcal{A}$满足$\varepsilon$-Edge-LDP，则有

$$\frac{Pr[\mathcal{A}(\boldsymbol{B}_i)=\boldsymbol{S}]}{Pr\left[\mathcal{A}\left(\boldsymbol{B}_i'\right)=\boldsymbol{S}\right]}\leqslant e^{\varepsilon},$$

其中$\boldsymbol{S}=(0,1)^n\in range(\mathcal{A})$，即算法$\mathcal{A}$的输出.

定义 3. ε-Node-LDP[8]. 当随机算法$\mathcal{A}$作用在任意 2 个邻接向量$\boldsymbol{B}_i=(0,1)^n$和$\boldsymbol{B}_i'=(0,1)^n$上时，若算法$\mathcal{A}$满足$\varepsilon$-Node-LDP，则有

$$\frac{Pr[\mathcal{A}(\boldsymbol{B}_i)=\boldsymbol{S}]}{Pr\left[\mathcal{A}\left(\boldsymbol{B}_i'\right)=\boldsymbol{S}\right]}\leqslant e^{\varepsilon},$$

其中$\boldsymbol{S}=(0,1)^n\in range(\mathcal{A})$，即算法$\mathcal{A}$的输出.

在分布式图结构数据应用场景下，ε-LDP 仅满足采集数据的隐私性，并不考虑数据间的关联性，如顶点之间关联的边. 在此基础之上，ε-Edge-LDP 和ε-Node-LDP 进一步考虑数据间的关联性，需要在采集过程中保护用户间的边信息（关联信息）；不同的是，ε-Edge-LDP 是ε-Node-LDP 的松弛，它将任意 2 个邻接向量的邻接定义限制在 1 位（即 1 条边）[12]. 在此背景下，设Q_1,Q_2,Q_3和U_1,U_2,U_3分别代表满足ε-Node-LDP,ε-Edge-LDP 和ε-LDP 时的隐私保护强度和效用精度. 根据上述描述可得出结论，在分布式图结构数据应用场景下，隐私约束性（即隐私保护强度）上

$$Q_1>Q_2>Q_3;$$

反之，在最终效用精度上

$$U_1<U_2<U_3.$$

定义 4. ε-RR[18]. 设算法$\mathcal{RR}$为ε-RR，用户通过算法$\mathcal{RR}$扰动数据的过程为：对任意二进制数据$x\in\{0,1\}$，用户以p的概率发送真实值x，以q的概率发送真实值$1-x$，即

$$Pr\left[\mathcal{RR}(x)=y\right]=\begin{cases}p=\dfrac{e^{\varepsilon}}{e^{\varepsilon}+1}, & \text{if } y=x,\\ q=\dfrac{1}{e^{\varepsilon}+1}, & \text{其他},\end{cases}$$

满足以上条件，则ε-RR 满足ε-LDP，其中$x,y\in\{0,1\}$分别为输入数据和输出数据.

定义 5. ε-SUE[11]. 用户通过ε-SUE 扰动数据的过

程为：对任意o维的实值数据$x\in\{1,2,\cdots,o\}$. 首先，用户将其编码为一个o维的二元向量，即$Encode(x)=(0,\cdots,0,1,0,\cdots,0)$，其中除第$x$位为1外，其余位全为0. 随后，用户对编码数据进行逐位扰动($\varepsilon/2$-RR)，即

$$Pr\left[Encode'(x)_i=1\right]=\begin{cases}p, \text{ if } Encode(x)_i=1,\\ q, \text{ if } Encode(x)_i=0,\end{cases}$$

该过程满足ε-LDP. 其中，$p=\dfrac{e^{\varepsilon/2}}{e^{\varepsilon/2}+1}$，$q=1-p$.

定义 6. 全局敏感度(global sensitivity, GS)[27]. 对于任意一个实际值的查询函数$f:D\to\mathbb{R}$，其全局敏感度定义为

$$GS_f=\max_{D,D'}\|f(D)-f(D')\|_1,$$

其中D和D'表示为相邻数据集.

定义 7. Laplace机制[24]. 对于给定的数据集D和实值查询函数f，GS_f是f在数据集D上的全局敏感度，则随机算法$\mathcal{M}$：$\mathcal{M}(D)=f(D)+X$满足ε-DP. 其中$X\sim Lap\left(GS_f/\varepsilon\right)$是服从尺度参数为$GS_f/\varepsilon$的Laplace分布噪声.

值得注意的是，在LDP视角上，相邻数据集D和D'被转换成了相邻的邻接向量$\boldsymbol{B}_i$和$\boldsymbol{B}'_i$. 例如，在Edge-LDP场景下采集度值d_i，删除任意一条边仅影响$\boldsymbol{B}_i$中的1个比特位，因此敏感度为1，即向d_i中添加$Lap(1/\varepsilon)$的噪声便可满足ε-Edge-LDP.

定理 1. 若通过ε-RR扰动一个n维的邻接向量$\boldsymbol{B}_i=(0,1)^n$，其满足ε-Edge-LDP[8].

证明. 设随机算法$\mathcal{A}$为ε-RR，$Pr\left[\mathcal{A}(x)=y\right]$表示$x\in\{0,1\}$随机翻转为$y\in\{0,1\}$的概率. 其中，$\boldsymbol{B}_i=(b_1,b_2,\cdots,b_n)$和$\boldsymbol{B}'_i=\left(b'_1,b'_2,\cdots,b'_n\right)$($b_i,b'_i\in\{0,1\}$)为仅相差1位的邻接向量. 不失一般性，假设$b_1\neq b'_1$，给定算法$\mathcal{A}$的任意输出$\boldsymbol{S}=(s_1,s_2,\cdots,s_n)$，$s_i\in\{0,1\}$，有

$$\frac{Pr[\mathcal{A}(\boldsymbol{B}_i)=\boldsymbol{S}]}{Pr\left[\mathcal{A}\left(\boldsymbol{B}'_i\right)=\boldsymbol{S}\right]}=\frac{Pr[\mathcal{A}(b_1)=s_1]Pr[\mathcal{A}(b_2)=s_2]\cdots Pr[\mathcal{A}(b_n)=s_n]}{Pr\left[\mathcal{A}\left(b'_1\right)=s_1\right]Pr\left[\mathcal{A}\left(b'_2\right)=s_2\right]\cdots Pr\left[\mathcal{A}\left(b'_n\right)=s_n\right]}=\frac{Pr[\mathcal{A}(b_1)=s_1]}{Pr\left[\mathcal{A}\left(b'_1\right)=s_1\right]}\leqslant\frac{p}{q}=e^{\varepsilon}.$$

因此，算法$\mathcal{A}$(ε-RR)满足ε-Edge-LDP. 证毕.

定理 2. 若通过ε-RR扰动一个n维的邻接向量$\boldsymbol{B}_i=(0,1)^n$，其满足$n\varepsilon$-Node-LDP.

证明. 设随机算法$\mathcal{A}$为ε-RR，$Pr\left[\mathcal{A}(x)=y\right]$表示$x\in\{0,1\}$随机翻转为$y\in\{0,1\}$的概率. 其中，$\boldsymbol{B}_i=(b_1,b_2,\cdots,b_n)$和$\boldsymbol{B}'_i=\left(b'_1,b'_2,\cdots,b'_n\right)$($b_i,b'_i\in\{0,1\}$)为2个相邻的邻接向量. 不失一般性，假设$\boldsymbol{B}_i,\boldsymbol{B}'_i$中每一位对应值都不一样，因此给定算法$\mathcal{A}$的任意输出$\boldsymbol{S}=(s_1,s_2,\cdots,s_n)$，$s_i\in\{0,1\}$，则有

$$\frac{Pr[\mathcal{A}(\boldsymbol{B}_i)=\boldsymbol{S}]}{Pr[\mathcal{A}(\boldsymbol{B}'_i)=\boldsymbol{S}]}=\frac{Pr[\mathcal{A}(b_1)=s_1]Pr[\mathcal{A}(b_2)=s_2]\cdots Pr[\mathcal{A}(b_n)=s_n]}{Pr\left[\mathcal{A}\left(b'_1\right)=s_1\right]Pr\left[\mathcal{A}\left(b'_2\right)=s_2\right]\cdots Pr\left[\mathcal{A}\left(b'_n\right)=s_n\right]}\leqslant\frac{p^n}{q^n}=e^{n\varepsilon}.$$

因此，算法$\mathcal{A}$(ε-RR)满足$n\varepsilon$-Node-LDP. 证毕.

根据定理1和定理2可知，在相同ε情况下对n维的邻接向量进行扰动时，满足ε-Node-LDP和ε-Edge-LDP分别需要通过ε/n-RR和ε-RR进行扰动. 显然，相同隐私预算的情况下，满足Node-LDP时所需注入的噪声量要远高于Edge-LDP. 同理，安全级别上Node-LDP也要远优于Edge-LDP.

2.3 DP的组合性质

LDP与DP的区别在于LDP将隐私保护处理过程从数据采集者端移至用户端，二者的核心仍旧是通过噪声机制将数据扰动加噪，从而达到隐私保护效果. 同时，LDP的后处理过程并不会影响最终的隐私保证，因此LDP也满足DP的组合性质[28].

引理 1. 序列组合性[29]. 给定任意n个随机算法$\{\mathcal{A}_i\}_{1\leqslant i\leqslant n}$. 其中，任意算法$\mathcal{A}_i$满足$\varepsilon_i$-DP，则按指定顺序组合后的算法$\{\mathcal{A}_i\}_{1\leqslant i\leqslant n}$满足$\sum_{i=1}^{n}\varepsilon_i$-DP.

引理 2. 并行组合性[29]. 给定任意n个随机算法$\{\mathcal{A}_i\}_{1\leqslant i\leqslant n}$. 其中，任意算法$\mathcal{A}_i$满足$\varepsilon_i$-DP，同时$n$个算法所作用的数据集不相交，则由上述算法所构成的算法$(\mathcal{A}_1(D_1),\mathcal{A}_2(D_2),\cdots,\mathcal{A}_n(D_n))$满足$\max_{1\leqslant i\leqslant n}\{\varepsilon_i\}$-DP.

3 基于LDP的图统计采集框架

本节阐述基于LDP的图结构数据统计采集(GS-LDP)框架. 分布式图结构数据应用场景下，所有用户仅持有包含自身节点的一阶邻居子图，而非拥有整个图结构数据集，数据采集者(服务提供者)希望能采集到整个图结构数据的度分布、三角计数序列和聚类系数等多个统计指标；为了避免自身的身份隐私和与他人间的关系隐私遭受泄露，用户不愿将自身和他人的原始数据直接共享发布给数据采集者. 故而在数据采集的过程中需要考虑用户的不同隐私需求和数据采集者采集不同统计指标时的数据效用需求.

为了同时实现分布式图结构数据的多个统计指标隐私保护采集，结合Node-LDP和Edge-LDP设计GS-LDP算法，具体构建框架如图2所示.

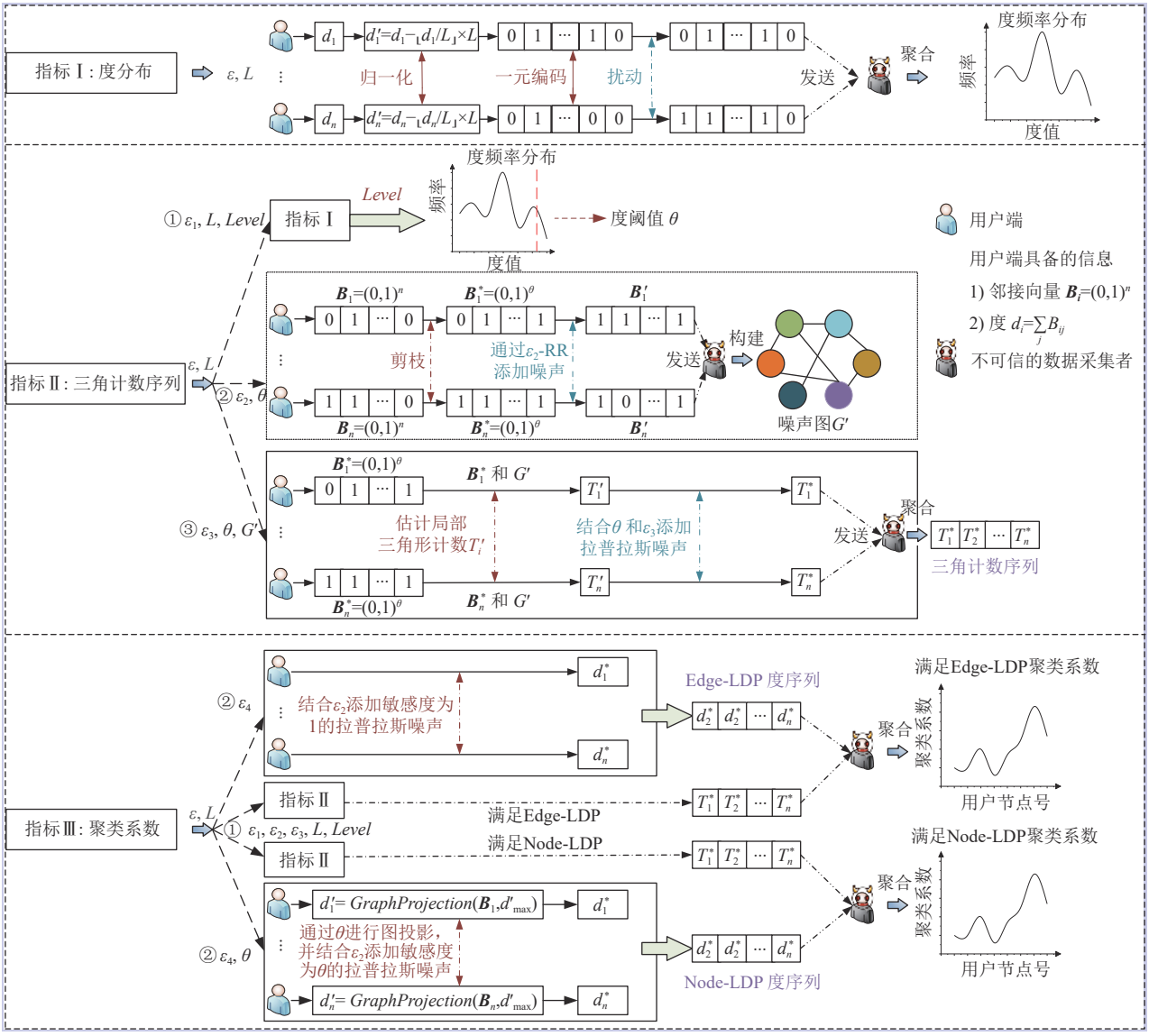


Fig. 2 The framework of statistical indicators collecting of graph-structure data via LDP

图 2 基于 LDP 的图结构数据统计指标采集框架

图 2 中，所提 GS-LDP 算法主要实现 3 个统计指标的隐私保护采集，即度分布、三角计数序列和聚类系数. 这 3 个统计指标差异性较大，其中度分布相对单一，三角计数序列和聚类系数具有复杂结构性. 因此，针对不同的统计指标采集需求，分别设计了基于 Node-LDP 的度分布、基于 Node-LDP 的三角计数序列和聚类系数采集，以及基于 Edge-LDP 的三角计数序列和聚类系数采集算法. 在 GS-LDP 算法中，三角计数序列和聚类系数 2 个统计指标的采集依赖于度分布设置阈值，同时聚类系数需要基于三角计数结果进行估计.

1）基于 Node-LDP 的度分布采集算法. 由于各用户持有子图定点的度值属于各用户可直接计算的数值类数据，采用 Edge-LDP 采集此类数据将无法有效保护关系隐私. 因此，引入分组机制和 SUE 机制设计基于 Node-LDP 的度分布采集算法. 具体如图 2 中指标Ⅰ所示，以单个用户与数据采集者的交互为例，过程为：首先，数据采集者将初始隐私预算 ε 和组距 L 发布给用户；其次，用户 i 收到 ε 和 L 后，根据自身度值 d_i 和 L 计算组号 g_i（$g_i=\lfloor d_i/L \rfloor$）；再次，用户 i 根据 g_i 将 d_i 归一化（d_i-g_iL），并通过一元编码将归一化的 d_i 编码成一个维度为 L 的二元度向量；然后，用户 i 通过对称一元编码扰动机制对度向量进行逐位扰动，并将扰动结果发送给数据采集者；最后，数据采集者分组聚合估计各度值的频率，得到满足 Node-LDP 的度分布结果. 具体算法和证明过程见第 4 节.

