计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(1): 256−268, 2025

基于本地边差分隐私的有向图聚类算法

付 楠 倪巍伟 姜泽鹏 侯立贺 张东月 张如玉
(东南大学计算机科学与工程学院 南京 211102)
(funan1988@seu.edu.cn)

Directed Graph Clustering Algorithm with Edge Local Differential Privacy

Fu Nan, Ni Weiwei, Jiang Zepeng, Hou Lihe, Zhang Dongyue, and Zhang Ruyu
(*School of Computer Science and Engineering*, *Southeast University*, *Nanjing* 211102)

Abstract Graph clustering based on local differential privacy has become a hot research topic in recent years. Most existing solutions are mainly realized through modular aggregation using bit vector technology. The linear relationship between the amount of noise and the vector dimension makes balancing clustering quality and privacy challenges. Aiming at the above problems, a directed graph clustering algorithm, DGC-LDP (directed graph clustering under LDP), is proposed based on edge local differential privacy (Edge-LDP). Concretely, the direct encoding method replaces the bit vector encoding method to reduce the amount of data in privacy processing. Meanwhile, a dynamic perturbation mechanism is designed based on the graph structure to balance privacy and statistical utility by adaptively adding noise. Then, according to the individual information uploaded by the terminal, the collector extracts the similarity information between nodes and designs a node aggregation algorithm based on the silhouette coefficient measurement model to generate clusters. Finally, an iterative mechanism is built between the terminal and the collector, and the collector iteratively optimizes the node aggregation form based on the statistical information fed back by the mechanism to achieve high-quality clustering. Theoretical analysis and experimentation on real-world datasets demonstrate that our proposed algorithm can obtain desirable clustering results while satisfying Edge-LDP.

Key words privacy-preserving; local differential privacy; graph clustering; random response; similarity

摘 要 基于本地差分隐私的图聚类工作成为近年来的一个研究热点. 已有工作主要针对的是无向图,且大多利用位向量技术通过模块化聚合实现. 由于噪声量与向量维度成线性关系,使得聚类质量和隐私性难以很好地兼顾. 此外,针对无向图中边的有/无设计的 2 元扰动机制在面对有向图时,因无法对边的方向性进行处理而无法适用. 针对上述问题,提出一种基于本地边差分隐私(edge local differential privacy, Edge-LDP)的有向图聚类算法 DGC-LDP (directed graph clustering under LDP). 具体来说,为了降低噪音量同时适用于有向图,基于直接编码方式设计了一种适用于有向星型图的动态扰动机制,通过自适应添加噪声来平衡隐私性和统计效用. 在此基础上,在终端和收集者之间构建迭代机制. 收集者依据终端上传的噪声数据提取节点间的相似性信息,并设计基于轮廓系数测量模型的节点聚合算法,通过迭代机制不断地优化节点聚合形式形成高质量簇. 理论分析和实验结果表明,所提算法在满足 Edge-LDP 的同时能够有效兼顾聚类精度.

关键词 隐私保护;本地差分隐私;图聚类;随机响应;相似度

中图法分类号 TP309

收稿日期: 2023-03-27; **修回日期**: 2023-11-24

基金项目: 国家自然科学基金项目(61772131)
This work was supported by the National Natural Science Foundation of China (61772131).

通信作者: 倪巍伟(wni@seu.edu.cn)

DOI: 10.7544/issn1000-1239.202330193 **CSTR:** 32373.14.issn1000-1239.202330193

随着人们通过互联网进行购物、办公、交流的常态化，互联网上产生了大量反映个体联系的图结构数据. 采集分布在不同终端上的这些数据进行聚类分析，可以支撑诸如故障预警、行为模式分析等应用. 然而这些数据涉及个体信任关系、交易情况等敏感信息. 直接采集并分析这些数据存在隐私泄露风险. 如何在不侵犯个体隐私的情况下，从终端安全采集这些图数据并进行聚类成为亟待解决的问题.

本地差分隐私(local differential privacy, LDP)[1-6]作为差分隐私(differential privacy, DP)[7]的本地化扩展，是一种被广泛采用的隐私模型. 可以在无需可信第三方干预的情况下实现隐私保护个体数据收集. 应用于图结构数据的LDP模型分为本地节点差分隐私(node local differential privacy, Node-LDP)和本地边差分隐私(edge local differential privacy, Edge-LDP)两类. 相较Edge-LDP模型仅能保护单条边隐私，Node-LDP能对多条边以及节点进行保护，但需要注入大量噪声，降低了数据的可用性. 因此，已有的解决方案大都基于Edge-LDP以提升采集数据的可用性. 例如，Ye等人[8]先利用位向量技术设计一种合成图发布框架LF-GDPR来结合模块度聚合机制实现隐私保护社区探测.Qin等人[9]给出了一种依据位向量技术统计度值并构建社区结构的方案. 虽然文献[9]方案可以实现隐私保护图数据聚类，但位向量技术会使噪声量随向量维度增加而猛增，造成聚类质量和隐私性难以兼顾. 此外，已有工作主要聚焦于无向图，所设计的隐私保护机制针对的是边存在与否这种二元关系. 然而，在有向图中，对边的隐私保护处理是一个三元关系，除了边的有无，还有方向. 这使得传统针对无向图设计的隐私保护算法无法直接应用于有向图.

针对上述问题，提出基于Edge-LDP的隐私保护有向图聚类框架DGC-LDP(directed graph clustering under LDP).采用直接编码取代位向量编码，以减少在隐私保护处理时的数据量；设计基于星型图结构的动态扰动机制，通过自适应添加噪声平衡隐私性和统计效用；在此基础上，在终端和收集者间构建迭代机制. 收集者依据终端上传的个体信息，提取节点间的相似性信息，并设计基于轮廓系数测量模型的节点聚合算法，通过迭代机制不断地优化节点聚合形式以形成高质量簇.

本文的主要贡献包括4个方面：

1)提出基于Edge-LDP模型的有向图聚类算法DGC-LDP，实现个体隐私与有向图聚类质量的兼顾.

2)提出基于星型图结构的有向图动态扰动机制，实现噪声的自适应添加以提高图数据统计准确性. 设计基于轮廓系数测量模型的节点聚合算法来实现簇的构建.

3)在终端与收集者之间构建循环迭代机制，收集者依据反馈的统计信息迭代优化节点聚合形式，实现高质量聚簇.

4)理论上分析了DGC-LDP的隐私性和效用，并设计实验对算法在真实数据集上的成效进行验证.