2）基于 Node-LDP 和 Edge-LDP 的三角计数序列采集算法. 该算法过程如图 2 的统计指标Ⅱ所示，具体包含一个参数设置阶段和一个 2 轮交互过程. 在参

数设置阶段，首先，数据采集者将初始隐私预算ε拆分为ε_1，ε_2，ε_3，并将ε_1、ε_2、ε_3、组距L和频率上限$Level$发布给每位用户；其次，用户和数据采集者以ε_1和L为参数按照基于 Node-LDP 的度分布采集算法过程获得图结构数据的度分布估计，并将度频率和达到$Level$时的度值设为度阈值θ(即$\sum_{j=0}^{\theta} f'_j \geqslant Level$，其中$f'_j$表示度值为$j$的频率估计)．在第 1 轮交互中，首先，用户通过$\theta$和 Prune 算法将自身一阶邻居子图的邻接向量$\boldsymbol{B}_i$进行裁剪；其次，用户通过$\varepsilon_2$-RR 机制对裁剪后的邻接向量$\boldsymbol{B}'_i$进行扰动，将扰动结果$\tilde{\boldsymbol{B}}_i$发送给数据采集者；然后，数据采集者基于所有用户发送的扰动邻接向量$\tilde{\boldsymbol{B}}_i$构建噪声图G'，并将G'发送给各用户．在第 2 轮交互中，首先，用户根据$\boldsymbol{B}'_i$和G'对自身局部三角计数T_i进行估计；其次，用户通过ε_3-拉普拉斯机制将估计值扰动后发送给数据采集者；然后，数据采集者对所有用户的估计值进行后处理，获取每个用户的局部三角计数估计T'_i，进而采集满足 Node-LDP 或 Edge-LDP 的三角计数序列估计$T'=\left(T'_1,T'_2,\cdots,T'_n\right)$．

在实际中，根据采集三角计数序列的隐私或效用需求，设置不同的ε_2应用于 RR 及不同的敏感度应用于拉普拉斯机制，可使得算法满足 Node-LDP 或 Edge-LDP．具体算法和证明见第 5 节．

3)基于 Node-LDP 和 Edge-LDP 的聚类系数采集算法．算法过程如图 2 的统计指标Ⅲ所示，包含 3 个步骤．首先，数据采集者将初始隐私预算ε拆分为ε_1，ε_2，ε_3，ε_4，并将ε_1、ε_2、ε_3、ε_4、组距L和频率上限$Level$发布给每位用户；其次，用户以ε_1，ε_2，ε_3为隐私参数向数据采集者报告三角计数序列估计$T'=\left(T'_1,T'_2,\cdots,T'_n\right)$，并以$\varepsilon_4$为参数通过拉普拉斯机制向数据采集者报告度值序列$\tilde{D}=\left(d'_1,d'_2,\cdots,d'_n\right)$；最后，数据采集者根据所获得的$T'=\left(T'_1,T'_2,\cdots,T'_n\right)$和$\tilde{D}=\left(d'_1,d'_2,\cdots,d'_n\right)$计算全局聚类系数估计$CC'=\left(CC'_1,CC'_2,\cdots,CC'_n\right)$．

类似地，在实际中，可以根据采集分布式图结构数据聚类系数时的隐私或效用需求，设置不同的加噪参数，使算法满足 Node-LDP 或 Edge-LDP．具体算法和证明见第 6 节．

需要注意的是，当算法满足 Node-LDP 时，各分布式用户度值的敏感度为整体图结构数据的最大度值$d_{\max}$．在真实分布式场景下：一方面，用户无法在分布式端直接获取全局图结构数据的最大度值$d_{\max}$；另一方面，真实场景图结构数据往往是稀疏的，敏感度为$d_{\max}$的拉普拉斯加噪将使噪声过大．因此，在所设计的算法中，仍采用θ为图结构数据的度阈值，并通过投影技术采集各分布式用户的度值，降低敏感度．

4 基于 Node-LDP 的度分布采集算法

面向分布式图结构数据的统计指标采集过程中，现有基于 LDP 或 Edge-LDP 的度分布采集算法隐私保护效果不佳；同时，现有的 Node-LDP 度分布采集算法往往采用拉普拉斯机制加噪[19]，其加噪敏感度大且投影阈值获取困难．本节设计基于 Node-LDP 的分布式图结构数据度分布采集算法来提高隐私保护强度，同时引入分组扰动机制和对称一元编码来提高数据效用．首先，详细介绍所提算法；其次，给出算法的分析与证明．

4.1 分布式图结构数据的度采集算法

本节给出具体的基于 Node-LDP 的分布式图结构数据度分布采集算法，即图 2 中的指标Ⅰ，具体过程如算法 1 所示．

在算法 1 中，具体分为 5 个步骤．首先，数据采集者向用户发送度频率分布查询请求，并设置用户隐私预算ε和组距L．其次，用户依次计算自身组号g_i，归一化度值(d_i-g_iL)，并将归一化后的度值编码为一个L维的二元度向量$\boldsymbol{D}_i$(行②③)．再次，用户通过$\varepsilon/2$-RR 逐位扰动$\boldsymbol{D}_i$，并将扰动后的度向量$\boldsymbol{D}'_i$和g_i发送给数据采集者(行④⑤)．然后，数据采集者获取所有用户数据后，计算总的分组数量和各组用户数量，并按分组依次校正各度值频率(行⑦~⑫)．例如，以度值i为例，设$v=\lfloor i/L\rfloor$和n_v分别为其所处组号及当前组用户数量，$j=i-\lfloor i/L\rfloor$为归一化后的度值．数据采集者计算第v组中所有噪声度向量中第j位为 1 的个数，并估计频率：

$$f'_i=\frac{\sum_{c=0}^{n_v} D'_c[j]-n_vq}{n(p-q)},$$

其中$\sum_{c=0}^{n_v} D'_c[j]$为第v组中所有噪声度向量中第j位为 1 的个数．最后，数据采集者可聚合采集所有度值的频率估计$F'=\left(f'_0,f'_1,\cdots,f'_{\max L-1}\right)$(行⑬)．

算法 1. 基于 Node-LDP 的度分布算法 NLDP-DD．

输入：隐私预算ε，组距L；

输出：度频率分布估计F'．

/*用户*/

① for i in $\{1,2,\cdots,n\}$ do

② $g_i=\lfloor d_i/L\rfloor$; /*用户i的组号*/

③ $\boldsymbol{D}_i=Unary-Encoding(d_i-g_iL)$;/*对用户$i$的度值进行归一化，并将其编码为$L$维二元向量 */

④ 通过$\varepsilon/2$-RR扰动$\boldsymbol{D}_i$中的每一位

$$D_i'[j]=\begin{cases}D_i[j], & p=\dfrac{\mathrm{e}^{\varepsilon/2}}{\mathrm{e}^{\varepsilon/2}+1},\\ 1-D_i[j], & q=\dfrac{1}{\mathrm{e}^{\varepsilon/2}+1};\end{cases}$$

⑤ 将$\boldsymbol{D}_i'$和g_i发送给数据采集者;

⑥ end for

/*数据采集者*/

⑦ for v in $\{0,1,\cdots,\max\{g_1,g_2,\cdots,g_n\}\}$ do

⑧ $n_v=\sum_{i=0}^{n}(g_i=v)$; /*组号为$v$的用户数量*/

⑨ for j in $\{0,1,\cdots,L-1\}$ do

⑩ $f'_{j+vL}=\dfrac{\sum_{c=0}^{n_v}D_c'[j]-n_vq}{n(p-q)}$; /*度值为$j+vL$的频率估计*/

⑪ end for

⑫ end for

⑬ $F'=\left(f'_0,f'_1,\cdots,f'_{\max L-1}\right)$;

⑭ return F'.

在算法1中,采用分组编码扰动的方式控制每个用户度值的报告域值大小,使得算法1在满足Node-LDP的同时采集精度得到提升.此外,算法1还能进一步避免随机化过程中扰动阈值过大的问题.

4.2 度分布采集算法的隐私效用分析

本节对所提基于Node-LDP的分布式图结构数据度分布采集算法的隐私及效用进行分析.证明了算法1满足ε-Node-LDP,采集的度值频率估计f_i'具有无偏性,以及f_i'所满足的方差.

定理3. 算法1满足ε-Node-LDP.

证明. 设$\boldsymbol{D}_i$和$\boldsymbol{D}_j$为任意组中的2个L维二元度向量,同时设算法$\mathcal{R}$为算法1中各用户扰动度向量时的随机算法.易知,$\boldsymbol{D}_i$和$\boldsymbol{D}_j$中元素最多有2位不同,其余位上的元素均为0.设$\boldsymbol{S}\subseteq range(\mathcal{R})$($\boldsymbol{S}=(s_1,s_2,\cdots,s_L)$,$s_i\in\{0,1\}$)为算法$\mathcal{R}$的随机输出.不失一般性,设$\boldsymbol{D}_i$和$\boldsymbol{D}_j$仅前2位不同,则有

$$\frac{Pr[\mathcal{R}(\boldsymbol{D}_i)=\boldsymbol{S}]}{Pr\left[\mathcal{R}\left(\boldsymbol{D}_j\right)=\boldsymbol{S}\right]}=$$

$$\frac{Pr[\mathcal{R}(b_1)=s_1]Pr[\mathcal{R}(b_2)=s_2]\cdots Pr[\mathcal{R}(b_L)=s_L]}{Pr\left[\mathcal{R}\left(b_1'\right)=s_1\right]Pr\left[\mathcal{R}\left(b_2'\right)=s_2\right]\cdots Pr\left[\mathcal{R}\left(b_L'\right)=s_L\right]}=$$

$$\frac{Pr[\mathcal{R}(b_1)=s_1]Pr[\mathcal{R}(b_2)=s_2]}{Pr\left[\mathcal{R}\left(b_1'\right)=s_1\right]Pr\left[\mathcal{R}\left(b_2'\right)=s_2\right]}\leqslant\frac{p^2}{q^2}=\mathrm{e}^{2\frac{\varepsilon}{2}}=\mathrm{e}^{\varepsilon},$$

其中$p=\dfrac{\mathrm{e}^{\varepsilon/2}}{\mathrm{e}^{\varepsilon/2}+1}$,$q=1-p$.因此,任意分布式用户报告度向量的随机化过程均满足ε-Node-LDP.

进一步,根据引理2,算法1整体满足ε-Node-LDP. 证毕.

定理4. 算法1所估计的度值频率f_i'是真实频率f_i的无偏估计.

证明. 设$z=\sum_{c=0}^{n_v}D_c'[j]$,即第$v$组中各用户度向量第$j$位为1的数量,则$f_i'=\dfrac{z-qn_v}{n(p-q)}$,有

$$E\left(f_i'\right)=E\left[\frac{z-qn_v}{n(p-q)}\right]=\frac{1}{n(p-q)}\left[E(z)-qn_v\right].$$

设f_i为真实度值为i的频率,$w=nf_i$为对应的频数,则有

$$E(z)=wp+(n_v-w)q=wp-wq+n_vq,$$

$$E\left(f_i'\right)=\frac{1}{n(p-q)}\left[E(z)-qn_v\right]=$$

$$\frac{1}{n(p-q)}(wp-wq+qn_v-qn_v)=\frac{w}{n}=f_i.$$

因此,度值频率f_i'是真实频率f_i的无偏估计. 证毕.

定理5. 算法1中估计的任意度值频率f_i'的方差满足

$$Var\left(f_i'\right)=\frac{n_vq(1-q)}{n^2(p-q)^2},$$

其中n_v为度值i所处组v的用户数量,$p=\dfrac{\mathrm{e}^{\varepsilon/2}}{\mathrm{e}^{\varepsilon/2}+1}$,$q=1-p$.

证明. 设$w=nf_i$为对应度值i的频数,即度值为i的用户数量,同时$z=\sum_{c=0}^{n_v}D_c'[j]$,则有

$$Var\left(f_i'\right)=Var\left[\frac{z-qn_v}{n(p-q)}\right]=\frac{1}{n^2(p-q)^2}Var(z)=$$

$$\frac{wp(1-p)+(n_v-w)q(1-q)}{n^2(p-q)^2}=$$

$$\frac{n_vq(1-q)}{n^2(p-q)^2}+\frac{w(1-p-q)}{n^2(p-q)}.$$

由于$w=nf_i$,则

$$Var\left(f_i'\right)=\frac{n_vq(1-q)}{n^2(p-q)^2}+\frac{f_i(1-p-q)}{n(p-q)}.$$

同时,$p+q=1$.因此,f_i'的方差为

$$Var\left(f_i'\right)=\frac{n_vq(1-q)}{n^2(p-q)^2}.$$

证毕.

在算法1中,每个用户需要先计算自身组号g_i以及将度值d_i编码为L维的二元向量,并对其二元度向量中的每一位进行RR扰动,因此用户的时间复杂度和空间复杂度分别为$O(2L+1)$和$O(L+1)$.对采集者而言,他们需要将每个用户的噪声度向量和组号收集,计算各组用户数量,并依次校正各度值的频率分布,

采集者的时间复杂度和空间复杂度均为$O(n(L+1))$.

5 三角计数序列采集算法

在真实分布式图结构数据应用场景下，隐私保护和数据效用需求多样化，分布式用户无法直接获取分布式图结构数据的三角计数序列等与图结构特征密切相关的统计指标，故此类信息采集过程中的用户隐私保护非常具有挑战性. 现有的算法在本地扰动加噪时，还存在添加大量噪声边导致数据效用不高的问题. 本节针对上述问题，首先提出一种 Prune 算法，用于控制噪声边添加的上限，进而降低实际噪声图中的密集程度；其次，针对隐私及效用需求，基于 Prune 算法和 2 轮交互模型分别提出基于 Node-LDP 的三角计数序列采集算法及基于 Edge-LDP 的三角序列采集算法；最后，给出三角计数序列采集算法的分析与证明.

5.1 基于 Node-LDP 的三角计数序列采集算法

诸如社交网络、物联网等真实场景中产生的图结构数据通常是稀疏图，即$n \gg d_{\max} \gg d_{\text{ave}}$. 其中，$d_{\max}$和$d_{\text{ave}}$分别为最大度和平均度. 为了应对 LDP 随机加噪过程中产生过多虚假边，降低采集统计指标的效用问题，本节提出 Prune 算法，通过将用户邻接向量缩减成一个θ维的二元向量，从而减少随机过程中产生的噪声边，具体过程如算法 2 所示.

算法 2 中包含 2 个步骤：1)在用户持有度值上限θ后，用户可依据θ和邻居真实情况裁剪自己的邻接向量. 以用户i为例，若$d_i > \theta$，则用户i随机抽取θ个邻居组成裁剪后的邻接向量$\boldsymbol{B}_i'$，每一位邻居均为 1(行①②). 相反，若$d_i \leqslant \theta$，则用户将d_i个邻居组成裁剪后邻接向量的前d_i位，并随机抽取$\theta - d_i$个非邻居充当邻接向量的后几位，从而组成一个θ维的二元向量(行③④). 其中，真实邻居对应的位为 1，其余为 0. 2)在填充$\boldsymbol{B}_i'$完毕之后，用户对$\boldsymbol{B}_i'$进行洗牌，从而避免数据采集者按顺序推导用户的邻居信息(行⑥). 在此基础之上，用户便可将一个n维的邻接向量缩减为θ维. 依据真实图的稀疏性，许多非邻居位将会被移除，从而可一定程度上减少随机化过程产生的噪声边.

算法 2. 剪枝算法 Prune.

输入：邻接向量$\boldsymbol{B}_i = (b_1, b_2, \cdots, b_n)$，$b_j \in \{0,1\}$，阈值$\theta$；

输出：剪枝后的邻接向量$\boldsymbol{B}_i' = \left(b_1', b_2', \cdots, b_\theta'\right)$.

① if $\sum_{j=1}^{n} b_j > \theta$ do /*$\sum_{j=1}^{n} b_j$为用户i的度值d_i*/

② $\boldsymbol{B}_i' = Cut(\boldsymbol{B}_i) = \left(b_1', b_2', \cdots, b_\theta'\right)$;

/*随机选择$\boldsymbol{B}_i$中θ个邻居，并均填充为 1，即$b_j' \in \{1\}$*/

③ else if $\sum_{j=1}^{n} b_j \leqslant \theta$ do

④ $\boldsymbol{B}_i' = \left(b_1', b_2', \cdots b'_{\sum_{j=1}^{n} b_j}, \cdots, b_\theta'\right)$;

/*$b'_{j \in \left\{1, \sum_{j=1}^{n} b_j\right\}} \in \{1\}$，选取$\sum_{j=1}^{n} b_j$个邻居填充为前$\sum_{j=1}^{n} b_j$位，计为 1；$b'_{j \in \left\{\sum_{j=1}^{n} b_j, \theta\right\}} \in \{0\}$，即随机抽取$\theta - d_i$个非邻居充当邻接向量的后几位，计为 0*/

⑤ end if

⑥ $Shuffle\left(\boldsymbol{B}_i'\right)$; /*对$\boldsymbol{B}_i'$进行洗牌*/

⑦ return $\boldsymbol{B}_i'$.

为了能在强隐私保护下采集三角计数序列，我们在算法 2 的基础之上提出基于 Node-LDP 的三角计数序列采集算法，具体如算法 3 所示. 在算法 3 中，引入 2 轮交互模型来进一步提升采集结果的数据效用. 其中：第 1 轮交互用于获取分布式用户的噪声边信息，从而聚合获得噪声图结构数据；随后，数据采集者将噪声图结构数据发送给各分布式用户. 第 2 轮交互过程中，各分布式用户利用自身一阶邻居子图的邻居信息和噪声图结构数据，计算各自的局部三角计数，并通过拉普拉斯机制对三角计数加噪，加噪后的三角计数发送给数据采集者.

算法 3. 基于 Node-LDP 的三角计数序列采集算法 NLDP-TS.

输入：邻接矩阵$\boldsymbol{M} = (\boldsymbol{B}_1, \boldsymbol{B}_2, \cdots, \boldsymbol{B}_n)$，$\boldsymbol{B}_i = (b_1, b_2, \cdots, b_n)$，$b_j \in \{0,1\}$，频率上限$Level$，隐私预算$\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3$，组距$L$；

输出：具备 Node-LDP 保护的三角计数序列$T' = \left(T_1', T_2', \cdots, T_n'\right)$.

/*第 1 轮交互*/

① $\theta = NLDP\text{-}DD(Level, \varepsilon_1, L)$;

/*调用算法 1 获取度分布估计，并结合$Level$计算度阈值θ*/

/*用户*/

② for i in $\{1,2,\cdots,n\}$ do

③ $\boldsymbol{B}_i' = Prune(\boldsymbol{B}_i, \theta)$; /*裁剪邻接向量*/

④ 通过ε_2/θ-RR 扰动$\boldsymbol{B}_i'$中的每一位

$$\tilde{B}_i\left[j\right] = \begin{cases} B_i'\left[j\right], & p = \dfrac{\mathrm{e}^{\varepsilon_2/\theta}}{\mathrm{e}^{\varepsilon_2/\theta}+1}; \\ 1 - B_i'\left[j\right], & q = \dfrac{1}{\mathrm{e}^{\varepsilon_2/\theta}+1}; \end{cases}$$

/*$B_i'[j]$表示$\boldsymbol{B}_i'$中的第j位，$\tilde{B}_i[j]$（$\tilde{\boldsymbol{B}}_i$中的第j位）表示扰动后的$B_i'[j]$，$\tilde{\boldsymbol{B}}_i$为经过扰动的$\boldsymbol{B}_i'$*/

⑤ 将$\tilde{\boldsymbol{B}}_i$发送给数据采集者；

⑥ end for

/*数据采集者*/

⑦ $G' = construct\left(\tilde{\boldsymbol{B}}_1, \tilde{\boldsymbol{B}}_2, \cdots, \tilde{\boldsymbol{B}}_n\right)$; /*构建噪声图*/

⑧ 将噪声图G'发送给每个用户；

/*第 2 轮交互*/

/*用户*/

⑨ for i in $\{1,2,\cdots,n\}$ do

⑩ /*计算噪声 2-star 计数*/

$t_i \leftarrow \frac{1}{2}\left[\sum_{j=1}^{\theta} B'_{ij}\left(\sum_{j=1}^{\theta} B'_{ij} - 1\right)\right]$;

⑪ /*计算噪声局部三角计数*/

$s_i = |\left(v_i, v_j, v_k\right) : j < k, a_{i,j} = a_{i,k} = 1,$ $\left(v_j, v_k\right)$ in $G'|$;

/*$a_{i,j}$表示用户i本地端的边缘(i,j)是否存在，若存在记为 1，否则记为 0*/

⑫ $w_i \leftarrow s_i - qt_i$;

⑬ $w_i' \leftarrow w_i + Lap\left(\frac{\theta(\theta-1)}{2\varepsilon_3}\right)$;

⑭ 将w_i'发送给数据采集者；

⑮ end for

/*数据采集者*/

⑯ for i in $\{1,2,\cdots,n\}$ do

⑰ $T_i' = \frac{w_i'}{2p-1}$;

⑱ end for

⑲ $T' = \left(T_1', T_2', \cdots, T_n'\right)$;

⑳ return T'.

算法 3 中，行①~⑧为第 1 轮交互，行⑨~⑲为第 2 轮交互，各分为 4 个步骤. 第 1 轮交互中，首先，数据采集者将隐私预算ε拆分为ε_1，ε_2，ε_3，并将ε_1、ε_2、ε_3、组距L和频率上限$Level$发布给每位用户. 其次，用户和数据采集者以ε_1和L为参数，通过算法 1 获取度频率分布估计. 数据采集者根据$Level$计算频率和达到$Level$时的最大度值θ，将其设为度阈值（行①）. 再次，用户根据θ和自身真实邻接向量$\boldsymbol{B}_i$，调用算法 2 对邻接向量$\boldsymbol{B}_i$进行裁剪，并通过ε_2/θ-RR 扰动裁剪后的邻接向量中的每一位，将加噪后的邻接向量$\tilde{\boldsymbol{B}}_i$发送给数据采集者（行③~⑤）. 最后，数据采集者根据采集到的噪声邻接向量$\tilde{\boldsymbol{B}}_i$构建噪声图结构数据G'，并将其发送给用户（行⑦⑧）. 值得注意的是，θ无需用户或数据采集者自行设置，其由$Level$和度分布估计计算而得，与$Level$成正比，即$Level$越大，θ越大.

第 2 轮交互中，首先，各分布式用户结合噪声图结构数据G'和自身裁剪后的邻接向量$\boldsymbol{B}_i'$计算其局部三角计数. 以用户i为例，用户i根据裁剪后的$\boldsymbol{B}_i'$计算其真实的 2-star 计数（行⑩）. 其中，用户i的 2-star 可表示为以其自身为中心节点来连接 2 条边的子图. 因此，用户i的 2-star 计数可表示为

$$t_i \leftarrow \frac{1}{2}\left[\sum_{j=1}^{\theta} B'_{ij}\left(\sum_{j=1}^{\theta} B'_{ij} - 1\right)\right].$$

同时，用户i可通过$\boldsymbol{B}_i'$上的真实边缘信息和G'中的噪声边信息计算他的噪声局部三角计数（行⑪），即

$$s_i = \left|\left(v_i, v_j, v_k\right) : j < k, a_{i,j} = a_{i,k} = 1, \left(v_j, v_k\right) \text{ in } G'\right|,$$

其中$a_{i,j}$表示用户i本地端的边缘(i,j)是否存在，若存在则$a_{i,j}$=1，否则$a_{ij} = 0$. 其次，用户i将$s_i - qt_i$发送给数据采集者. 由数据采集者根据用户发送t_i和s_i，并结合 RR 校正公式估计其三角计数，即

$$T_i' = \frac{s_i - qt_i}{2p-1}.$$

为了避免直接发送$s_i - qt_i$引发的边缘信息泄露问题，用户向该值添加拉普拉斯噪声，使算法满足 Node-LDP. 同时，为了降低满足 Node-LDP 时添加的噪声量，算法不采用$d_{\max}(d_{\max}-1)/2$作为敏感度. 用户i向$s_i - qt_i$添加敏感度为$\theta(\theta-1)/2$的拉普拉斯噪声，使算法满足 Node-LDP（行⑫⑬）. 再次，数据采集者估计该用户的三角计数（行⑰），即

$$T_i' = \frac{s_i - qt_i + Lap\left(\frac{\theta(\theta-1)}{2\varepsilon_3}\right)}{2p-1}.$$

最后，数据采集者根据获取的所有分布式用户的局部三角计数，聚合得到满足 Node-LDP 的三角计数序列$T' = \left(T_1', T_2', \cdots, T_n'\right)$（行⑲）.

5.2 基于 Edge-LDP 的三角计数序列采集算法

为了满足各类场景下不同隐私保护和数据效用需求，进一步提升数据效用并放松隐私保护强度，本节在算法 2 的基础上提出基于 Edge-LDP 的三角计数序列采集算法，如算法 4 所示.

算法 4 包含 2 轮交互，行①~⑧为第 1 轮交互，行⑨~⑲为第 2 轮交互，各分为 4 个步骤. 第 1 轮交互中，首先，数据采集者将隐私预算ε拆分为ε_1，ε_2，ε_3，并将ε_1、ε_2、ε_3、组距L和频率上限$Level$发布给每位用户. 其次，用户和数据采集者以ε_1和L为参数，通过算法 1 获取度频率分布估计. 在算法 4 中，数据采集者根据$Level$计算频率和达到$Level$时的最大度值θ，并将其设为度阈值（行①）；再次，用户通过θ将其邻接向量$\boldsymbol{B}_i$进行剪枝，获取到剪枝后的邻接向量$\boldsymbol{B}_i'$，并通过ε_2-

RR 扰动$\boldsymbol{B}_i'$中的每一位并将其发送给数据采集者(行③④)；最后，数据采集者通过所有用户的噪声向量$\tilde{\boldsymbol{B}}_i$聚合获得噪声图结构数据G'，并将G'发给所有分布式用户(行⑦⑧)．值得注意的是，由于 Edge-LDP 和 Node-LDP 下的隐私约束有所不同，因此在满足 Edge-LDP 时，算法 4 中用户在第 1 轮交互中仅需使用ε_2-RR(行④)即可．

第 2 轮交互中，首先，所有分布式用户各自依据G'和$\boldsymbol{B}_i'$计算其真实 2-star 计数t_i和噪声三角计数s_i(行⑩⑪)．其次，用户通过拉普拉斯机制将s_i-qt_i加噪扰动后报告给数据采集者．此处与算法 3 不同的是，用户仅需向s_i-qt_i中添加$Lap(\theta/\varepsilon_3)$的噪声即可满足 Edge-LDP．即用户分别向数据采集者报告$s_i-qt_i+Lap(\theta/\varepsilon_3)$(行⑫⑬)．再次，数据采集者根据用户发送的数据，结合 RR 校正公式估计其局部三角计数(行⑰)，即

$$T_i'=\frac{s_i-qt_i+Lap(\theta/\varepsilon_3)}{2p-1}.$$

最后，数据采集者可获得每个用户的局部三角计数估计T_i'，进一步得到满足 Edge-LDP 的三角计数序列估计$T'=\left(T_1',T_2',\cdots,T_n'\right)$(行⑲)．与前述同理，在满足 Edge-LDP 时，算法 4 中用户在第 2 轮交互的报告过程中(行⑬)所使用的参数与算法 3 不同．(与算法 3 的不同之处为算法 4 中行④⑬．)

算法 4. 基于 Edge-LDP 的三角计数序列采集算法 ELDP-TS.

输入：邻接矩阵$\boldsymbol{M}=(\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_n)$，$\boldsymbol{B}_i=(b_1,b_2,\cdots,b_n)$，$b_j\in\{0,1\}$，频率上限$Level$，隐私预算$\varepsilon=\varepsilon_1+\varepsilon_2+\varepsilon_3$；

输出：具备 Edge-LDP 保护的三角计数序列$T'=\left(T_1',T_2',\cdots,T_n'\right)$．

/*第 1 轮交互*/

① $\theta=NLDP\text{-}DD(Level,\varepsilon_1,L)$；

/*用户*/

② for i in $\{1,2,\cdots,n\}$ do

③ $\boldsymbol{B}_i'=Prune(\boldsymbol{B}_i,\theta)$；

④ 通过ε_2-RR 扰动$\boldsymbol{B}_i'$中的每一位

$$\tilde{B}_i\left[j\right]=\begin{cases}B_i'\left[j\right],\ p=\dfrac{\mathrm{e}^{\varepsilon_2}}{\mathrm{e}^{\varepsilon_2}+1},\\ 1-B_i'\left[j\right],\ q=\dfrac{1}{\mathrm{e}^{\varepsilon_2}+1};\end{cases}$$

⑤ 将$\tilde{\boldsymbol{B}}_i$发送给数据采集者；

⑥ end for

/*数据采集者*/

⑦ $G'=construct\left(\tilde{\boldsymbol{B}}_1,\tilde{\boldsymbol{B}}_2,\cdots,\tilde{\boldsymbol{B}}_n\right)$；

⑧ 将噪声图G'发送给每个用户；

/*第 2 轮交互*/

/*用户*/

⑨ for i in $\{1,2,\cdots,n\}$ do

⑩ /*计算噪声 2-star 计数*/

$$t_i\leftarrow\frac{1}{2}\left[\sum_{j=1}^{\theta}B_{ij}'\left(\sum_{j=1}^{\theta}B_{ij}'-1\right)\right];$$

⑪ /*计算噪声局部三角计数*/

$$s_i=|\left(v_i,v_j,v_k\right):j<k,a_{i,j}=a_{i,k}=1,\ \left(v_j,v_k\right)\ \text{in}\ G'|;$$

⑫ $w_i\leftarrow s_i-qt_i$；

⑬ $w_i'\leftarrow w_i+Lap(\theta/\varepsilon_3)$；

⑭ 将w_i'发送给数据采集者；

⑮ end for

/*数据采集者*/

⑯ for i in $\{1,2,\cdots,n\}$ do

⑰ $T_i'=\dfrac{w_i'}{2p-1}$；

⑱ end for

⑲ $T'=\left(T_1',T_2',\cdots,T_n'\right)$；

⑳ return T'.

5.3 三角计数序列采集算法隐私效用分析

本节对所提基于 Node-LDP 和基于 Edge-LDP 的三角计数序列采集算法(算法 3 和算法 4)所具备的隐私保证、结果无偏性及估计结果满足的方差上限进行理论分析．

定理 6. 算法 3 满足ε-Node-LDP.

证明．算法 3 中存在 3 个过程提供隐私保障，即通过算法 1 获取度值上限θ(行①)、通过 RR 扰动邻接向量(行④)以及向估计值中添加拉普拉斯噪声(行⑬)．

首先，依据定理 3 可知，通过算法 1 获取度值上限θ满足ε_1-Node-LDP.

其次，在通过 RR 扰动位向量过程中，易知剪枝后的邻接向量$\boldsymbol{B}_i'$是θ维的．不失一般性，我们假设任意 2 个$\boldsymbol{B}_i'$和$\boldsymbol{B}_j'$为任意位都不一致的相邻位向量，则通过ε_2/θ-RR 扰动时，有

$$\frac{Pr\left[\mathcal{A}\left(\boldsymbol{B}_i'\right)=\boldsymbol{S}\right]}{Pr\left[\mathcal{A}\left(\boldsymbol{B}_j'\right)=\boldsymbol{S}\right]}=$$

$$\frac{Pr[\mathcal{A}(b_1)=s_1]Pr[\mathcal{A}(b_2)=s_2]\cdots Pr[\mathcal{A}(b_\theta)=s_\theta]}{Pr\left[\mathcal{A}\left(b_1'\right)=s_1\right]Pr\left[\mathcal{A}\left(b_2'\right)=s_2\right]\cdots Pr\left[\mathcal{A}\left(b_\theta'\right)=s_\theta\right]}\leqslant$$

$$\frac{p^\theta}{q^\theta}=\mathrm{e}^{\theta\frac{\varepsilon_2}{\theta}}=e^{\varepsilon_2},$$

其中$p=\dfrac{\mathrm{e}^{\varepsilon_2/\theta}}{\mathrm{e}^{\varepsilon_2/\theta}+1}$，$q=1-p$，$\boldsymbol{S}$为$\varepsilon_2/\theta$-RR 的任意输出值．因此，该过程满足$\varepsilon_2$-Node-LDP.

再次，在 Node-LDP 约束下，删除任意顶点对三角计数的敏感度为$\theta(\theta-1)/2$，因此，在报告三角计数过程中添加$Lap\left(\frac{\theta(\theta-1)}{2\varepsilon_3}\right)$满足$\varepsilon_3$-Node-LDP.

根据引理 1，算法 3 提供$\varepsilon_1+\varepsilon_2+\varepsilon_3$-Node-LDP. 其中，$\varepsilon_1+\varepsilon_2+\varepsilon_3=\varepsilon$，因此算法 3 满足$\varepsilon$-Node-LDP. 证毕.

定理 7. 算法 3 收集的任意用户局部三角计数T_i'都是真实计数T_i的无偏估计.

证明.

$$E\left(T_i'\right)=E\left(\frac{w_i'}{2p-1}\right)=E\left[\frac{s_i-qt_i+Lap\left(\frac{\theta(\theta-1)}{2\varepsilon_3}\right)}{2p-1}\right]=\frac{1}{2p-1}E(s_i-qt_i).$$

设t_i^*为用户i的 2-star 中不包括三角形的计数，即$a_{i,j}=a_{i,k}=1$，且$a_{k,j}=0(j<k)$. T_i为用户i的真实局部三角计数，即$a_{i,j}=a_{i,k}=a_{k,j}=1(j<k)$. s_i为用户结合G'和$\boldsymbol{B}_i'$计数的噪声三角计数，即$s_i=\left|\left(v_i,v_j,v_k\right):j<k,a_{i,j}=a_{i,k}=1,\left(v_j,v_k\right)\text{ in }G'\right|$. 同时，设$t_i$为用户$i$的真实 2-star 计数，则有$t_i=t_i^*+T_i$. 根据 RR 的性质，有

$$E(s_i)=T_ip+t_i^*q,$$

则

$$E\left(T_i'\right)=\frac{1}{2p-1}E(s_i-qt_i)=\frac{1}{2p-1}\left[T_ip+t_i^*q-qt_i\right]=\frac{1}{2p-1}\left[T_ip+t_i^*q-q\left(T_i+t_i^*\right)\right]=\frac{(p-q)T_i}{2p-1}=T_i.$$

因此，T_i'是真实计数T_i的无偏估计. 证毕.

定理 8. 算法 3 收集的任意用户局部三角计数T_i'的方差满足

$$Var\left(T_i'\right)\leqslant[\theta(\theta-1)]\frac{\varepsilon_3^2pq+\theta(\theta-1)}{2(2p-1)^2\varepsilon_3^2},$$

其中$p=\frac{\mathrm{e}^{\varepsilon_2/\theta}}{\mathrm{e}^{\varepsilon_2/\theta}+1}$，$q=\frac{1}{\mathrm{e}^{\varepsilon_2/\theta}+1}$.

证明. 设t_i为用户i的 2-star 计数，t_i^*为用户i的 2-star 中不包括三角形的计数，即$a_{i,j}=a_{i,k}=1$，且$a_{k,j}=0(j<k)$. T_i为用户i的局部三角计数，即$a_{i,j}=a_{i,k}=a_{k,j}=1(j<k)$. 有

$$Var\left(T_i'\right)=Var\left(\frac{w_i'}{2p-1}\right)=Var\left[\frac{s_i-qt_i+Lap\left(\frac{\theta(\theta-1)}{2\varepsilon_3}\right)}{2p-1}\right]=\frac{1}{(2p-1)^2}Var\left[s_i-qt_i+Lap\left(\frac{\theta(\theta-1)}{2\varepsilon_3}\right)\right]=\frac{1}{(2p-1)^2}\left\{Var(s_i)+Var\left[Lap\left(\frac{\theta(\theta-1)}{2\varepsilon_3}\right)\right]\right\}=\frac{T_ip(1-p)+t_i^*q(1-q)}{(2p-1)^2}+\frac{\theta^2(\theta-1)^2}{2(2p-1)^2\varepsilon_3^2}=\frac{\left(T_i+t_i^*\right)pq}{(2p-1)^2}+\frac{\theta^2(\theta-1)^2}{2(2p-1)^2\varepsilon_3^2}=\frac{t_ipq}{(2p-1)^2}+\frac{\theta^2(\theta-1)^2}{2(2p-1)^2\varepsilon_3^2}\leqslant\frac{\theta(\theta-1)pq}{2(2p-1)^2}+\frac{\theta^2(\theta-1)^2}{2(2p-1)^2\varepsilon_3^2}=[\theta(\theta-1)]\frac{\varepsilon_3^2pq+\theta(\theta-1)}{2(2p-1)^2\varepsilon_3^2}.$$

证毕.