1 相关工作

近年来，基于LDP的图发布[9-12]与分析[8,13-18]工作得到了研究者的持续关注. 在图发布研究中，Qin等人[9]提出了基于度值的LDPGen算法生成合成图. 然而，度值不能准确反映节点间的关联关系，导致合成图中的节点关联性较差. Zhang等人[10]提出了LDPGM合成图算法，通过使用安全多方计算(SMC)[19]和优化的随机响应(RR)[20]机制来实现隐私强度和数据可用性的提高. 但图的稀疏性使得基于优化后的RR机制形成的邻接矩阵仍然包含很多伪边，导致图精度不足. Wei等人[11]创建了用于合成属性图的AsgLDP算法，基于属性的联合分布以及接受-拒绝采样算法来保证节点之间的关联. 然而，采样易丢失数据，降低了合成图的准确性. Yang等人[12]使用分层随机图(HRG)模型发布满足LDP的分层社交网络. 尽管使用蒙特卡洛马尔可夫链来减少HRG采样空间规模，但图的高维性使得候选空间仍然很大，无法生成高质量社交网络.

在基于LDP的图分析方面，Sun等人[13]假设每个用户允许他/她的朋友看到其所有连接，提出一种统计子图算法. 然而，这种假设在实际应用中具有挑战性. Ye等人[8]基于LF-GDPR框架设计社区检测算法实现聚类. 模块化的准确性取决于复合图的边缘信息. 然而使用位向量技术编码和RR扰动机制生成的复合图会增加许多虚拟边，降低聚簇的准确性. 2022年Zhang[16]提出了社区结构探测算法LDPCD，利用优化的Laplace机制降低噪声注入量，使用极致优化的社区划分策略减少交互轮次以及隐私预算分割次数，提升社区探测质量.Hou等人[18]基于度向量

设计了 Wdt-SCAN 算法, 利用 2/8 定律减少迭代次数, 从而实现高效准确的社区结构构建.

综上所述, 已有研究主要聚焦于图的发布和统计分析任务, 对图聚类的研究较少. 已有的少量隐私保护图聚类探索主要针对的是无向图, 且大多利用位向量技术通过模块化聚合实现. 存在噪声量随维度增大而增大的问题, 导致聚类质量与隐私难以兼顾. 此外, 基于无向图所设计的隐私保护机制针对的是边存在与否这种二元关系. 然而, 在有向图中, 对边的隐私保护处理是一个三元关系, 除了边的有无, 还有方向. 这使得传统针对无向图设计的隐私保护算法无法直接应用于有向图. 因此, 如何实现有向图的隐私保护高质量聚类仍然是一个开放性问题.

2 预备知识及设计动机

在本节中, 我们主要介绍图以及本地差分隐私等相关概念.

2.1 图相关概念

给定有向图$G=(V,E)$, 其中$V=\{v_1,v_2,\cdots,v_n\}$表示节点集合, $E\subseteq V\times V$代表边集合.

定义 1. 星型图. 对于任意图$G=(V,E)$, 如果存在节点$v_i\in V$, 使得公式$\left|\{e_{v_i,v_j}\left|\exists v_j\in V,e_{v_i,v_j}\in E\right.\}\right|=|E|$成立, 则图 G 是以 v_i 为中心点的星型图, 用$G_{v_i}=\left\{V_{v_i},E_{v_i}\right\}$表示.

定义 2. k-hop 图. 如果在图 G 中存在一个节点 v_i, 从该点出发, 到图中任意其他节点所经过的最少节点数不超过 k−1, 则该图被称为节点 v_i 的 k-hop 图.

图聚类是依据节点相似性将具有相似特征的节点分组成簇的过程. 节点间的相似性评估算法对确定节点是否具有相似性至关重要. 共同邻居(common neighbor, CN)索引作为解决此问题的经典方案, 其相关定义有定义 3 和定义 4.

定义 3. 出度边/入度边. 在有向星型图中, 将从中心节点出发指向其他节点的边称为出度边; 而指向中心节点的边称为入度边.

定义 4. 共同邻居(common neighbor, CN). 对图$G=(V,E)$, 给定 V 中任意 2 个节点 v_i 和 v_j, 这 2 个节点的 CN 定义为

$$s_{i,j}=\left|\varpi(v_i)\cap\varpi(v_j)\right|, \tag{1}$$

其中$\varpi(v_i)$表示在以 v_i 为中心构成的有向星型图中出度边连接的节点所形成的节点集. 同理$\varpi\left(v_j\right)$表示在以 v_j 为中心构成的有向星型图中出度边连接的节点所形成的节点集.

2.2 LDP 相关概念

不同于中心差分隐私, 本地环境中的每个人通过机制 A 扰动他/她的数据 v, 并将 $A(v)$ 报告给不被个人信任的收集者进行数据查询和分析. 对于图数据, 其既有节点又有边, LDP 可以应用于节点或边, 这便产生了 Node-LDP[8] 和 Edge-LDP[8]. Edge-LDP 保证随机算法的输出不会揭露 G 中是否存在任何单条边.Edge-LDP 的正式定义如定义 5.

定义 5. Edge-LDP. 一个随机函数 A 满足 ε-Edge-LDP, 当且仅当对于任意 2 个仅有 1 条边不同的星型子图 G^1 和 G^2 以及任意可能的输出$o\in Range(A)$满足

$$Pr\left[A\left(G^1\right)=o\right]\leqslant\exp(\varepsilon)\times Pr\left[A\left(G^2\right)=o\right], \tag{2}$$

其中ε表示隐私预算. ε越小, 函数 A 的隐私保护强度越高.

性质 1. 序列组合性[21-22]. 给定数据集和 n 个扰动算法$M_1,M_2,\cdots,M_i$, 且 $M_i(1\leqslant i\leqslant n)$满足 ε_i-Node-LDP, 则在该数据集上的序列组合满足$\sum_{i=1}^{n}\varepsilon_i$-Edge-LDP.

一元编码(UE)[23]是 LDP 机制下常用的编码方法. 它将输入数据 x 转换为长度为 d 的位向量 $\boldsymbol{B}$=(0, ⋯, 0, 1, 0, ⋯, 0), 其中只有第 j 位(即 j=$Hash(x)$)为 1, 其余位为 0.然后通过随机响应机制对其扰动, 从而实现数据安全保护.