定理 9. 算法 4 满足ε_1-Node-LDP$+\varepsilon_2+\varepsilon_3$-Edge-LDP.

证明. 与算法 3 一致，算法 4 也存在 3 个过程提供隐私保障，即通过算法 1 获取度值上限θ(行①). 通过 RR 扰动位向量(行④)以及通过向估计值中添加拉普拉斯噪声(行⑬).

首先，依据定理 3 可知，通过算法 1 获取度值上限θ满足ε_1-Node-LDP.

其次，在通过 RR 扰动位向量过程中，易知剪枝后的位向量$\boldsymbol{B}_i'$是θ维的. 不失一般性，我们假设任意 2 个$\boldsymbol{B}_i'$和$\boldsymbol{B}_j'$为仅第 1 位不一致的相邻位向量，则通过ε_2-RR 扰动时，有

$$\frac{Pr\left[\mathcal{A}\left(\boldsymbol{B}_i'\right)=\boldsymbol{S}\right]}{Pr\left[\mathcal{A}\left(\boldsymbol{B}_j'\right)=\boldsymbol{S}\right]}=\frac{Pr[\mathcal{A}(b_1)=s_1]Pr[\mathcal{A}(b_2)=s_2]\cdots Pr[\mathcal{A}(b_\theta)=s_\theta]}{Pr\left[\mathcal{A}\left(b_1'\right)=s_1\right]Pr\left[\mathcal{A}\left(b_2'\right)=s_2\right]\cdots Pr\left[\mathcal{A}\left(b_\theta'\right)=s_\theta\right]}=\frac{Pr[\mathcal{A}(b_1)=s_1]}{Pr\left[\mathcal{A}\left(b_1'\right)=s_1\right]}\leqslant\frac{p}{q}=\mathrm{e}^{\varepsilon_2},$$

其中$p=\frac{\mathrm{e}^{\varepsilon_2}}{\mathrm{e}^{\varepsilon_2}+1}$，$q=1-p$，$\boldsymbol{S}$为$\varepsilon_2$-RR 的任意输出值. 因此，该过程满足$\varepsilon_2$-Edge-LDP.

再次，在 Edge -LDP 约束下，删除任意顶点对三角计数的影响为θ，即敏感度为θ，因此，在报告三角计数过程中添加$Lap(\theta/\varepsilon_3)$满足$\varepsilon_3$-Edge-LDP.

综上，根据引理 1，算法 4 满足ε_1-Node-LDP$+\varepsilon_2+\varepsilon_3$-Edge-LDP. 证毕.

定理 10. 算法 4 收集的任意用户局部三角计数T_i'都是真实计数T_i的无偏估计.

证明.

$$E\left(T_i'\right)=E\left(\frac{w_i'}{2p-1}\right)=E\left[\frac{s_i-qt_i+Lap\left(\frac{\theta}{\varepsilon_3}\right)}{2p-1}\right]=\frac{1}{2p-1}E(s_i-qt_i).$$

与定理 7 同理, $E(s_i-qt_i)=(p-q)T_i$, 因此

$$E\left(T_i'\right)=\frac{1}{2p-1}E(s_i-qt_i)=\frac{(p-q)T_i}{2p-1}=T_i.$$

综上, T_i'是真实计数T_i的无偏估计. 证毕.

定理 11. 算法 4 收集的任意用户局部三角计数T_i'的方差满足

$$Var\left(T_i'\right)\leqslant\frac{\varepsilon_3^2pq[\theta(\theta-1)]+4\theta^2}{2(2p-1)^2\varepsilon_3^2},$$

其中$p=\frac{\mathrm{e}^{\varepsilon_2}}{\mathrm{e}^{\varepsilon_2}+1}$, $q=\frac{1}{\mathrm{e}^{\varepsilon_2}+1}$.

证明. 设t_i和t_i^*分别为用户i的 2-star 计数和 2-star 中不包括三角形的计数, T_i为用户i的局部三角计数, 有

$$Var\left(T_i'\right)=Var\left(\frac{w'_i}{2p-1}\right)=Var\left[\frac{s_i-qt_i+Lap\left(\frac{\theta}{\varepsilon_3}\right)}{2p-1}\right]=\frac{1}{(2p-1)^2}Var\left[s_i-qt_i+Lap\left(\frac{\theta}{\varepsilon_3}\right)\right]=\frac{1}{(2p-1)^2}\left\{Var(s_i)+Var\left[Lap\left(\frac{\theta}{\varepsilon_3}\right)\right]\right\}=\frac{T_ip(1-p)+t_i^*q(1-q)}{(2p-1)^2}+\frac{2\theta^2}{(2p-1)^2\varepsilon_3^2}=\frac{\left(T_i+t_i^*\right)pq}{(2p-1)^2}+\frac{2\theta^2}{(2p-1)^2\varepsilon_3^2}=\frac{t_ipq}{(2p-1)^2}+\frac{2\theta^2}{(2p-1)^2\varepsilon_3^2}\leqslant\frac{\theta(\theta-1)pq}{2(2p-1)^2}+\frac{2\theta^2}{(2p-1)^2\varepsilon_3^2}=\frac{\varepsilon_3^2\theta(\theta-1)pq+4\theta^2}{2(2p-1)^2\varepsilon_3^2}.$$

证毕.

在算法 3 和算法 4 中, 首先需要调用算法 1, 因此该过程中用户的时间复杂度和空间复杂度分别为$O(2L+1)$和$O(L+1)$, 数据采集者的时间复杂度和空间复杂度均为$O(n(L+1))$. 其次, 用户需先对邻接向量进行剪枝并扰动剪枝后的位向量, 该过程时间复杂度和空间复杂度均为$O(2\theta)$, 数据采集者收集所有用户的噪声位向量并聚合噪声图, 时间复杂度和空间复杂度均为$O(n\theta)$. 再次, 用户需结合噪声图计算三角计数并加噪, 时间复杂度和空间复杂度分别为$O\left(\frac{d_i(d_i-1)}{2}+1\right)$和$O(n\theta)$. 最后, 数据采集者采集各用户的三角计数, 时间复杂度和空间复杂度均为$O(n)$.

综上, 算法 3 和算法 4 中用户所需的时间复杂度和空间复杂度分别为$O\left(2L+2\theta+\frac{d_i(d_i-1)}{2}+2\right)$和$O(L+1+(n+2)\theta)$, 数据采集者所需的时间复杂度和空间复杂度均为$O(n(L+2+\theta))$.

6 聚类系数采集算法

本节在第 5 节所提算法的基础之上, 进一步引入拉普拉斯机制采集分布式用户的度值, 设计满足 Node-LDP 和 Edge-LDP 的聚类系数采集算法.

6.1 基于 Node-LDP 的聚类系数采集算法

算法 5 为所提的基于 Node-LDP 的聚类系数采集算法, 具体分为 4 个步骤. 首先, 数据采集者将隐私预算ε拆分为ε_1, ε_2, ε_3, ε_4, 并将ε_1、ε_2、ε_3、ε_4、组距L和频率上限$Level$发布给每位用户. 其次, 用户和数据采集者以ε_1, ε_2, ε_3, L, $Level$为参数调用算法 3 获取度阈值θ以及三角计数序列$T'=\left(T_1',T_2',\cdots,T_n'\right)$(行①). 再次, 每个用户通过$\varepsilon_4$-拉普拉斯机制向数据采集者报告其度值$d_i$. 值得注意的是, 在满足 Node-LDP 时, 度值的敏感度为$d_{\max}$. 显然, 向常规图结构数据度值中添加$d_{\max}$的拉普拉斯噪声将使得噪声量过大. 因此我们仍然采用算法 3 中取得的阈值θ作为度值上限, 通过投影方式采集度值. 具体而言, 以用户i为例, 当度值$d_i>\theta$时, 用户i报告$\theta+Lap(\theta/\varepsilon_4)$, 否则直接向真实度值上添加$Lap(\theta/\varepsilon_4)$的噪声并报告(行③~⑦). 最后, 数据采集者便可获取每个用户的局部三角计数估计T_i'和度值估计d_i', 并依据T_i'和d_i'计算每个用户的局部聚类系数估计CC_i'(行⑫), 即

$$CC_i'=\frac{2T_i'}{d_i'\left(d_i'-1\right)},$$

并最终聚合采集满足 Node-LDP 保护的聚类系数估计$CC'=\left(CC_1',CC_2',\cdots,CC_n'\right)$(行⑭).

算法 5. 基于 Node-LDP 的聚类系数采集算法 NLDP-CC.

输入: 邻接矩阵$\boldsymbol{M}=(\boldsymbol{B}_1,\boldsymbol{B}_2,\cdots,\boldsymbol{B}_n)$, $\boldsymbol{B}_i=(b_1,b_2,\cdots,b_n)$, $b_j\in\{0,1\}$, 频率上限$Level$, 隐私预算$\varepsilon=\varepsilon_1+\varepsilon_2+\varepsilon_3+\varepsilon_4$;

输出: 具备 Node-LDP 保护的聚类系数$CC'=\left(CC_1',CC_2',\cdots,CC_n'\right)$.

① $T',\theta=NLDP\text{-}TS(\boldsymbol{M},\varepsilon_1+\varepsilon_2+\varepsilon_3,Level)$;

/*调用算法 3 获取满足 Node-LDP 的三角计数序列T'及度阈值θ*/

/*用户*/

② for i in $\{1,2,\cdots,n\}$ do

/*对度值进行投影*/

③ if $d_i > \theta$ do /*度值大于阈值θ*/

④ $d_i' = \theta + Lap(\theta/\varepsilon_4)$;

⑤ else if $d_i \leqslant \theta$ do/*度值小于阈值θ*/

⑥ $d_i' = d_i + Lap(\theta/\varepsilon_4)$;

⑦ end if

⑧ 将噪声度值d_i'发送给数据采集者;

⑨ end for

⑩ $\tilde{D} = \{d_1', d_2', \cdots, d_n'\}$;

/*数据采集者*/

⑪ for i in $\{1,2,\cdots,n\}$ do

⑫ $CC_i' = \dfrac{2T_i'}{d_i'(d_i'-1)}$;

⑬ end for

⑭ $CC' = (CC_1', CC_2', \cdots, CC_n')$;

⑮ return CC'.

6.2 基于 Edge-LDP 的聚类系数采集算法

本节在算法 4 和算法 5 的基础之上，提出基于 Edge-LDP 的聚类系数采集算法，如算法 6 所示.

类似地，算法 6 也可分为 4 个步骤. 首先，数据采集者将隐私预算ε拆分为ε_1，ε_2，ε_3，ε_4，并将ε_1、ε_2、ε_3、ε_4、组距L和频率上限$Level$发布给每位用户. 其次，用户和数据采集者以ε_1，ε_2，ε_3，L，$Level$为参数调用算法 4，获取满足 Edge-LDP 的三角计数序列$T' = (T_1', T_2', \cdots, T_n')$（行①）. 再次，用户通过$\varepsilon_4$-拉普拉斯机制报告其度值$d_i$（行③④）. 与算法 5 所不同，在满足 Edge-LDP 时，删除任意一条边敏感度仅为 1，敏感度相对较小，因此算法 6 无需删边投影. 用户在报告度值前仅需向度值中添加敏感度为 1 的拉普拉斯噪声，即$Lap(1/\varepsilon_4)$. 最后，数据采集者利用每个用户的局部三角计数估计T_i'和度值估计d_i'计算其局部聚类系数估计CC_i'（行⑦~⑨），并聚合采集满足 Edge-LDP 的全局聚类系数$CC' = (CC_1', CC_2', \cdots, CC_n')$（行⑩）.

与算法 5 不同的是，在采集满足 Edge-LDP 的聚类系数时，首先算法 6 需调用算法 4 采集满足 Edge-LDP 的三角计数序列（行①）；其次，在 Edge-LDP 约束下采集度值的敏感度仅为 1，因此无需通过投影限制采集过程中的敏感度. 进而，在算法 6 的度值报告过程中（行③）仅需添加$Lap(1/\varepsilon_4)$即可.（与算法 5 的不同之处为算法 6 中的行①③.）

算法 6. 基于 Edge-LDP 的聚类系数采集算法 ELDP-CC.

输入：邻接矩阵$\boldsymbol{M} = (\boldsymbol{B}_1, \boldsymbol{B}_2, \cdots, \boldsymbol{B}_n)$，$\boldsymbol{B}_i = (b_1, b_2, \cdots, b_n)$，$b_j \in \{0,1\}$，频率上限$Level$，隐私预算$\varepsilon = \varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$;

输出：具备 Edge-LDP 保护的聚类系数$CC' = (CC_1', CC_2', \cdots, CC_n')$.

① $T' = ELDP\text{-}TS(\boldsymbol{M}, \varepsilon_1 + \varepsilon_2 + \varepsilon_3, Level)$;

/*调用算法 5 获取满足 Edge-LDP 的三角计数序列T'*/

/*用户*/

② for i in $\{1,2,\cdots,n\}$ do

③ $d_i' = d_i + Lap(1/\varepsilon_4)$;

④ 将噪声度值d_i'发送给数据采集者;

⑤ end for

⑥ $\tilde{D} = \{d_1', d_2', \cdots, d_n'\}$;

/*数据采集者*/

⑦ for i in $\{1,2,\cdots,n\}$ do

⑧ $CC_i = \dfrac{2T_i'}{d_i'(d_i'-1)}$;

⑨ end for

⑩ $CC' = (CC_1', CC_2', \cdots, CC_n')$;

⑪ return CC'.

6.3 聚类系数采集算法隐私效用分析

本节对基于 Node-LDP 和基于 Edge-LDP 的聚类系数采集算法（算法 5 和算法 6）具备的隐私保证、结果无偏向及估计结果满足的方差上限进行理论分析.

定理 12. 算法 5 满足ε-Node-LDP.

证明. 算法 5 中，需调用算法 3 获取用户三角计数序列估计（行①），并由用户通过拉普拉斯机制报告自身度值（行③~⑦）.

首先，根据定理 6 可知，算法 3 满足$\varepsilon_1 + \varepsilon_2 + \varepsilon_3$-Node-LDP.

其次，在 Node-LDP 约束下，删除任意顶点敏感度为θ，因此在报告度值时添加$Lap(\theta/\varepsilon_4)$的噪声满足$\varepsilon_4$-Node-LDP.

综上，依据引理 1 可知，算法 5 满足$\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4$-Node-LDP（$\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 = \varepsilon$），因此算法 5 满足$\varepsilon$-Node-LDP. 证毕.

定理 13. 算法 5 收集的任意用户局部聚类系数CC_i'都是真实局部聚类系数CC_i的无偏估计.

证明.

$$E(CC_i') = E\left[\frac{2T_i'}{d_i'(d_i'-1)}\right] = \frac{2}{d_i(d_i-1)}E(T_i').$$

依据定理 7 可知，$E(T_i') = T_i$，因此有

$$E\left(CC_i'\right)=\frac{2}{d_i(d_i-1)}E\left(T_i'\right)=\frac{2T_i}{d_i(d_i-1)}=CC_i.$$

因此，任意用户局部聚类系数CC_i'都是真实局部聚类系数CC_i的无偏估计. 证毕.

定理 14. 算法 5 收集的任意用户局部聚类系数CC_i'的方差满足

$$Var\left(CC_i'\right)\leqslant 4[\theta(\theta-1)]\frac{\left[\varepsilon_3^2pq+\theta(\theta-1)\right]\left[\theta^2\left(10d_i^2-10d_i+3\right)\right]}{(2p-1)^2\varepsilon_3^2d_i^4(d_i-1)^4\varepsilon_4^2},$$

其中$p=\dfrac{\mathrm{e}^{\varepsilon_2/\theta}}{\mathrm{e}^{\varepsilon_2/\theta}+1}$，$q=1-p$.

证明.

$$Var\left(CC_i'\right)=Var\left[\frac{2T_i'}{d_i'\left(d_i'-1\right)}\right]=4Var(T')Var\left[\frac{1}{d_i'\left(d_i'-1\right)}\right].$$

依据定理 8 可知

$$Var\left(T_i'\right)\leqslant[\theta(\theta-1)]\frac{\varepsilon_3^2pq+\theta(\theta-1)}{2(2p-1)^2\varepsilon_3^2}.$$

同时，

$$Var\left[\frac{1}{d_i'\left(d_i'-1\right)}\right]=E\left[\frac{1}{d_i'\left(d_i'-1\right)}\right]^2-\left\{E\left[\frac{1}{d_i'\left(d_i'-1\right)}\right]\right\}^2.$$

根据文献 [12] 可知

$$E\left[\frac{1}{d_i'\left(d_i'-1\right)}\right]^2\approx\frac{1}{d_i^2(d_i-1)^2}+\frac{2\theta^2\left(10d_i^2-10d_i+3\right)}{d_i^4(d_i-1)^4\varepsilon_4^2}.$$

进一步地，

$$\left\{E\left[\frac{1}{d_i'\left(d_i'-1\right)}\right]\right\}^2=\frac{1}{d_i^2(d_i-1)^2}.$$

因此，有

$$Var\left[\frac{1}{d_i'\left(d_i'-1\right)}\right]\approx\frac{1}{d_i^2(d_i-1)^2}+\frac{2\theta^2\left(10d_i^2-10d_i+3\right)}{d_i^4(d_i-1)^4\varepsilon_4^2}-\frac{1}{d_i^2(d_i-1)^2}=\frac{2\theta^2\left(10d_i^2-10d_i+3\right)}{d_i^4(d_i-1)^4\varepsilon_4^2}.$$

综上，

$$Var\left(CC_i'\right)\leqslant 4[\theta(\theta-1)]\frac{\varepsilon_3^2pq+\theta(\theta-1)}{2(2p-1)^2\varepsilon_3^2}\frac{2\theta^2\left(10d_i^2-10d_i+3\right)}{d_i^4(d_i-1)^4\varepsilon_4^2}=4[\theta(\theta-1)]\frac{\left[\varepsilon_3^2pq+\theta(\theta-1)\right]\left[\theta^2\left(10d_i^2-10d_i+3\right)\right]}{(2p-1)^2\varepsilon_3^2d_i^4(d_i-1)^4\varepsilon_4^2}.$$

证毕.

定理 15. 算法 6 满足ε_1-Node-LDP+$\varepsilon_2+\varepsilon_3+\varepsilon_4$-Edge-LDP.

证明. 算法 6 中，需调用算法 4 获取用户三角计数序列估计(行①)，并由用户通过拉普拉斯机制报告自身度值(行③).

首先，根据定理 9 可知，算法 4 满足ε_1-Node-LDP+$\varepsilon_2+\varepsilon_3$-Edge-LDP.

其次，在 Edge-LDP 约束下，删除任意顶点敏感度为 1，因此在报告度值时添加$Lap(1/\varepsilon_4)$满足ε_4-Edge-LDP.

综上，依据引理 1 可知，算法 6 满足ε_1-Node-LDP+$\varepsilon_2+\varepsilon_3+\varepsilon_4$-Edge-LDP. 证毕.

定理 16. 算法 6 收集的任意用户局部聚类系数CC_i'都是真实局部聚类系数CC_i的无偏估计.

证明.

$$E\left(CC_i'\right)=E\left[\frac{2T_i'}{d_i'\left(d_i'-1\right)}\right]=\frac{2}{d_i(d_i-1)}E\left(T_i'\right).$$

依据定理 10 可知，$E\left(T_i'\right)=T_i$，因此有

$$E\left(CC_i'\right)=\frac{2}{d_i(d_i-1)}E\left(T_i'\right)=\frac{2T_i}{d_i(d_i-1)}=CC_i.$$

因此，任意用户局部聚类系数CC_i'都是真实局部聚类系数CC_i的无偏估计. 证毕.

定理 17. 算法 6 收集的任意用户局部聚类系数CC_i'方差满足

$$Var\left(CC_i'\right)\leqslant\frac{4\left[\varepsilon_3^2\theta(\theta-1)pq+4\theta^2\right]\left(10d_i^2-10d_i+3\right)}{(2p-1)^2\varepsilon_3^2d_i^4(d_i-1)^4\varepsilon_4^2},$$

其中$p=\dfrac{\mathrm{e}^{\varepsilon_2}}{\mathrm{e}^{\varepsilon_2}+1}$，$q=1-p$.

证明.

$$Var\left(CC_i'\right)=Var\left[\frac{2T_i'}{d_i'\left(d_i'-1\right)}\right]=4Var(T')Var\left[\frac{1}{d_i'\left(d_i'-1\right)}\right].$$

依据定理 11 可知

$$Var\left(T_i'\right)\leqslant\frac{\varepsilon_3^2pq[\theta(\theta-1)]+4\theta^2}{2(2p-1)^2\varepsilon_3^2}.$$

同时，依据文献 [12]，

$$Var\left[\frac{1}{d_i'\left(d_i'-1\right)}\right]\approx\frac{1}{d_i^2(d_i-1)^2}+\frac{2\left(10d_i^2-10d_i+3\right)}{d_i^4(d_i-1)^4\varepsilon_4^2}-\frac{1}{d_i^2(d_i-1)^2}=\frac{2\left(10d_i^2-10d_i+3\right)}{d_i^4(d_i-1)^4\varepsilon_4^2}.$$

综上，

$$Var\left(CC_i'\right)=4Var(T')Var\left[\frac{1}{d_i'\left(d_i'-1\right)}\right]\leqslant4\frac{\varepsilon_3^2pq[\theta(\theta-1)]+4\theta^2}{2(2p-1)^2\varepsilon_3^2}\frac{2\left(10d_i^2-10d_i+3\right)}{d_i^4(d_i-1)^4\varepsilon_4^2}=$$

$$\frac{4\left[\varepsilon_3^2\theta(\theta-1)pq+4\theta^2\right]\left(10d_i^2-10d_i+3\right)}{(2p-1)^2\varepsilon_3^2d_i^4(d_i-1)^4\varepsilon_4^2}.$$

证毕.

在算法 5 和算法 6 中，需要调用算法 3 和算法 4. 由 5.3 节可知，该过程中用户所需的时间复杂度和空间复杂度分别为 $O\left(2(L+\theta+1)+\frac{d_i(d_i-1)}{2}\right)$ 和 $O(L+1+(n+2)\theta)$，数据采集者所需的时间复杂度和空间复杂度均为 $O(n(L+2+\theta))$. 此外，每个用户需报告自身度值，时间复杂度和空间复杂度均为 $O(1)$. 数据采集者需根据用户的度值序列和三角计数序列依次计算每个用户的局部聚类系数，其时间复杂度和空间复杂度分别为 $O(n)$ 和 $O(2n)$.

综上所述，在算法 5 和算法 6 中，用户所需的时间复杂度和空间复杂度分别为 $O\left(2(L+\theta)+\frac{d_i(d_i-1)}{2}+3\right)$ 和 $O(L+2+(n+2)\theta)$，数据采集者的时间复杂度和空间复杂度分别为 $O(n(L+3+\theta))$ 和 $O(n(L+4+\theta))$.

7 实验分析

为评估所提出的 GS-LDP 算法在采集分布式图结构数据的不同统计指标（度分布、三角计数序列及聚类系数）时的效果，本节选取多个公开真实数据集及生成数据集进行实验分析，并与其他同类图结构数据统计指标采集算法进行对比分析. 本节所有实验均在 Windows 10 上运行，采用 Inter Core i5-6200U CPU, A4000-16G-ARM 和 Python3.8 实现.

7.1 实验设置

本文实验数据集选取了 Stanford Large Network Dataset Collection 网站上 3 个具有代表性的真实图结构数据集，这些数据集也被广泛应用于图结构数据隐私保护研究中[12,14,16-17]. 由于公开数据集都是非分布式的图结构数据，为了有效评估不同算法在分布式图机构数据采集中的隐私保护效果，本文采用与文献 [9, 12, 14, 16−17] 相同的数据处理方法，对实验图结构数据集进行了分布式处理，使得每个分布式用户仅持有单个顶点的一阶邻居子图.

实验选取的图结构数据集分别为 Facebook, AstroPh, Email-Enron, 表 2 简要分析了数据集包含的节点数、边数，以及最大度和平均度等特征信息.

为对比分析图结构数据的不同统计指标的采集结果效用，在实验中采用均方误差（mean squared error, *MSE*）和平均绝对误差（mean absolute error, *MAE*）两个指标进行量化对比. 以度分布为例，设 $F=(f_1,f_2,\cdots,f_l)$ 和 $F'=(f_1',f_2',\cdots,f_l')$ 分别为原始图结构数据的度频率分布和隐私保护采集的度频率分布，$MSE=\frac{1}{l}\sum_{i=1}^{l}(f_i-f_i')^2$ 和 $MAE=\frac{1}{l}\sum_{i=1}^{l}|f_i-f_i'|$. 其中，*MSE* 和 *MAE* 值越小，则表明度分布采集结果 F' 的效用越优.

Table 2 Main Characteristics of the Real Graph-Structure Datasets

表 2 真实图结构数据集的主要特征

数据集	节点数	边数	最大度	平均度
Facebook	4 039	88 234	1 045	43.7
AstroPh	18 772	198 110	504	21.1
Email-Enron	36 692	367 662	1 383	10

7.2 算法对比与分析

本文所提 GS-LDP 算法可满足不同隐私保护强度和数据效用需求的分布式图结构数据多统计指标采集. 在此方面，与现有的基于 DP 的分布式图结构数据统计指标采集算法具有明显差异. 为了更好地表现这些算法间的不同，凸显 GS-LDP 的优势，本节进行了理论对比与分析，如表 3 所示.

由表 3 知，GS-LDP 可采集满足 Node-LDP 的度分布，还可采集满足 Node-LDP 和 Edge-LDP 约束的三角计数序列及聚类系数. 算法 RJ、NodeProj[19] 和 EdgeProj[19] 仅能采集度分布这一单一图结构数据统计指标. 其中，RJ 直接对分布式节点的度值进行本地差分加噪，其仅满足 LDP，无法有效保护图结构数据节点间的结构隐私；NodeProj 和 EdgeProj 均可采集满足 Node-LDP 的度分布估计，但二者需要花费更多的隐私预算以获取度阈值，进而依据度阈值对度值进行加噪扰动，因此其整体数据效用会受到影响；GS-LDP 在采集度分布时满足 Node-LDP，采用分组机制和一元编码机制降低差分加噪量，提高度分布的数据效用，相较于 NodeProj 和 EdgeProj 具有更优的效用优势.

算法 RNL[8] 和 Local2Round$_\triangle$[9] 都仅满足 Edge-LDP，且能同时应用于采集三角计数序列和聚类系数，但是不能有效采集度分布. 其中，RNL 通过 RR 机制扰动子图的邻接向量并进一步聚合得到噪声图结构数据，以此对三角技术序列和聚类系数进行估计；Local2Round$_\triangle$ 在 RNL 基础上引入 2 轮交互模型降低估计中的噪声量；GS-LDP 在采集三角计数序列和聚类系数时，通过引入 Prune 算法和 2 轮交互算法减少随机化过程及报告过程中的噪声量，相较于 RNL 和 Local2Round$_\triangle$ 具有更优的效用优势. 此外，在采集三

Table 3 Comparison of Different Statistics Collecting Algorithms of Distributed Graph Structural Data

表 3 不同分布式图结构数据统计指标采集算法对比

统计指标	采集不同指标时具备的特征及时间复杂度	算法						
		GS-LDP（本文）	RJ[16]	NodeProj[19]	EdgeProj[19]	RNL[8]	Local2Round$_{\triangle}$[9]	LF-GDPR[12]
度分布	是否支持采集	√	√	√	√	×	×	×
	满足的隐私约束	Node-LDP	LDP	Node-LDP	Node-LDP			
	用户时间复杂度	$O(2L+1)$	$O(1)$	$O(3K+1)$	$O(3K+\theta)$			
	数据采集者时间复杂度	$O(n(L+1))$	$O(n)$	$O\left(K(n+1)+n\right)$	$O((K+n)(n+1))$			
三角计数序列	是否支持采集	√	×	×	×	√	√	×
	满足的隐私约束	Node-LDP Edge-LDP				Edge-LDP	Edge-LDP Node-LDP	
	用户时间复杂度	$O\left(\frac{2(L+\theta+1)+}{\frac{d_i(d_i-1)}{2}}\right)$				$O(n)$	$O\left(\frac{2+\frac{n}{2}+}{\frac{d_i(d_i-1)}{2}}\right)$	
	数据采集者时间复杂度	$O(n(L+2+\theta))$				$O\left(\frac{2n+}{n^2}\right)$	$O\left(\frac{2n+n^2}{2}\right)$	
聚类系数	是否支持采集	√	×	×	×	√	√	√
	满足的隐私约束	Node-LDP Edge-LDP				Edge-LDP	Edge-LDP Node-LDP	Edge-LDP
	用户时间复杂度	$O\left(\frac{2(L+\theta)+}{\frac{d_i(d_i-1)}{2}}+3\right)$				$O(n)$	$O\left(\frac{2+\frac{n}{2}+}{\frac{d_i(d_i-1)}{2}}\right)$	$O(n+1)$
	数据采集者时间复杂度	$O(n(L+3+\theta))$				$O\left(\frac{4n+}{n^2}\right)$	$O\left(\frac{4n+n^2}{2}\right)$	$O\left(3n+\frac{n^2}{2}\right)$

注：采集满足的隐私约束，若未标记下划线，则原始文献所提的算法满足该隐私约束；若标记下划线，则原始文献所提的算法不满足该隐私约束，在本文实验中将其扩展为满足该隐私约束（具有一定实用性）以便进行实验对比.

角计数序列和聚类系数时，GS-LDP 可根据不同的参数设置，进而满足 Node-LDP 和 Edge-LDP 这 2 种隐私约束. 考虑到 Local2Round$_{\triangle}$在 Node-LDP 设定下仍具备一定实用性，在 7.4 节和 7.5 节的对比实验中，将 Local2Round$_{\triangle}$扩展使得其满足 Node-LDP，以便与 GS-LDP 在满足 Node-LDP 情况下的三角计数序列和聚类系数采集效用进行对比.

算法 IF-GDPR[12] 仅适用于采集聚类系数的算法，且仅满足 Edge-LDP，其在算法 RNL 的基础上引入拉普拉斯机制和优化机制采集度值，从而结合噪声图信息提升采集聚类系数的精度. 而 GS-LDP 除了能采集聚类系数，还能同时采集度分布和三角计数序列，且能根据数据效用和隐私需求采用满足 Edge-LDP 或 Node-LDP 的隐私参数.

为了明确分析算法 GS-LDP 与其他不同算法在采集分布式图结构数据的不同统计指标时的时间复杂度差异，分为不同统计指标进行分析，如表 3 所示.

1）度分布. GS-LDP 在采集度分布时用户和数据采集者的时间复杂度分别为$O(2L+1)$和$O(n(L+1))$，其中，L为组距，且$L \ll d_{\max}$；算法 RJ 中用户和数据采集者的时间复杂度分别为$O(1)$和$O(n)$；算法 NodeProj 中用户和数据采集者的时间复杂度分别为$O(3K+1)$和$O(K(n+1)+n)$；算法 EdgeProj 中用户和数据采集者的时间复杂度分别为$O(3K+\theta)$和$O((K+n)(n+1))$，其中$K \leqslant d_{\max}, \theta \leqslant d_{\max}, d_{\max} \ll n$. 可见，GS-LDP 的时间复杂度高于 RJ，优于 NodeProj 和 EdgeProj. 但由于 RJ 满足传统 LDP，不能保护图结构数据的结构性隐私，而 GS-LDP，NodeProj，EdgeProj 都满足 Node-LDP，故 GS-LDP 在同类度采集算法中具有效率优势.

2）三角计数序列. GS-LDP 在采集三角计数序列时，用户和数据采集者的时间复杂度分别为$O\left(2(L+\theta+1)+\frac{d_i(d_i-1)}{2}\right)$和$O(n(L+2+\theta))$，其中，$L \ll d_{\max}, d_i \leqslant d_{\max}, \theta \leqslant d_{\max}, d_{\max} \ll n$；算法 RNL 中用户和数据采集者的时间复杂度分别为$O(n)$和$O\left(2n+n^2\right)$；算法 Local2Round$_{\triangle}$中用户和采集者的时间复杂度分别$O\left(2+\frac{n+d_i(d_i-1)}{2}\right)$和$O\left(\frac{2n+n^2}{2}\right)$. 可见，3 种算法中用户的计算时间开销相差不大；而 GS-LDP 中数据采集者的计算时间开销要显著小于 RNL 和 Local2Round$_{\triangle}$. 故而，GS-LDP 在采集三角计数序列时具备显著的效率优势.

3）聚类系数. GS-LDP 在采集聚类系数时，用户和数据采集者的时间复杂度分别为$O\left(2(L+\theta)+\frac{d_i(d_i-1)}{2}+3\right)$和$O(n(L+3+\theta))$，其中$L \ll d_{\max}, d_i \leqslant$

$d_{\max},\theta\leqslant d_{\max},d_{\max}\ll n$；算法 RNL 中用户和数据采集者的时间复杂度分别为 $O(n)$ 和 $O\left(4n+n^2\right)$；算法 Local2Round$_\triangle$ 中用户和数据采集者的时间复杂度分别 $O\left(2+\dfrac{n+d_i(d_i-1)}{2}\right)$ 和 $O\left(\dfrac{2n+n^2}{2}\right)$；算法 LF-GDPR 中用户和数据采集者的时间复杂度分别 $O(n+1)$ 和 $O\left(3n+\dfrac{n^2}{2}\right)$. 可见，4 种不同算法中用户的计算时间开销相差不大；而 GS-LDP 中数据采集者的计算时间开销要显著小于其他 3 种算法.

7.3 度分布采集效用对比与分析

本节对比算法 GS-LDP, NodeProj, EdgeProj, RJ 在采集度分布时的效用差异.

在对比实验中，对 4 种不同的算法均设定隐私预算 $\varepsilon\in\{0.5,1.0,1.5,2.0,2.5,3.0\}$ 和 $L=10$，对比算法在 3 个数据集上采集度分布时的效用. 实验结果如图 3 所示，其中图 3(a)~(c)分别为不同算法在 3 个数据集上的 *MSE* 效用对比结果，图 3(d)~(f)分别为不同算法在 3 个数据集上的 *MAE* 效用对比结果.

由图 3 可知，所提算法 GS-LDP 在采集度分布时的效用显著优于 NodeProj 和 EdgeProj，但差于 RJ. 具体而言，如图 3(a)(d)所示，当数据集为 Facebook 时，GS-LDP 相较于 NodeProj 和 EdgeProj，*MSE* 和 *MAE* 分别降低了 63%~99% 和 19%~90%；相较于 RJ，*MSE* 和 *MAE* 分别提升了 68%~90% 和 43%~73%. 类似地，如图 3(b)(c)和图 3(e)(f)所示，当数据集为 AstroPh 和 Email-Enron 时，GS-LDP 的数据效用比 NodeProj 和 EdgeProj 具有明显优势，但略差于 RJ.