随机应答 (random response,RR)[20] 机制是实现 LDP 的常用技术. 该机制的基本思想是用户在响应敏感的布尔问题时, 以概率 p 真实应答, 以 1−p 的概率给出相反的应答. 为了使随机应答机制满足 ε-LDP, 通常设置$p=\exp(\varepsilon)/\left(1+\exp(\varepsilon)\right)$[24].

使用 RR 机制对任意一个向量 $\boldsymbol{B}$ 进行扰动, 其形式为:

$$B[i]=\begin{cases}B[i], & \text{w.p.} \quad p=\exp(\varepsilon/2)/\left(\exp(\varepsilon/2)+1\right),\\ 1-B[i], & \text{w.p.} \quad p=1/\left(\exp(\varepsilon/2)+1\right).\end{cases} \tag{3}$$

2.3 设计动机

本文主要研究去中心化环境下非属性有向图的隐私保护聚类问题. 社交网络作为去中心化的 1 个典型例子, 其各个用户只拥有 1 个有限的局部视图, 即以自我为中心形成的星型图. 在这些星型图上进行聚类可以为用户提供故障预警、行为模型分析等服务. 然而, 出于隐私考虑, 用户通常不愿意共享他们的联系人列表以形成全局图. 因此, 如何在去中心化环境下实现隐私保护聚类是一个亟待解决的问题.

大多数已有的解决方案都是针对无向图来设计的. 在面对诸如信任关系、交易信息等这类有向图数

据时，直接套用无向图方案可能存在诸多问题. 例如，在无向图中，所设计的隐私保护机制针对的是边存在与否这种二元关系. 然而，在有向图中，对边的隐私保护处理是一个三元关系，除了边的有无，还有方向. 这使得传统针对无向图设计的隐私保护算法无法直接应用于有向图. 为此，提出 DGC-LDP 方案. 该算法的整体结构和细节将在以下部分详细介绍，表 1 列出了本文中的基本符号.

Table 1 Common Symbols

表 1 常用符号

符号	说明
G	任意一张图或网络
E	图 G 中的边集
V	图 G 中的节点集
G_{v_i}	以节点 v_i 为中心的星型图
E_{v_i}	图G_{v_i}中的边集
V_{v_i}	图G_{v_i}中的节点集
e_{v_i,v_j}	从节点 v_i 出发到 v_j 的边
$s_{i,j}$	节点 v_i 和 v_j 间的相似度
V^*	节点集
p_1~p_6	扰动概率
L_1	数据截断长度
S_N	相似度集合
T_i^1, T_i^2	用户 i 的部分边数据集合
$\boldsymbol{B}_{v_i}$	用户 i 的邻接位向量
cc	单个簇
δ	阈值
α	隐私预算分割比例

3 基于 Edge-LDP 的有向图聚类算法

3.1 整体框架

DGC-LDP 方案的核心思想是在终端与收集者间构建一个循环交互机制，通过迭代不断地优化扰动机制和节点聚合方式来实现聚簇. 具体过程包括数据预处理、相似度计算以及构建簇结构 3 个部分. 主要步骤见算法 1，设计思路见图 1.

算法 1. DGC-LDP 算法.

输入：有向星型图$G_{v_i}=\left(V_{v_i},E_{v_i}\right),i\in[1,n]$，隐私预算$\varepsilon$，分配因子$\alpha\in[0,1]$，聚合阈值$\rho$，节点集$V^*=\{v_1,v_2,\cdots\}$，标识 $mark$=1；

输出：簇 C_{all}.

① $\varepsilon=\alpha\varepsilon+\varepsilon_r$；/*$\varepsilon_r=(1-\alpha)\varepsilon$*/

② $G_{v_i}^*\leftarrow F_{\text{DPP}}\left(\alpha\varepsilon,G_{v_i}\right)$；

③ $\varepsilon_r=\varepsilon_r/2$；

④ $S_N\leftarrow F_{\text{SC-LDP}}\left(G_{v_i},\varepsilon_r\right)$；

⑤ $m_{\text{sc}},C_{\text{all}}\leftarrow F_{\text{BC-LDP}}\left(S_N,V^*,\rho\right)$；

⑥ if $m_{\text{sc}}\leqslant mark$

⑦ $mark=m_{\text{sc}}$；

⑧ 重复行 ③~⑨；

⑨ else

⑩ return C_{all}；

⑪ end if

3.2 数据预处理

在 DGC-LDP 算法中，采用直接编码方式[23]处理用户数据. 根据 LDP 的定义，为了保证所有用户满足同等级别的隐私性，需要对每一位用户的星型图做截断或填充处理，以使各个星型图包含等量的边信息. 数据预处理步骤见算法 2.

算法 2. DPP (data-preprocessing) 算法.

输入：图$G_{v_i}=\left(V_{v_i},E_{v_i}\right)$，$i\in[1,n]$，隐私预算$\varepsilon$，分配因子$\alpha\in[0,1]$，截断因子$L_1\in[0,1]$；

输出：图$G_{v_i}=\left(V_{v_i},E_{v_i}\right)$.

/*用户端*/

① for $i\in[1,n]$

② 消除G_{v_i}中指向中心节点v_i的边；

③ 提取度值$d_i=\left|E_{v_i}\right|$；

④ $d_i^*=F_{\text{UE}}\left(\left|E_{v_i}\right|\right)$；/*编码度值*/

⑤ 发送$d_i'=F_{\text{RR}}\left(\alpha\varepsilon,\left|E_{v_i}\right|\right)$给收集者；

⑥ end for

/*收集端*/

⑦ 构建度分布并将与参数L_1对应的度值d^*共享给每个个体；

/*用户端*/

⑧ for $i\in[1,n]$

⑨ if $\left|E_{v_i}\right|\geqslant d^*$

⑩ 从G_{v_i}中随机选择$\left|E_{v_i}\right|-d^*$边删除；

⑪ else

⑫ 从$\left\{e_{v_i,\bar{v}_1},e_{v_i,\bar{v}_2},\cdots,e_{v_i,\bar{v}_\ell}\right\}$中随机选择$d^*-\left|E_{v_i}\right|$条伪边填充到$G_{v_i}$；

⑬ end if

⑭ end for

⑮ return G_{v_i}.