所提出的基于 Node-LDP 的度分布采集算法的效用大幅度提升的原因是，GS-LDP 采用了分组编码的方式，通过 SUE 机制扰动分布式用户的度向量，从而在满足 Node-LDP 的同时，减少所添加的噪声. 此外，RJ 仅为传统的 LDP 算法，并不能保护图结构数据的结构隐私，隐私约束相对宽松. 因此相较于 RJ，GS-LDP 在采集度分布时具备略微的效用劣势，但隐私保护强度要远优于 RJ.

总体而言，GS-LDP 满足 Node-LDP，在提高隐私保护效果的同时提升了采集度分布的效用.

7.4 三角计数序列采集效用对比与分析

本节将对比 GS-LDP 与 RNL, Local2Round$_\triangle$ 采集三角计数序列时的效用差异.

由 7.2 节可知，GS-LDP 和 Local2Round$_\triangle$ 可满足 2 种隐私约束，而 RNL 仅能在 Edge-LDP 约束下具备实用性. 因此，在对比实验中，对 3 种算法均设定隐私预算为 $\varepsilon\in\{1,2,3,4,5,6\}$ 和 $L=10$，对比不同算法在 3 个数据集上采集不同隐私约束(Node-LDP 和 Edge-LDP)三角计数序列时的效用，实验结果如图 4 和

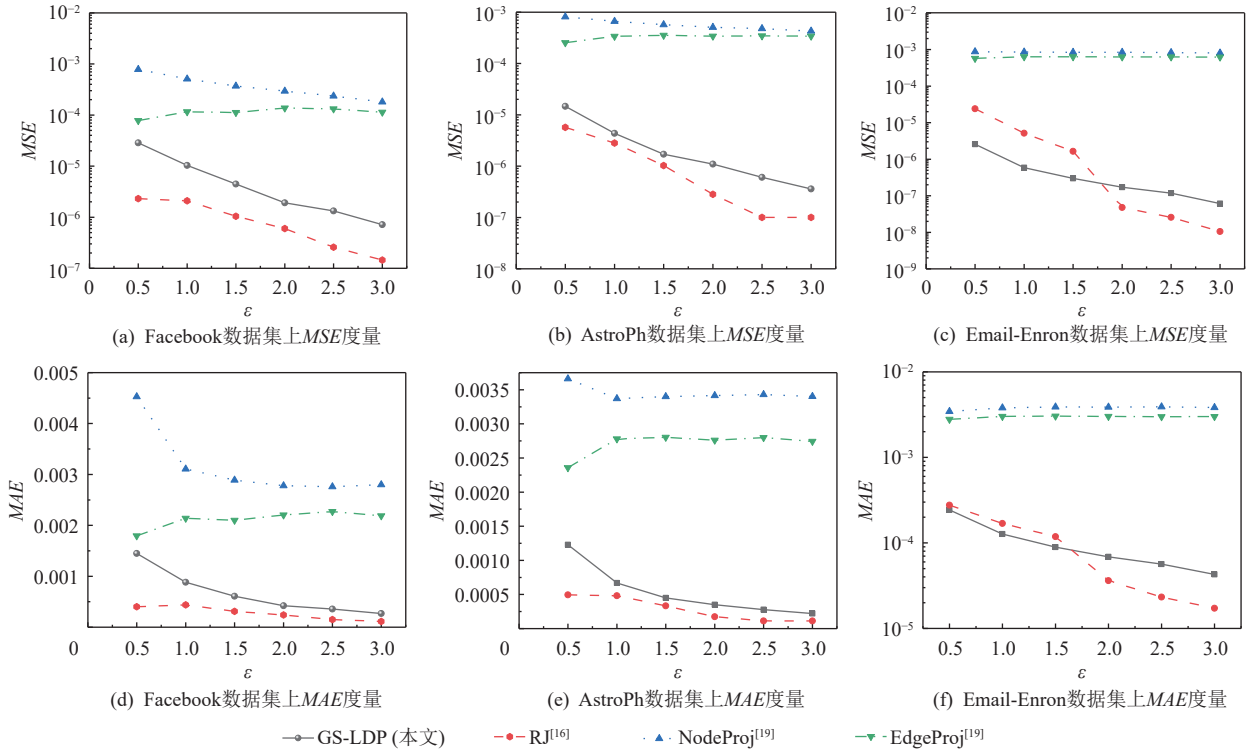


Fig. 3 Utility comparison of four different degree distribution collecting algorithms

图 3 4 种不同度分布采集算法的效用对比

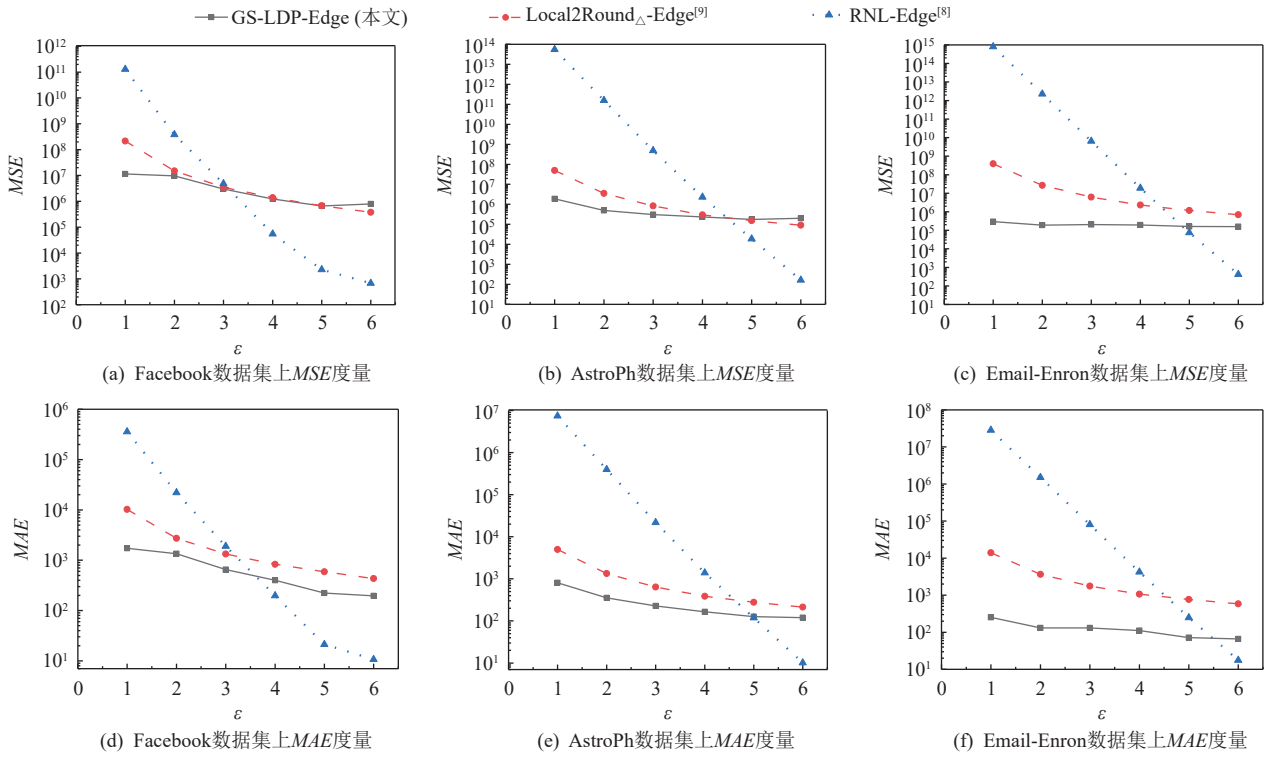


Fig. 4 Utility comparison of different triangle count sequence collecting algorithms when meeting Edge-LDP requirements

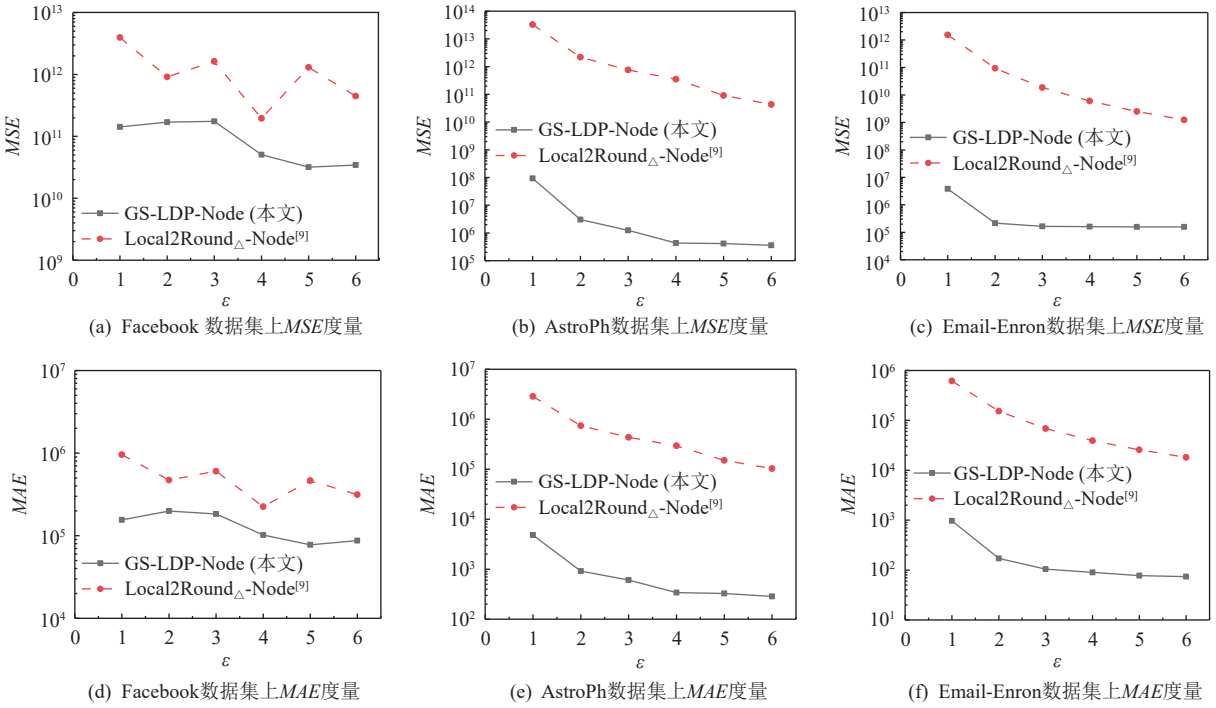
图 4 满足 Edge-LDP 约束时不同三角计数序列采集算法的效用对比

图 5 所示. 在此, 通过在算法名后加 Node 和 Edge 符号作为满足 Edge-LDP 和 Node-LDP 约束的区分. 其中图 4 为 GS-LDP-Edge, RNL-Edge, Local2Round$_{\triangle}$-Edge 的效用对比结果, 图 5 为 GS-LDP-Node 和 Local2Round$_{\triangle}$-Node 的效用对比结果, 其中, 满足 Edge-LDP 的实验中 $Level = 0.98$, 满足 Node-LDP 的实验中 $Level = 0.8$.

由图 4 可知, 在大多数情况下, 所提算法 GS-LDP 在 Edge-LDP 约束下采集分布式图结构数据三角计数序列时的效用在 2 个指标度量下较 RNL-Edge 和 Local2Round$_{\triangle}$-Edge 具备一定的优势. 具体而言, 如图 4(a)(d)所示, 在数据集 Facebook 上, 当 $\varepsilon \in \{1,2,3\}$时, GS-LDP-Edge 相比其他 2 个算法的 *MSE* 和 *MAE* 分别降低了 14%~99% 和 50%~99%, 而当 $\varepsilon \in \{4,5,6\}$时, GS-LDP-Edge 相较于 Local2Round$_{\triangle}$-Edge 的 *MSE* 有 −36%~10% 的波动, *MAE* 降低了 51%~62%, 但相较于 RNL-Edge 在 *MSE* 和 *MAE* 上具备一定的劣势. 同理, 如图 4(b)(c)和图 4(e)(f)所示, 在数据集 AstroPh 和 Email-Enron 上, 无论隐私预算 ε 为何值, GS-LDP-Edge 相较于 Local2Round$_{\triangle}$-Edge 在 *MSE* 和 *MAE* 上均存在不同程度的优势. 而与 RNL-Edge 相比, 当 $\varepsilon \in \{1,2,3,4\}$时 GS-LDP-Edge 具备明显的效用优势, 但当 $\varepsilon \in \{5,6\}$时 GS-LDP-Edge 的效用较 RNL-Edge 存在劣势.

整体而言, 在不同数据集上, GS-LDP-Edge 相较于 RNL-Edge 的 *MSE* 和 *MAE* 随着隐私预算 ε 的取值不同而各有优势, 而相较于 Local2Round$_{\triangle}$-Edge 具备明显优势; 特别是隐私预算 ε 较小时, GS-LDP-Edge 效用优势更为突出. 具体而言, 在满足 Edge-LDP 约束及隐私预算较小时, Prune 算法和 2 轮交互模型在一定程度上可以减少三角计数估计过程中的噪声量; 但当隐私预算 ε 较大时, 噪声量的减少使得 Prune 算法可能会删掉部分无需删除的边从而导致最终估计精度受一定影响. 但整体而言, 在大多数情况下, GS-LDP-Edge 仍然具备相应的效用优势.

由图 5 可知, 所提出的基于 Node-LDP 的分布式图结构数据三角计数序列采集算法 GS-LDP-Node 的效用在 2 个指标度量下都优于算法 Local2Round$_{\triangle}$-Node. 具体而言, 如图 5(a)(d)所示, 在数据集 Facebook 上, 当 $\varepsilon \in \{1,2,3,4,5,6\}$时, GS-LDP-Node 相较于 Local2Round$_{\triangle}$-Node 的 *MSE* 和 *MAE* 分别降低了 74%~96% 及 58%~84%. 类似地, 如图 5(b)(c)和图 5(e)(f)所示, 在数据集 AstroPh 和 Email-Enron 上, GS-LDP-Node 的数据效用也明显优于 Local2Round$_{\triangle}$-Node. 所提出的基于 Node-LDP 的三角计数序列采集算法的效用大幅度提升的原因是 Prune 算法可大幅度限制噪声边的数量以及解决 Node-LDP 随机化过程隐私预算分



(a) Facebook 数据集上*MSE*度量 (b) AstroPh数据集上*MSE*度量 (c) Email-Enron数据集上*MSE*度量

(d) Facebook数据集上*MAE*度量 (e) AstroPh数据集上*MAE*度量 (f) Email-Enron数据集上*MAE*度量

Fig. 5 Utility comparison of different triangle count sequence collecting algorithms when meeting Node-LDP requirements

图 5 满足 Node-LDP 约束时不同三角计数序列采集算法的效用对比

配过小等问题，进而提升估计精度.

7.5 聚类系数对比实验分析

本节将对比所提算法 GS-LDP 与算法 RNL, Local2Round$_{\triangle}$, IF-GDPR 在采集聚类系数时的效用差异.

由 7.2 节可知，GS-LDP 和 Local2Round$_{\triangle}$可满足 2 种隐私约束，而 RNL 和 IF-GDPR 仅能在 Edge-LDP 约束下具备实用性. 因此，在对比实验中，分别设置 2 组实验，测试满足 Edge-LDP 时算法 GS-LDP 与 RNL, Local2Round$_{\triangle}$, IF-GDPR 的效用差异以及满足 Node-LDP 时 GS-LDP 与 Local2Round$_{\triangle}$的效用差异.

首先，设置隐私预算 $\varepsilon \in \{1,2,3,4,5,6\}$, $L=10$和 $Level=0.98$，将所提基于 Edge-LDP 的分布式图结构数据聚类系数采集算法与算法 RNL-Edge, Local2Round$_{\triangle}$-Edge, IF-GDPR-Edge 在表 2 中 3 个真实数据集上进行效用对比，结果如图 6 所示；其次，由于真实数据集上的稀疏性，GS-LDP-Node 和 Local2Round$_{\triangle}$-Node 的效用差距也难以彰显. 因此，采取在生成数据集上测试算法 GS-LDP-Node 和 Local2Round$_{\triangle}$-Node 采集聚类系数的效用. 实验设置 $\varepsilon \in \{5,10,15\}$, $L=10$, $Level=0.5$，生成数据集规模 $n \in \{1\times10^3, 2\times10^3, \cdots, 15\times10^3\}$，将所提基于 Node-LDP 的分布式图结构数据聚类系数采集算法 GS-LDP-Node 与 Local2Round$_{\triangle}$-Node 的效用对比，结果如图 7 所示.

由图 6 可知，在不同的数据集上，GS-LDP-Edge 相较于其他算法效用表现各有不同. 在数据集 Facebook 和 Email-Enron 上时，如图 6(a)(c)和图 6(d)(f)所示，IF-GDPR-Edge 的数据效用最优，GS-LDP-Edge 与 Local2Round$_{\triangle}$-Edge 的数据效用相近，且大多数情况下都比 RNL-Edge 的数据效用更优. 特别是当隐私预算 $\varepsilon<5$ 时，算法 GS-LDP-Edge 相比于 RNL-Edge 的数据效用优势更加明显. 在数据集 AstroPh 上，如图 6(b)(e)所示，无论隐私预算 ε 如何取值，GS-LDP-Edge 的数据效用都比 RNL-Edge 和 Local2Round$_{\triangle}$-Edge 更优，*MSE* 和 *MAE* 分别降低了 0.3%~50% 和−1%~43%. 同时，与 IF-GDPR-Edge 的数据效用相比在隐私预算 ε 取值不同时各有优势，当 $\varepsilon \in \{2,3,4\}$, GS-LDP-Edge 相较于 IF-GDPR-Edge 在 *MSE* 和 *MAE* 上降低了 4%~13% 和 4%~8%，而当 $\varepsilon \in \{1,5,6\}$时，GS-LDP-Edge 在 *MSE* 和 *MAE* 具有−2%~−11% 和−4%~−10% 的劣势.