算法 2 描述了数据预处理的具体细节. 有向星型图既包括入度边又包括出度边. 在保证信息不丢失的同时减轻后期隐私处理数据时的难度，每一位用

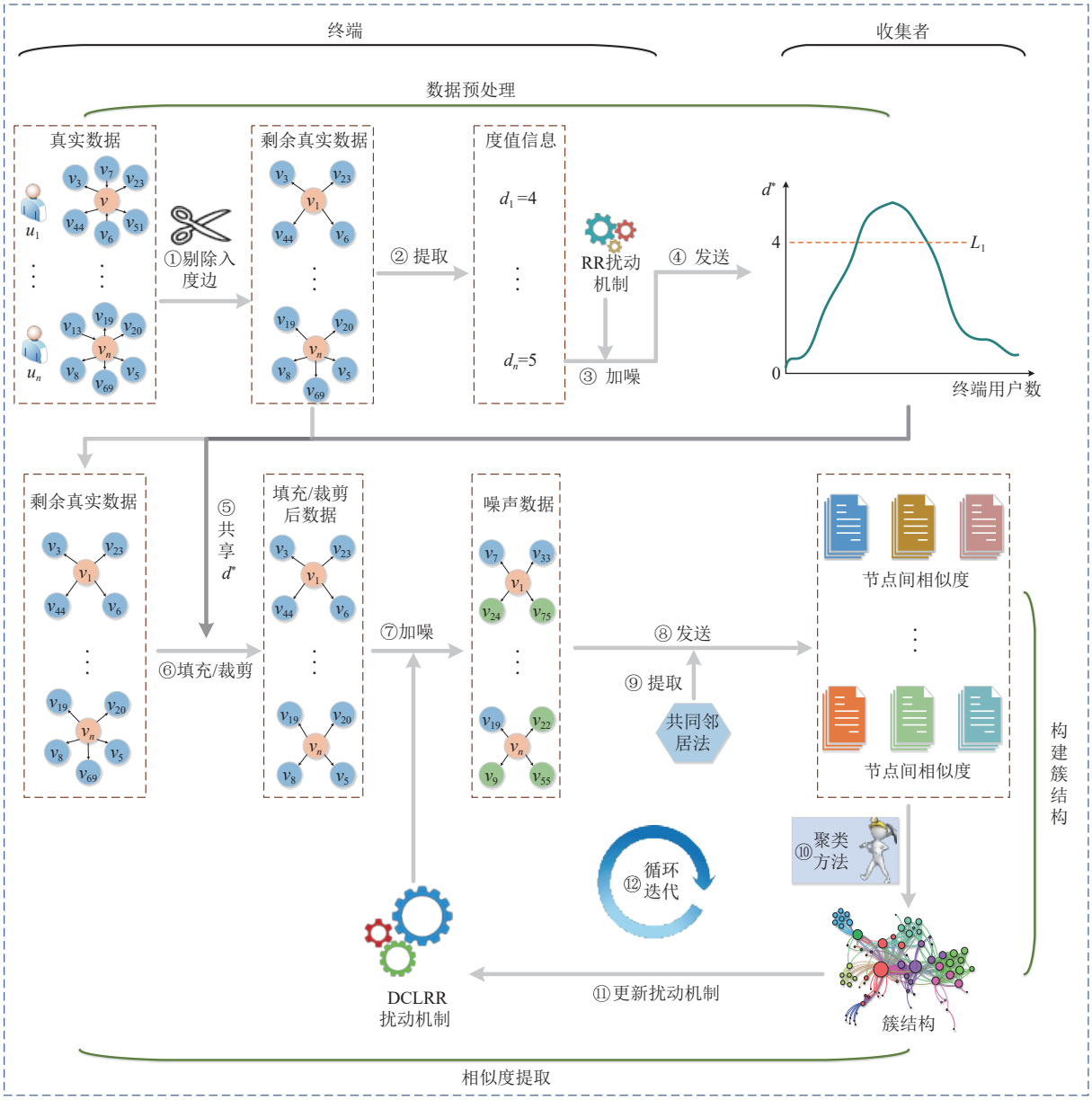


Fig. 1 Illustration of DGC-LDP

图 1 DGC-LDP 示意图

户将自身星型图中的入度边全部剔除(行②). 这样可以将后期隐私处理时面临的多元关系(即边的有无和边的正负方向)变成二元关系(即有无单个方向的边). 之后, 用户使用一元编码对其处理并使用 RR 机制对该信息进行保护(行③④). 收集者在收到用户上传的信息后统计度分布, 并将第 L_1 个百分位数占据的度值d^*共享给所有用户(行⑦). 用户在收到收集者分享的d^*后, 对自身数据进行分割处理. 为了使各个数据在整个算法中拥有同等级别的隐私保护程度, 根据 LDP 的定义, 结合文献 [25] 设计填充-抽样算法, 保证不同终端拥有的信息量(星型图中的边数)是相同的. 具体来说, 将自身数据量$|E_{v_i}|$与参数d^*比较, 如果$|E_{v_i}| > d^*$, 则从E_{v_i}中随机选择$|E_{v_i}| - d^*$条边剔除; 否则, 从哑元集合$\{e_{v_i,\bar{v}_1}, e_{v_i,\bar{v}_2}, \cdots, e_{v_i,\bar{v}_\ell}\}$中抽取对应数量的边数据填充到图$G_{v_i}$中(行 ⑧~⑬).

3.3 相似度计算

节点间相似性指标作为构建簇的一个至关重要因子, 通常由邻居数目来表示. 但每个用户仅有 1 个星图结构, 只能确定该图中心节点的邻居数, 无法完全获得其他节点的邻居信息. 此外, 收集者在不受信任的分布式场景中缺乏用户信息, 加剧了计算节点间邻居数量的挑战. 为了解决这些问题, 我们设计了 SC-LDP(similarity calculation under LDP)算法. 其核心思想是依据节点的部分连接信息构造 2-hop 图, 之后

使用 CN 索引法从图中提取相似度信息. 算法 3 描述了相似度提取的实现框架.

算法 3. SC-LDP 算法.

输入：$G_{v_i}=\left(V_{v_i},E_{v_i}\right),i\in[1,n]$,隐私预算 ε_r，节点集 $V^*=\{v_1,v_2,\cdots,v_n\}$；

输出：相似度集合 S_N.

① $S_N=\{\}$, $C_{\text{all}}=V^*$；

② for each G_{v_i}

③ for each edge $e_{v_i,v_x}\in E_{v_i}$

④ 发送 $y\leftarrow F_{\text{DCLRR}}\left(e_{v_i,v_x},\varepsilon_r/\left|E_{v_i}\right|,C_{\text{all}}\right)$ 到收集者；

⑤ end for

⑥ end for

/*Curator*/

⑦ 为每个节点构建 2-hop 图；

⑧ 使用 CN 计算任意 2 个节点 v_x 和 v_y 间的相似度 $s_{x,y}$；

⑨ $S_N\leftarrow S_N\cup s_{x,y}$；

⑩ return S_N.

SC-LDP 算法以伪代码形式展示了在非可信分布式环境下安全统计节点间相似度的执行过程. 用户使用 DCLRR 机制对星型图中每一条边进行隐私处理，并将处理后的数据上传到收集者（行②~④）. 收集者基于收集到的噪声边信息构建各个节点的 2-hop 图（行⑦），之后使用 CN 索引法统计各个节点间的相似度（行⑧⑨）.

众所周知，在 LDP 下，使用 RR 机制对一个由 n 个位组成的邻接向量表示的星型图进行隐私保护时，其添加的伪边数为 $q(n-L_1)$，其中 q 表示边存在与否的扰动概率. 当所有用户均使用该算法时，则对于合成图来说增加了 $nq(n-L_1)$ 条假边. 然而，采用直接编码方法时，其只对真实存在的边做隐私保护，不对虚拟边进行处理. 所以，对于每个用户来说，形成的伪边数为 qL_1，整张合成图的伪边数为 nqL_1，其远小于 $nq(n-L_1)$. 因此，采用直接编码方式来取代常用的位向量编码.

RR 机制添加的噪声幅度是固定的，继续使用该机制进行簇优化无法使聚簇质量得到更大的提高. 为了解决这个问题，我们针对用户数据形式在直接编码的基础上设计了 DCLRR 动态扰动机制. 该机制基于簇结构通过减少虚假边的产生概率以及增大簇内节点间联系性的概率来实现噪声的自适应添加. 具体来说，首先确定原始边 e_{v_i,v_x} 中节点 v_i 所在的簇 cc. 如果原始边 e_{v_i,v_x} 所连接的 2 个节点 v_i 和 v_x 同属于一个簇 cc. 为了维持已有的簇结构，同时增强簇内节点间的联系，以大概率 p_1 扰动成与原始边同属一个簇内的边 $e_{v_i,v_t}(v_t\in cc)$；反之则以概率 p_2 扰动成任意一条边 $e_{v_i,v_t}(v_t\notin cc)$. 如果原始边 e_{v_i,v_x} 所连接的 2 个节点 v_i 和 v_x 不属于同一个簇 cc. 以大概率 p_3 扰动成真实边 $e_{v_i,v_t}\in E_{v_i}$；反之以概率 p_4 扰动成任意一条边 $e_{v_i,v_t}\notin E_{v_i}$. 为了确保上述扰动满足 LDP，同时最大化数据可用性，根据定义 4 推导出 8 个公式. 为描述方便起见，用 ε_3 来代指隐私预算 $\varepsilon_r/\left|E_{v_i}\right|$.

$$(|cc|-1)p_1+(n-|cc|)p_2=1, \tag{4}$$

$$\left|E_{v_i}\right|p_3+\left(n-\left|E_{v_i}\right|-1\right)p_4=1, \tag{5}$$

$$p_1\leqslant p_2\exp(\varepsilon_3), \tag{6}$$

$$p_3\leqslant p_4\exp(\varepsilon_3), \tag{7}$$

$$p_1\leqslant p_3\exp(\varepsilon_3), \tag{8}$$

$$p_1\leqslant p_4\exp(\varepsilon_3), \tag{9}$$

$$p_2\leqslant p_3\exp(\varepsilon_3), \tag{10}$$

$$p_2\leqslant p_4\exp(\varepsilon_3). \tag{11}$$

通过最优化这些公式，发现概率设置与 $\left|E_{v_i}\right|$ 和 $|cc|$ 这 2 个变量相关. 为了使最终获得概率结果满足 LDP，随后按照这 2 个参数的大小分为 $\left|E_{v_i}\right|+1<|cc|$ 和 $\left|E_{v_i}\right|+1\geqslant|cc|$ 这 2 种情况进行概率的计算. 最终在第 1 种情况下获得概率分别为

$$p_1=\frac{\exp(\varepsilon_3)}{(|cc|-1)\exp(\varepsilon_3)+n-|cc|},$$

$$p_2=\frac{1}{(|cc|-1)\exp(\varepsilon_3)+n-|cc|},$$

$$p_3=\frac{\exp(\varepsilon_3)}{\left|E_{v_i}\right|\exp(\varepsilon_3)+n-1-\left|E_{v_i}\right|},$$

$$p_4=\frac{1}{\left|E_{v_i}\right|\exp(\varepsilon_3)+n-1-\left|E_{v_i}\right|}.$$

在第 2 种情况下获得概率分别为

$$p_1=p_5=\frac{\exp(\varepsilon_3)}{(|cc|-1)\exp(\varepsilon_3)+n-|cc|},$$

$$p_3=p_7=\frac{\exp(\varepsilon_3)}{\exp(\varepsilon_3)+n-2},$$

$$p_2=p_6=\frac{1}{(|cc|-1)\exp(\varepsilon_3)+n-|cc|},$$

$$p_4=p_8=\frac{1}{\exp(\varepsilon_3)+n-2}.$$

DCLRR 扰动机制如算法 4 所示.

算法 4. DCLRR 算法.

输入：隐私预算 $\varepsilon_3=\varepsilon_r/\left|E_{v_i}\right|,C_{\text{all}},e_{v_i,v_x}\in E_{v_i}$；

输出：噪声边 y^*.

① 定位节点 v_i 所属的社区 cc；

② if$\left|E_{v_i}\right|+1<|cc|$

③ if v_i 和 v_x 属于相同的社区 cc

④ $Pr\left[y^*=e_{v_i,v_t}\right]=$

$$\begin{cases} p_1=\dfrac{\exp(\varepsilon_3)}{(|cc|-1)\exp(\varepsilon_3)+n-|cc|}, \text{if } v_t\in cc; \\ p_2=\dfrac{1}{(|cc|-1)\exp(\varepsilon_3)+n-|cc|}, \text{其他}; \end{cases}$$

⑤ else $v_i\in cc$ and $v_x\notin cc$

⑥ $Pr\left[y^*=e_{v_i,v_t}\right]=$

$$\begin{cases} p_3=\dfrac{\exp(\varepsilon_3)}{\left|E_{v_i}\right|\exp(\varepsilon_3)+n-1-\left|E_{v_i}\right|}, \text{if } e_{v_i,v_t}\in E_{v_i}; \\ p_4=\dfrac{1}{\left|E_{v_i}\right|\exp(\varepsilon_3)+n-1-\left|E_{v_i}\right|}, \text{其他}; \end{cases}$$