整体而言，随着数据集的规模和隐私预算 ε 的逐步增大，GS-LDP-Edge 相较于 RNL-Edge 和 Local2Round$_{\triangle}$-Edge 的数据效用优势越大，且逐步与算法 IF-GDPR-Edge 的效用越接近. 具体原因与采集满足 Egde-LDP 约束下三角计数序列时大同小异，Prune 算法和 2 轮交互模型在一定程度上可以减少三角计数估计过程中的噪声量，整体可提升聚类系数的估计精度. 但与 IF-GDPR 相比，GS-LDP-Edge 无法进一步引入度值优化机制，因而最终的聚类系数估计精度较算法 IF-

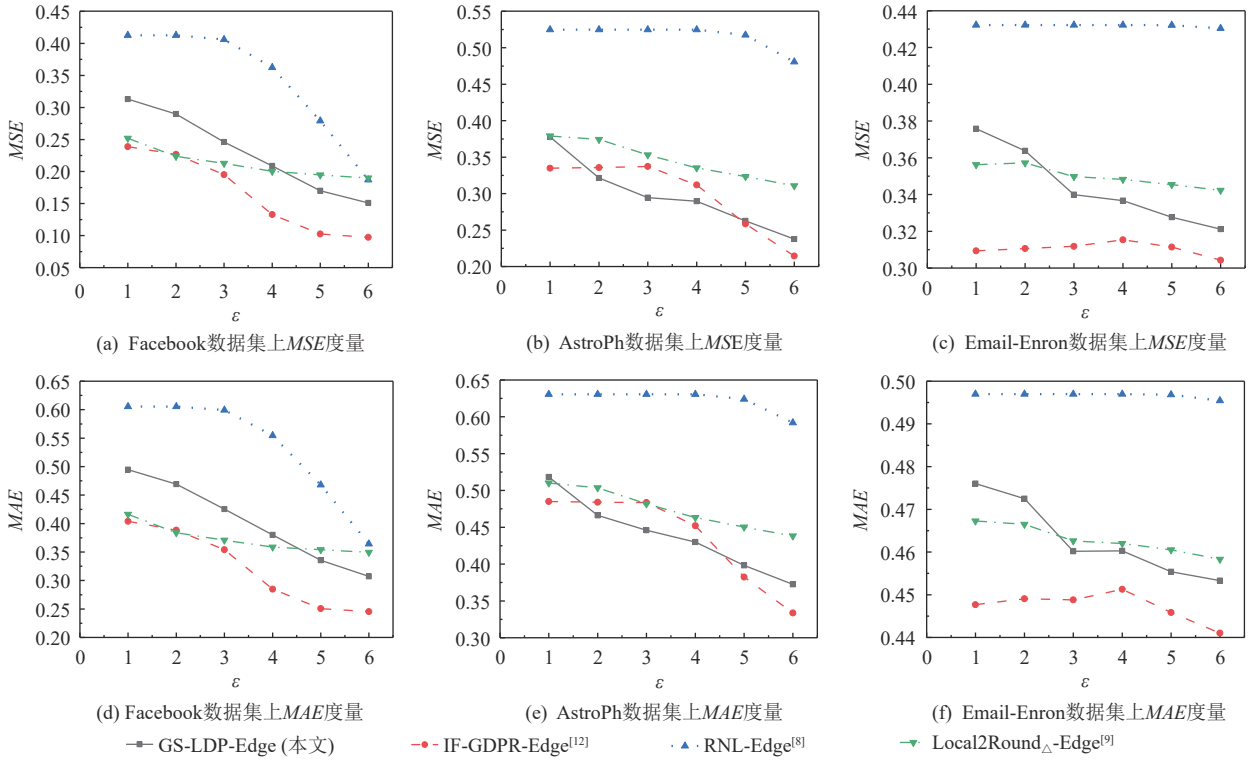


Fig. 6 Utility comparison of different clustering coefficient collecting algorithms when meeting Edge-LDP requirements

图 6 满足 Edge-LDP 约束时不同聚类系数采集算法的效用对比

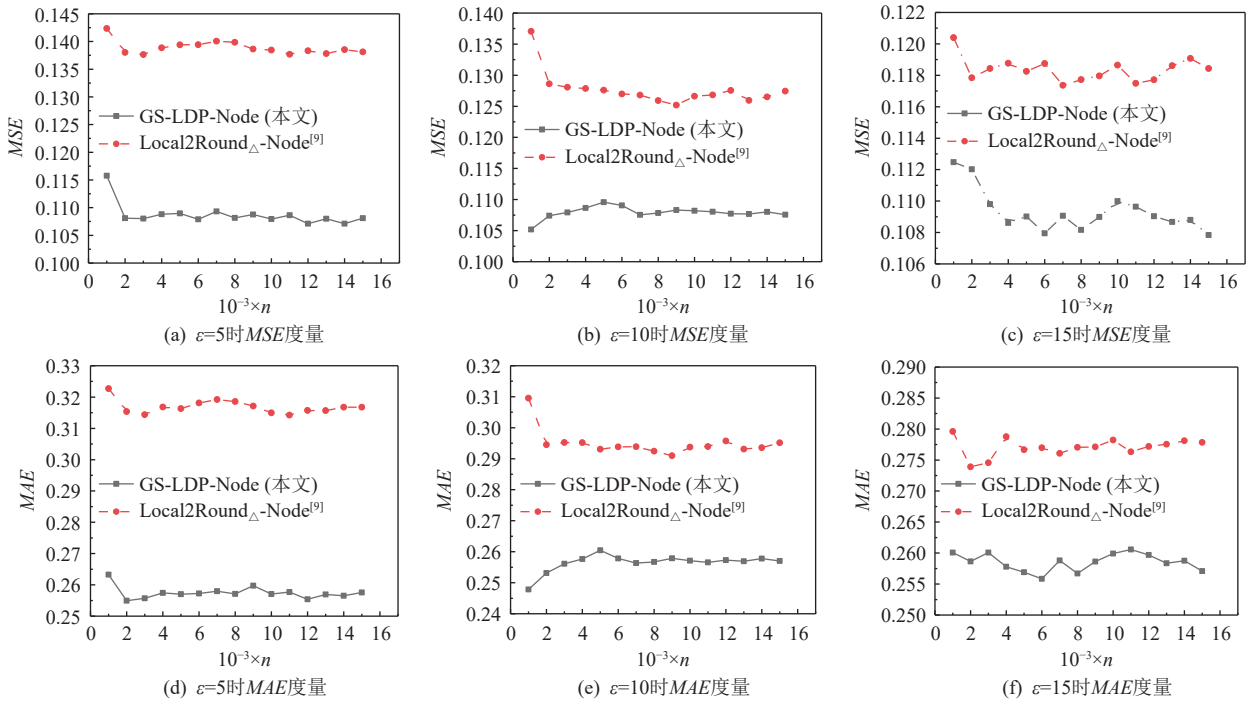


Fig. 7 Utility comparison of different clustering coefficient collecting algorithms when meeting Node-LDP requirements

图 7 满足 Node-LDP 约束时不同聚类系数采集算法的效用对比

GDPR 优势各异.

由图 7 可知, 随着生成图结构数据集的规模逐步扩大, 无论隐私预算ε取何值, GS-LDP-Node 的数据效用均优于 Local2Round$_{\triangle}$-Node, *MAE* 和 *MSE* 分别降低了 6%~23% 和 5%~20%. 与 7.4 节同理, Prune 算法可大幅度限制噪声边的数量以及 Node-LDP 随机

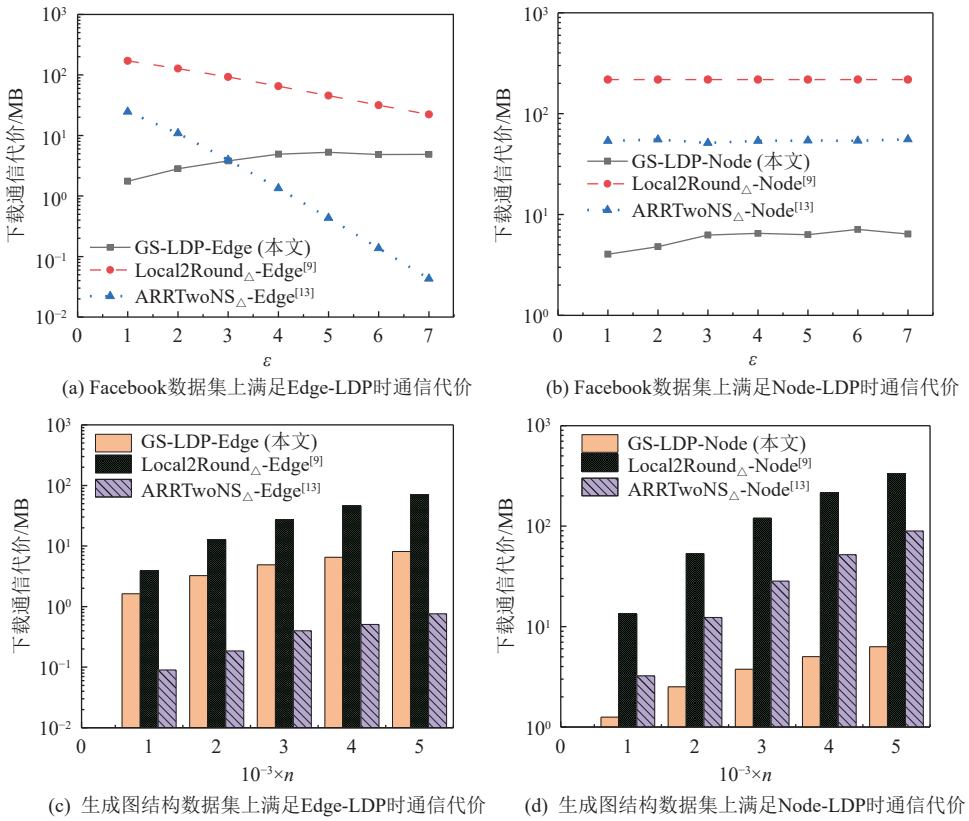
化过程隐私预算过小等问题；且 GS-LDP-Node 可通过度分布算法获取阈值，在一定程度上也可降低度值报告过程中所需添加的噪声量，进而可提升估计精度.

7.6 通信代价分析

GS-LDP 和 Local2Round$_\triangle$[9] 在采集三角计数序列和聚类系数时，都应用 2 轮交互机制在第 2 轮交互中需要数据采集者将噪声数据下发给各分布式用户，以便用户能对自身局部三角计数进行估计. GS-LDP 引入算法 Prune 降低了噪声结构数据中的噪声边的数量，能够降低噪声图结构数据下发时的通信代价. 为直观表明 GS-LDP 的通信优势，本节将 GS-LDP 与 Local2Round$_\triangle$进行通信代价实验对比分析. 同时，进一步引入文献 [13] 中的算法 ARRTwoNS$_\triangle$进行通信代价实验对比，其中 ARRTwoNS$_\triangle$是 Imola 等人[13] 在 Local2Round$_\triangle$基础上引入边采样方法所提出的一种低通信代价的三角计数采集算法，其通过对噪声图进行采样下载，从而减少通信代价和其他计算开销. 因此，本节分别对比上述 3 个算法交互过程中下载所需的通信代价，结果如图 8 所示.

图 8 中，设置了 2 组实验. 首先，在数据集 Facebook 上设置隐私预算$\varepsilon \in \{1,2,3,4,5,6,7\}$，$L=10$，$Level=0.98$，验证算法 GS-LDP，Local2Round$_\triangle$，ARRTwoNS$_\triangle$分别在满足 Edge-LDP 和 Node-LDP 时，在第 2 轮交互过程中噪声图结构数据下载的通信数据量，结果如图 8(a)(b)；其次，数据集规模为$n \in \{1\times10^3, 2\times10^3, 3\times10^3, 4\times10^3, 5\times10^3\}$的生成图设置隐私预算$\varepsilon=5$，$L=10$，$Level=0.5$，验证算法 GS-LDP，Local2Round$_\triangle$，ARRTwoNS$_\triangle$分别在满足 Edge-LDP 和 Node-LDP 时，在第 2 轮交互过程中噪声图结构数据下载的通信数据量，结果如图 8(c)(d).

由图 8(a)(b)可知，GS-LDP 在真实数据集上，无论满足 Edde-LDP 还是满足 Node-LDP 时，其通信代价都显著低于算法 Local2Round$_\triangle$，特别是隐私预算ε越小，GS-LDP 的优势越大；而相较于 ARRTwoNS$_\triangle$而言，满足 Node-LDP 时 GS-LDP 仍旧具备显著的优势；但在满足 Edge-LDP 时，GS-LDP 仅在隐私预算ε较小时具备通信代价优势，但随着隐私预算ε的逐步上升，GS-LDP 则具备一定的劣势. 进一步地，由图 8(c)(d)可知，在模拟的生成图结构数据集上，GS-LDP 的通信代价依旧远小于 Local2Round$_\triangle$，特别是数据规模越大，GS-LDP 的优势越大. 但相对于 ARRTwoNS$_\triangle$而言，



(a) Facebook数据集上满足Edge-LDP时通信代价　(b) Facebook数据集上满足Node-LDP时通信代价

(c) 生成图结构数据集上满足Edge-LDP时通信代价　(d) 生成图结构数据集上满足Node-LDP时通信代价

Fig. 8 Experimental analysis of communication cost

图 8 通信代价实验分析

实验结果与真实数据集结果一致，即满足 Node-LDP 时 GS-LDP 具备显著优势，而满足 Edge-LDP 时 GS-LDP 所需的通信代价较 ARRTwoNS$_\triangle$存在劣势.

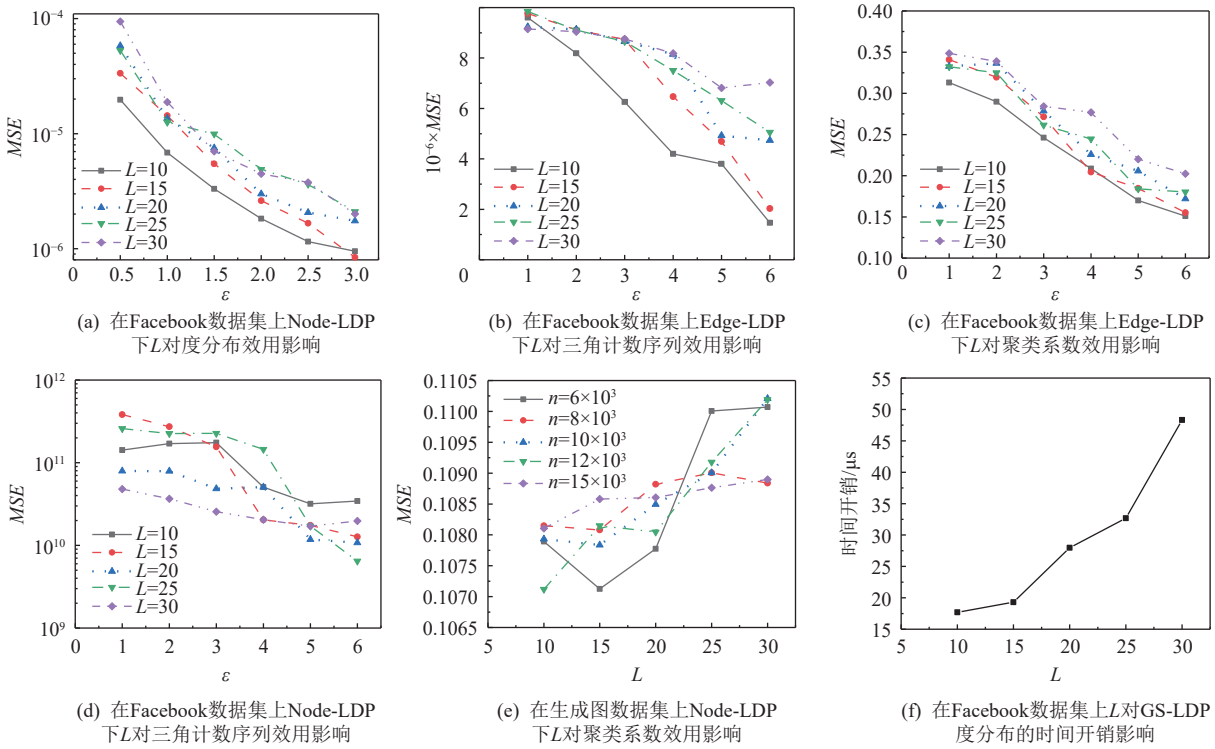
GS-LDP 在满足 Node-LDP 时相较于 Local2Round$_\triangle$和 ARRTwoNS$_\triangle$具备显著通信代价优势的原因在于，GS-LDP 分配了部分隐私预算以通过度分布过程获得更小的度值扰动阈值，并以此对各个用户的子图邻接向量进行裁剪，大幅度减少添加的噪声边，降低通信代价. 同理，在满足 Edge-LDP 时，GS-LDP 相较于 Local2Round$_\triangle$仍旧能够一定程度地通过裁剪算法优化随机化过程中存在的噪声边，降低下载过程中所需的通信代价. 但是，由于在 Edge-LDP 下，噪声边缘的增加量相较于 Node-LDP 会弱化许多，且噪声边缘的增加程度与隐私预算ε成反比，因此仍需直接下载噪声图的 GS-LDP 相较于使用采样方法的 ARRTwoNS$_\triangle$而言在满足 Edge-LDP 时，通信代价具备一定的劣势.