⑦ end if

⑧ else $\left|E_{v_i}\right|+1\geqslant|cc|$

⑨ if v_i 和 v_x 属于相同的社区 cc

⑩ $Pr\left[y^*=e_{v_i,v_t}\right]=$

$$\begin{cases} p_5=\dfrac{\exp(\varepsilon_3)}{(|cc|-1)\exp(\varepsilon_3)+n-|cc|}, \text{if } v_t\in cc; \\ p_6=\dfrac{1}{(|cc|-1)\exp(\varepsilon_3)+n-|cc|}, \text{其他}; \end{cases}$$

⑪ else $v_i\in cc$ and $v_x\notin cc$

$$\begin{cases} p_7=\dfrac{\exp(\varepsilon_3)}{\exp(\varepsilon_3)+n-2}, \text{if } v_t=v_x; \\ p_8=\dfrac{1}{\exp(\varepsilon_3)+n-2}, \text{其他}; \end{cases}$$

⑬ end if

⑭ end if

⑮ return 噪声边 y^*.

首先每一个终端用户确定原始边e_{v_i,v_x}中节点 v_i 所在的簇 cc(行①). 然后, 基于 $\left|E_{v_i}\right|+1<|cc|$和 $\left|E_{v_i}\right|+1\geqslant|cc|$这 2 种情况使用上述扰动策略对原始数据实施隐私保护(行②~⑬).

3.4 簇结构构建

考虑基于 2-hop 图先生成复合图, 再结合模块化聚合算法实现图聚类. 然而, 通过邻接位向量编码和 RR 扰动技术创建的复合图具有大量伪边, 会严重降低依赖边信息统计节点间相似性的准确度, 进而影响诸如模块化聚类算法这种依赖边信息才能形成簇的聚簇方案. 轮廓系数作为关系型数据中最常用的簇质量度量指标也可以反向用于聚簇. 鉴于此, 我们提出了一种结合相似度和轮廓系数进行聚类的节点聚合算法. 该算法的核心思想是用节点间的相似度代替轮廓系数中涉及的簇对象间的距离信息. 然后, 通过迭代形式融合高相似度的节点形成最终的簇结构. 该方法的具体实现细节如 BC-LDP(build clustering under LDP)算法所示.

算法 5. BC-LDP 算法.

输入: S_N, 节点集 $V^*=\{v_1,v_2,\cdots,v_n\}$, 聚合阈值$\rho$;

输出: 平均轮廓系数m_{sc}以及簇C_{all}.

① $C_{\text{all}}=\theta=\{\}$;

② while $|V^*|>0$

③ 选择 S_N 中最大值对应的 2 个节点形成社区 cc;

④ for each node $v_j\in V^*$

⑤ 计算$a_v=\dfrac{\sum\limits_{v_g\in cc}s_{j,g}}{|cc|}$, $b_v=\dfrac{\sum\limits_{v_z\in V^*,v_z\neq v_j}s_{j,z}}{|V^*|-1}$, $s_v=\dfrac{b_v-a_v}{\max\{a_v,b_v\}}$;

⑥ $\theta\leftarrow\theta\cup s_v$;

⑦ end for

⑧ 选择θ中超过ρ的最大值$c_{\max}$;

⑨ 向 cc 中添加$c_{\max}$对应的节点;

⑩ 重复行 ③~⑨直到无新节点添加为止;

⑪ 从 S_N 中删除 cc 中任意 2 个节点间的相似度;

⑫ 向 C_{all}添加 cc 并从 V^*中删除 cc 中的节点;

⑬ end while

⑭ 计算C_{all}中所有节点的平均轮廓系数m_{sc};

⑮ return m_{sc} and C_{all}.

算法 5 描述了簇的构建步骤. 首先, 收集者将所有节点视为一个个的局部簇. 之后, 将相似度最大的 2 个节点合并在一起, 形成初始簇(行③). 其次, 计算剩余未划入初始簇的节点与初始簇的轮廓系数, 并将轮廓系数值最大的节点并入初始簇(行④~⑨). 再者, 重复行③~⑨, 直到完全找到构建第 1 个簇的所有节点为止(行⑩). 最后, 对剩余未形成簇的节点重复上述步骤, 直至所有节点都划归到某一个簇为止(行②~⑬).

算法 1 概述了 DGC-LDP 算法的实现过程. 第 1 步, 隐私预算被分成 2 份被后边使用(算法 1 的行①); 之后, 利用算法 2 来对用户数据进行预处理, 使得所有用户手中的星型图具有同等规模, 即含有等量的边(算法 2 的行②). 第 2 步, 收集者利用算法 3 获取节点间的相似度信息(算法 3 的行④).第 3 步, 收集者利用算法 5 进行簇的构建(算法 5 的行⑤), 如果所得簇的聚类系数小于目标值, 则重复上述步骤继续构建簇; 否则输出簇结构(算法 5 的行⑥~⑩).

3.5 隐私和效用分析

定理 1. DGC-LDP 算法满足 ε-本地差分隐私.

证明.根据算法 1~5 的描述, DGC-LDP 机制总共使用了 RR 以及 DCLRR 这 2 种隐私机制来防范隐私

泄露. 接下来. 先分析这 2 种机制的隐私性, 进而再证明 DGC-LDP 算法的隐私性.

在本文中, RR 机制仅用于 DPP 算法的度分布计算步骤. 结合 LDP 的定义以及式(4), 在度分布统计步骤中, 对于任意 2 个不同的输入度值, 经 RR 处理后得到相同度值的概率 $p_1 \in \{\exp(-\alpha\varepsilon/2), \exp(\alpha\varepsilon/2), 1\}$. 该概率集合中的元素不超过 $\exp(\alpha\varepsilon)$, 根据 LDP 的定义, RR 机制在这部分满足 $\alpha\varepsilon$-LDP.

接下来证明 DCLRR 机制的隐私性.

对于任意 2 条从同一节点出发指向其他不同节点的边 e_{v_i,v_x} 和 e_{v_i,v_y}, e_{v_i,v_t} 为 DCLRR 机制的任一输出结果. 根据 LDP 的定义, 有

$$\frac{Pr\left[F_{\mathrm{DCLRR}}\left(e_{v_i,v_x}\right)=e_{v_i,v_t}\right]}{Pr\left[F_{\mathrm{DCLRR}}\left(e_{v_i,v_y}\right)=e_{v_i,v_t}\right]}=p^*.$$

根据算法 4 的描述, 可以得到

$$p^* \in \left\{\frac{p_1}{p_1}, \frac{p_1}{p_2}, \frac{p_1}{p_3}, \frac{p_1}{p_4}, \frac{p_2}{p_2}, \frac{p_2}{p_3}, \cdots, \frac{p_4}{p_4}\right\},$$

p^* 中任一元素均不超过 $\varepsilon_3=\varepsilon_r/\left|E_{v_i}\right|$. 因此, 有

$$\frac{Pr\left[F_{\mathrm{DCLRR}}\left(e_{v_i,v_x}\right)=e_{v_i,v_t}\right]}{Pr\left[F_{\mathrm{DCLRR}}\left(e_{v_i,v_y}\right)=e_{v_i,v_t}\right]}\leqslant\exp(\varepsilon_3).$$

综上可知 DCLRR 满足 $\varepsilon_r/\left|E_{v_i}\right|$-LDP.

接下来分析 DGC-LDP 算法的隐私性.

DPP 算法只在度分布计算时使用了 RR 扰动机制, 且该机制已被证明满足 $\alpha\varepsilon$-LDP.因此, DPP 算法根据后置处理性质[21-22,26] 满足 $\alpha\varepsilon$-LDP.

根据算法 3 的行③描述, SC-LDP 算法需要对每位用户的 $\left|E_{v_i}\right|$ 条边做处理. 上述描述已证明对任一条边的描述满足 $\varepsilon_r/\left|E_{v_i}\right|$-LDP.因此, 根据序列组合性(即性质 1)以及后置处理性质[21-22,26] 可得出 SC-LDP 机制满足 ε_r-LDP.

DGC-LDP 算法中只有行②④⑧涉及到隐私算法. 根据上述描述, 行②④满足各自的 LDP.如果不触发行⑥的条件, 则 DGC-LDP 算法满足 $(\alpha\varepsilon+\varepsilon_r/2)$-LDP. 由于 $\alpha\varepsilon+\varepsilon_r/2<\varepsilon$, 继而 DGC-LDP 算法满足 ε-LDP. 如果触发行⑥, DGC-LDP 算法满足 $(\alpha\varepsilon+\varepsilon_r/2+\varepsilon_r/4+\cdots)$-LDP.相似地, DGC-LDP 算法 ε-LDP.

综上分析, DGC-LDP 算法 ε-LDP.

证毕.

定理 2. $Z(d)$ 和 $Z^*(d)$ 分别代表度值为 d 的节点数的真实值和估计值. 给定扰动机制 RR 和 n 个星型子图(即 n 个用户), 可以以 $1-\beta$ 的概率得到

$$|Z(d)-Z^*(d)|=O\left(\frac{2\sqrt{n\mathrm{lb}(1/\beta)}}{\alpha\varepsilon}\right). \tag{12}$$

证明.令 d_i 表示任一星型子图中心点的度值, 其经一元编码之后, 对应度值为 d 的位信息为 $B[d_i]\in\{0,1\}$. 根据 RR 的扰动规则, $B[d_i]$ 以概率 $p=\exp(\alpha\varepsilon/2)/\left(\exp(\alpha\varepsilon/2)+1\right)$ 保持不变, 以概率 $q=1/\left(\exp(\alpha\varepsilon/2)+1\right)$ 取相反值, 即 $\overline{B}[d_i]$. 则: $\left|B[d_i]-\overline{B}[d_i]\right|\leqslant 1$, $\left|\frac{B[d_i]-\overline{B}[d_i]}{p-q}\right|\leqslant\frac{\exp(\alpha\varepsilon/2)+1}{\exp(\alpha\varepsilon/2)-1}$. 进而可得到 $B[d_i]-\overline{B}[d_i]$ 的方差为 $Var\left(B[d_i]-\overline{B}[d_i]\right)=O\left(1/\exp(\alpha\varepsilon/2)\right)$.

根据 RR 机制的统计规则, 任一度值的分布均需所有用户参与, 根据伯恩斯坦不等式可得到:

$$Pr(|Z(d)-Z^*(d)|\geqslant\lambda)\leqslant$$

$$2\exp\left\{-\frac{\lambda^2 n}{2n\sum_{i=1}^{n}Var\left(B[d_i]-\overline{B}[d_i]\right)+\frac{2}{3}\lambda\frac{\exp(\alpha\varepsilon/2)+1}{\exp(\alpha\varepsilon/2)-1}}\right\}=$$

$$2\exp\left\{-\frac{\lambda^2 n}{O\left(1/\exp(\alpha\varepsilon/2)\right)+O\left(1/\exp(\alpha\varepsilon/2)\right)}\right\},$$

其中存在 $\lambda=O\left(\sqrt{n\mathrm{lb}(1/\beta)}/\exp(\alpha\varepsilon/2)\right)$ 使得度值分布统计以 $1-\beta$ 的概率保证 $|Z(d)-Z^*(d)|<\lambda$.

证毕.

定理 3.在迭代构建簇结构过程中, 任意一轮收集者获得的任意 2 个节点间的相似度误差上界为

$$\left|Z\left(s_{x,y}\right)-Z^*\left(s_{x,y}\right)\right|\leqslant O\left(\frac{\sqrt{n\mathrm{lb}(1/\beta)}}{4\varepsilon^*}\right), \tag{13}$$

其中 $Z\left(s_{x,y}\right)$ 和 $Z^*\left(s_{x,y}\right)$ 分别表示任意 2 个节点 v_x 与 v_y 间相似度的真实值和估计值; ε^* 表示每一轮统计花费的隐私预算量.

证明.在第 1 轮统计节点间相似性时, 由于没有簇结构, 根据 DCLRR 机制, 采用的是近似 KRR[27] 的算法来求解的, 即 DCLRR 机制中第 2 类扰动算法中的后一种情况. 在此后各轮的相似性估计中, 采用的扰动机制是在 KRR 机制基础上改进的扰动形式, 即 DCLRR 机制中除第 1 轮使用外的其他情况. 改进后的扰动机制从概率角度分析好于 KRR 算法. 因此, 在给定各轮的隐私预算下, 任何一轮得出的统计误差上界均不超过采用 KRR 方式统计的上界. 因此, 我们可以根据 KRR 方式得到误差上界定义各轮的界限.