7.7 其他参数影响

本文所提算法 GS-LDP 在采集度分布、三角计数序列及聚类系数时，均需设置合理的组距L和频率上限*Level*，以达到更优的效用. 因此，本节分别测试L和*Level*对 GS-LDP 在 Edge-LDP 和 Node-LDP 下收集的统计指标的效用影响，主要分为 2 个部分. 此处，本节采用 Facebook 数据集和生成图结构数据集进行实验，并使用 *MSE* 作为度量指标.

1)无论是采集度分布，还是采集三角计数序列和聚类系数时，组距L均需被调用. 因此，设置$L\in\{10,15,20,25,30\}$，分别测试其对各统计指标的影响，实验结果如图 9 所示. 首先，针对度分布采集，设置$\varepsilon\in\{0.5,1.0,1.5,2.0,2.5,3.0\}$进行实验，结果如图 9(a)所示. 可直观看出，随着L的逐步上升，GS-LDP 在收集度分布时的效用随之降低. 具体原因在于，更小的L对应于 SUE 扰动的阈值更小，可一定程度减少噪声输入量，进而提升采集精度. 其次，设置$\varepsilon\in\{1,2,3,4,5,6\}$和*Level* = 0.98对 Edge-LDP 约束下采集三角计数序列和聚类系数时的效用进行测试，实验结果如图 9(b)(c)所示. 与度分布同理，二者的效用也随着L的逐步下降而上升. 原因在于，随着L的下降，度分布效用逐步上升，使得度阈值更精确，从而可避免删除过多有效边. 再次，我们分别设置$\varepsilon\in\{1,2,3,4,5,6\}$和*Level* = 0.8，在 Facebook 数据集上测试 GS-LDP 在 Node-LDP 约束下采集三角计数序列的效用，并设置$\varepsilon=5$及*Level* = 0.5 在生成图上测试 Node-LDP 约束下 GS-LDP 采集聚类系数的效用，其中生成图的大小为$n\in\{6\times10^3,8\times10^3,10\times10^3,12\times10^3,15\times10^3\}$，实验结果分别为图 9(d)(e). 如图 9(d)(e)所示，GS-LDP 在满足 Node-LDP 时采集三角计数序列和聚类系数的效用与L不具备线性关系. 同时，随着隐私预算ε的逐步上升，无论L为何值，GS-LDP 的效用均相近. 具体原因为，GS-LDP 在 Node-LDP 下采集三角计数序列和聚类系数时，随机化过程的噪声量更大，尽管L会影响度分布的精度，但相较于 Node-LDP 约束下随机化过程所产生的噪声量仍较小. 最后，由于L的值还对应于算法 1 中用户报告的噪声度向量的长度，进而将会影响各分布式用户度向量报告过程所需的时间开销. 因此，在 Facebook 数据集上测试采集度分布时，L对分布式用户时间开销的影响的实验结果如图 9(f)所示. 如图 9(f)所示，可直观看出，GS-LDP 采集度分布时，用户所需的时间开销随着L的增大而上升. 整体而言，在条件允许的情况下，L可设置相对较小，以保证低时间开销的同时提供更优的效用.

2)GS-LDP 采集三角计数序列和聚类系数时，需通过频率上限*Level*设定合理的度阈值θ. 因此，设置*Level*$\in\{0.5,0.6,0.7,0.8,0.9,0.98\}$，分别测试*Level*对三角计数序列和聚类系数指标采集的影响. 其中，由于计算θ时使用的是度分布估计，可能造成阈值溢出问题，因此使用*Level* = 0.98代替*Level* = 1. 实验结果如图 10 所示. 首先，设置$\varepsilon\in\{1,2,3,4,5,6\}$和$L=10$，测试 Edge-LDP 约束下*Level*对 GS-LDP 采集三角计数序列和聚类系数的效用影响，如图 10(a)(b)所示. 可直观发现，三角计数序列和聚类系数的效用随*Level*的增大而逐步上升，当*Level* = 0.98时，效用最优. 具体原因为：θ与*Level*成正比，随着*Level*的逐步上升，θ值越大. 因此，在满足 Edge-LDP 时，其能够在通过剪枝算法弱化噪声量的同时避免删除过多有效边，进而可将效用最大化. 其次，设置$\varepsilon\in\{1,2,3,4,5,6\}$和$L=10$，在 Facebook 数据集上测试 Node-LDP 约束下*Level*对 GS-LDP 采集三角计数序列的效用影响；同时，设置$\varepsilon=5$和$L=10$，在生成图上测试 Node-LDP 约束下聚类系数的采集效用，其中生成图大小$n\in\{6\times10^3,8\times10^3,10\times10^3,12\times10^3,15\times10^3\}$，实验结果分别为图 10(c)(d). 如图 10(c)(d)所示，在满足 Node-LDP 时，GS-LDP 采集三角计数序列和聚类系数时的效用与*Level*成反比. 具体原因为：在满足 Node-LDP 时，随机响应机制所产生的噪声量更高. 此时，更小的θ可更好地限制噪声边的上限，进而可采集更高效用的指标. 因此，可得出结论，当需要在 Edge-LDP 约束下采集统计时，可在条件允许的情况下将*Level*设置大一些(但不能设为 1，因为可能造成阈值溢出)；而当需要在

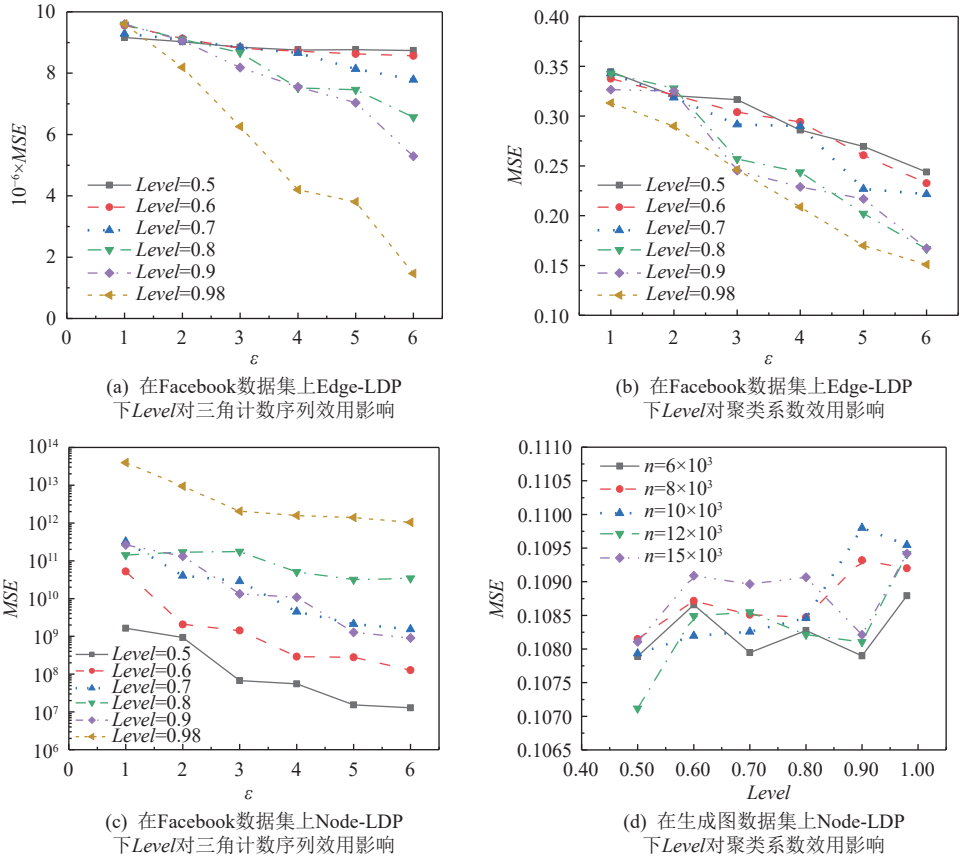


(a) 在Facebook数据集上Node-LDP下L对度分布效用影响　(b) 在Facebook数据集上Edge-LDP下L对三角计数序列效用影响　(c) 在Facebook数据集上Edge-LDP下L对聚类系数效用影响

(d) 在Facebook数据集上Node-LDP下L对三角计数序列效用影响　(e) 在生成图数据集上Node-LDP下L对聚类系数效用影响　(f) 在Facebook数据集上L对GS-LDP度分布的时间开销影响

Fig. 9　The impact of L on *MSE*

图 9　L对 *MSE* 的影响



(a) 在Facebook数据集上Edge-LDP下*Level*对三角计数序列效用影响　(b) 在Facebook数据集上Edge-LDP下*Level*对聚类系数效用影响

(c) 在Facebook数据集上Node-LDP下*Level*对三角计数序列效用影响　(d) 在生成图数据集上Node-LDP下*Level*对聚类系数效用影响

Fig. 10　The impact of *Level* on *MSE*

图 10　*Level*对 *MSE* 的影响

Node-LDP 约束下采集统计时, 则可将*Level*设置小一些, 以保证更优的采集效用.

8 结 论

本文提出一种基于 LDP 的分布式图结构数据统计指标采集算法 GS-LDP, 旨在满足不同隐私保护强度和数据效用需求的同时实现图结构数据多统计指标高效用采集. 首先, 针对特性单一且易泄露隐私的度分布, 采取分组机制和对称一元编码机制进行扰动采集, 使算法满足 Node-LDP 的同时采集数据效用更高的度分布. 其次, 针对强结构性统计指标(三角计数序列和聚类系数), 提出剪枝算法解决随机过程中噪声边添加过多的问题. 进一步地, 根据分布式图结构场景不同的隐私性和安全性需求, 结合剪枝算法和 2 轮交互模型分别提出基于 Node-LDP 和 Edge-LDP 的三角计数序列和聚类系数采集算法, 同时提高算法安全性和数据效用. 最后, 利用真实数据集和生成数据集对不同的图结构数据统计指标采集算法进行理论和实验对比, 结果表明所提算法在采集度分布、三角计数序列和聚类系数时有显著的优势, 同时大幅度降低了 2 轮交互采集算法的通信代价.

参 考 文 献

[1] Wu Dapeng, Zhang Zhihao, Wu Shaoen, et al. Biologically inspired resource allocation for network slices in 5G-enabled Internet of things[J]. IEEE Internet of Things Journal, 2018, 6(6): 9266−9279

[2] Wu Shaoen, Guo Hanqing, Xu Junhong, et al. In-band full duplex wireless communications and networking for IoT devices: Progress, challenges and opportunities[J]. Future Generation Computer Systems, 2019, 92: 705−714

[3] Sharma V, You I, Jayakody D N K, et al. Cooperative trust relaying and privacy preservation via edge-crowdsourcing in social Internet of things[J]. Future Generation Computer Systems, 2019, 92: 758−776

[4] Tian Youliang, Zhang Zhiying, Xiong Jinbo, et al. Achieving graph clustering privacy preservation based on structure entropy in social IoT[J]. IEEE Internet of Things Journal, 2021, 9(4): 2761−2777

[5] Chang Tao, Li Li, Wu Meihan, et al. GraphCS: Graph-based client selection for heterogeneity in federated learning[J]. Journal of Parallel and Distributed Computing, 2023, 177: 131−143

[6] Alrahis L, Sengupta A, Knechtel J, et al. GNN-RE: Graph neural networks for reverse engineering of gate-level netlists[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2021, 41(8): 2435−2448

[7] Liu Yuhan, Chen Hong, Liu Yixuan, et al. State-of-the-art privacy attacks and defenses on graphs[J]. Chinese Journal of Computers, 2022, 45(4): 702−734 (in Chinese)
(刘宇涵, 陈红, 刘艺璇, 等. 图数据上的隐私攻击与防御技术[J]. 计算机学报, 2022, 45(4): 702−734)

[8] Zhan Qin, Ting Yu, Yin Yang, et al. Generating synthetic decentralized social graphs with local differential privacy[C]//Proc of the 24th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2017: 425-438

[9] Imola J, Murakami T, Chaudhuri K. Locally differentially private analysis of graph statistics[C]//Proc of the 30th USENIX Security Symp (USENIX Security 21). Berkeley, CA: USENIX Association, 2021: 983−1000

[10] Ye Qingqing, Meng Xiaofeng, Zhu Minjie, et al. Survey on local differential privacy[J]. Journal of Software, 2018, 29(7): 1981−2005 (in Chinese)
(叶青青, 孟小峰, 朱敏杰, 等. 本地化差分隐私研究综述[J]. 软件学报, 2018, 29(7): 1981−2005)

[11] Wang Tianhao, Blocki J, Li Ninghui, et al. Locally differentially private protocols for frequency estimation[C]//Proc of the 26th USENIX Security Symp (USENIX Security 17). Berkeley, CA: USENIX Association, 2017: 729−745

[12] Ye Qinging, Hu Haibo, Au M H, et al. LF-GDPR: A framework for estimating graph metrics with local differential privacy[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(10): 4905−4920

[13] Imola J, Murakami T, Chaudhuri K. Communication-efficient triangle counting under local differential privacy[C]//Proc of the 31st USENIX Security Symp (USENIX Security 22). Berkeley, CA: USENIX Association, 2022: 537−554

[14] Hou Lihe, Ni Weiwei, Zhang Sen, et al. Wdt-SCAN: Clustering decentralized social graphs with local differential privacy[J]. Computers & Security, 2023, 125: 103036

[15] Gao Tianchong, Li Feng, Chen Yu, et al. Local differential privately anonymizing online social networks under HRG-based model[J]. IEEE Transactions on Computational Social Systems, 2018, 5(4): 1009−1020

[16] Wei Chengkun, Ji Shouling, Liu Changchang, et al. AsgLDP: Collecting and generating decentralized attributed graphs with local differential privacy[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 3239−3254

[17] Fu Nan, Ni Weiwei, Zhang Sen, et al. GC-NLDP: A graph clustering algorithm with local differential privacy[J]. Computers & Security, 2023, 124: 102967

[18] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias[J]. Journal of the American Statistical Association, 1965, 60(309): 63−69

[19] Liu Shang, Cao Yang, Murakami T, et al. A crypto-assisted approach for publishing graph statistics with node local differential privacy[C]//Proc of the 10th IEEE Int Conf on Big Data (Big Data). Piscataway, NJ: IEEE, 2022: 5765−5774

[20] Dwork C. Differential privacy[C]//Proc of the 33rd Int Colloquium on Automata, Languages, and Programming. Berlin: Springer, 2006: 1−12

[21] Jian Xun, Wang Yue, Chen Lei. Publishing graphs under node differential privacy[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(4): 4164−4177

[22] Hay M, Li Chao, Miklau G, et al. Accurate estimation of the degree

distribution of private networks[C]//Proc of the 9th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2009: 169−178

[23] Day W Y, Li Ninghui, Lyu M. Publishing graph degree distribution with node differential privacy[C]//Proc of the 35th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2016: 123−138

[24] Zhang Yuxuan, Wei Jianghong, Li Ji, et al. Graph degree histogram publication method with node-differential privacy[J]. Journal of Computer Research and Development, 2019, 56(3): 508−520 (in Chinese)
(张宇轩, 魏江宏, 李霁, 等. 点差分隐私下图数据的度直方图发布方法[J]. 计算机研究与发展, 2019, 56(3): 508−520)

[25] Sun Haipei, Xiao Xiaokui, Khalil I, et al. Analyzing subgraph statistics from extended local views with decentralized differential privacy[C]//Proc of the 26th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2019: 703−717

[26] Liu Yuhan, Zhao Suyun, Liu Yixuan, et al. Collecting triangle counts with edge relationship local differential privacy[C]//Proc of the 38th IEEE Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2022: 2008−2020

[27] Jiang Honglu, Pei Jian, Yu Dongxiao, et al. Applications of differential privacy in social network analysis: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(1): 108−127

[28] Lin Wanyu, Li Baochun, Wang Cong. Towards private learning on decentralized graphs with local differential privacy[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 2936−2946

[29] Mcsherry F D. Privacy integrated queries: An extensible platform for privacy preserving data analysis [C]//Proc of the 28th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2009: 19−30

Fu Peiwang, born in 2000. Master. Student member of CCF. His main research interests include data security and privacy protection, and differential privacy.

傅培旺, 2000 年生. 硕士. CCF 学生会员. 主要研究方向为数据安全与隐私保护、差分隐私.

Ding Hongfa, born in 1988. PhD, associate professor, master supervisor. Member of CCF. His main research interests include big data security and privacy protection, cryptographic algorithms and protocols, and data security governance.

丁红发, 1988 年生. 博士, 副教授, 硕士生导师. CCF 会员. 主要研究方向为大数据安全与隐私保护、密码算法与协议、数据安全治理.

Liu Hai, born in 1984. PhD, associate professor, master supervisor. His main research interests include data security and privacy protection, and location privacy protection.

刘　海, 1984 年生. 博士, 副教授, 硕士生导师. 主要研究方向为数据安全与隐私保护、位置隐私保护.

Jiang Heling, born in 1983. PhD candidate, lecturer. His main research interests include data security governance and privacy computing protection.

蒋合领, 1983 年生. 博士, 讲师. 主要研究方向为数据安全治理、隐私计算保护.

Tang Mingli, born in 1998. Master. His main research interest includes data security and privacy protection.

唐明丽, 1998 年生. 硕士. 主要研究方向为数据安全与隐私保护.

Yu Yingying, born in 1998. Master. Her main research interest includes data security and privacy protection.

于莹莹, 1998 年生. 硕士. 主要研究方向为数据安全与隐私保护.