对于任一节点 v_z, 令 $\hbar(v_z)\in\{0,1\}, \overline{\hbar}(v_z)\in\{0,1\}$ 分别表示其为节点 v_x 与 v_y 间的一个公共邻居节点的真实计数和估计计数. 其中, $\hbar(v_z)=1$ 表示公共节点, 且真实计数为 1; $\overline{\hbar}(v_z)=1$ 表示公共节点, 且估计数为 1.根据 DCLRR 扰动规则, 给定每一轮的隐私预算 ε^*, $\hbar(v_z)$ 以概率 $p=\frac{\left(p_7^2+2p_7p_8+p_8^2\right)}{2}=\frac{\exp(2\varepsilon^*)}{2\left(\exp(2\varepsilon^*)+n-2\right)^2}$ 保

持不变, 以概率$q=(1-p)/(n-1)$变成非公共邻居节点, 即$\bar{\hbar}(v_z)=0$. 为了方便, 同时也为了统计v_x与v_y间公共邻居节点的个数, 将$\hbar(v_z)-\bar{\hbar}(v_z)$看成一个随机变量, 则$\left|\hbar(v_z)-\bar{\hbar}(v_z)\right|\leqslant 1$. 根据文献[27], 可以计算得到$\hbar(v_z)-\bar{\hbar}(v_z)$的方差为

$$Var\left(\hbar(v_z)-\bar{\hbar}(v_z)\right)=O\left(1/\exp(4\varepsilon^*)\right).$$

令$Z\left(s_{x,y}\right)$和$Z^*\left(s_{x,y}\right)$分别表示任意2个节点v_x与v_y间相似度的真实值和估计值. 根据DCLRR机制的统计规则, 任意2个节点v_x与v_y间的相似度估计涉及到所有用户, 根据伯恩斯坦不等式可得到:

$$Pr\left(\left|Z\left(s_{x,y}\right)-Z^*\left(s_{x,y}\right)\right|\geqslant\eta\right)\leqslant 2\exp\left\{-\frac{\eta^2 n}{2n\sum_{i=1}^{n}Var\left[\hbar(v_v)-\bar{\hbar}(v_z)\right]+\frac{2}{3}\eta}\right\}=2\exp\left\{-\frac{\eta^2 n}{O\left(1/\exp(4\varepsilon^*)\right)+\frac{2}{3}\eta}\right\},$$

其中存在$\eta=O\left(\sqrt{n\mathrm{lb}(1/\beta)}\Big/\exp(4\varepsilon^*)\right)$使得相似度统计以$1-\beta$的概率保证$\left|Z\left(s_{x,y}\right)-Z^*\left(s_{x,y}\right)\right|<\lambda$.

证毕.

4 实验结果与分析

实验平台是Intel® Xero®, 64 GB内存, Win 10系统. 所有算法均采用Python实现. 实验采用3个数据集LastFm[28], Fediverse[28], Government[16]. 其中, LastFm是一个有向网络数据集, 总共有136 409个节点和1 685 524条边.Fediverse也是一个有向数据集, 总共有4 860个节点和484 164条边. Government[16]是关于政府的Facebook页面网络数据集. 该数据集涉及7 057个关于政府信息的网页, 网页之间的边表示它们的共同喜好. 删除原始数据中的自循环边后, 网络中的边总数为89 429.为了满足有向性, 在原始数据上为每一条边随机加入了方向, 以满足实验需要.

结合上述数据集, 为充分衡量各隐私保护聚类算法的聚类质量, 采用ARI(adjusted random index)[29]和AMI(adjusted mutual information)[30]作为聚类评估指标.ARI值或AMI值越大, 表示聚类精度越高.

令V表示n个节点的集合.$P=\{P_1,P_2,\cdots,P_\hbar\}$和$Q=\{Q_1,Q_2,\cdots,Q_\psi\}$是2个不同的簇, 但满足$\bigcup_{i=1}^{\hbar}P_i=\bigcup_{i=1}^{\psi}Q_i=V$. 此外, 令$n_{ij}$表示簇$P_i$和$Q_i$中相同数据的数目, 则ARI以及AMI的定义为:

$$F_{\mathrm{ARI}}(P,Q)=\frac{\sum_{ij}\binom{n_{ij}}{2}-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right]-\left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]\Big/\binom{n}{2}}$$

$$F_{\mathrm{AMI}}(P,Q)=\frac{\sum_{i=1}^{\hbar}\sum_{j=1}^{\psi}\frac{n_{ij}}{n}\mathrm{lb}\left(\frac{nn_{ij}}{a_ib_j}\right)-F_{\mathrm{EMI}}(P,Q)}{\max\{H\{P\},H\{Q\}\}-F_{\mathrm{EMI}}(P,Q)},$$

其中, $H(P)=-\sum_{i=1}^{\hbar}\frac{a_i}{n}\mathrm{lb}\left(\frac{a_i}{n}\right)$, $a_i=\sum_{j=1}^{\psi}n_{ij}$,

$H(Q)=-\sum_{j=1}^{\psi}\frac{b_j}{n}\mathrm{lb}\left(\frac{b_j}{n}\right)$, $b_j=\sum_{i=1}^{\hbar}n_{ij}$以及

$$F_{\mathrm{EMI}}(P,Q)=\sum_{i=1}^{\hbar}\sum_{j=1}^{\psi}\sum_{n_{ij}=\max\{a_i+b_j-n,0\}}^{\min\{a_i,b_j\}}\frac{n_{ij}}{n}\mathrm{lb}\left(\frac{nn_{ij}}{a_ib_j}\right)\times\frac{a_i!b_j!(n-a_i)!\left(n-b_j\right)!}{n!n_{ij}!\left(a_i-n_{ij}\right)!\left(b_j-n_{ij}\right)!\left(n-a_i-b_j+n_{ij}\right)!}.$$

为了评估DGC-LDP算法的性能, 采用ARI和AMI度量N-DGC, LF-GDPR-D, LDPGen[9], LDPCD[16], DGC-R算法的聚簇精度. 其中, N-DGC是DGC-LDP的无隐私保护版; LF-GDPR-D算法是将文献[8]中满足无向图的Edge-LDP的聚簇方案, 通过调整敏感度和扰动概率修改为适用于有向图的聚类算法; DGC-R算法是通过使用RR扰动机制替换DGC-LDP算法中的DCLRR机制得到的.

在本节实验当中, 为了方便观察, 给出的隐私预算为每条边被扰动时所消耗的隐私预算.

实验1. 在Fediverse, Government, LastFm三个数据集上, 评估隐私预算对聚类质量的影响.

实验结果如图2所示. 随着ε的逐渐增加, 4种算法的ARI和AMI均呈现上升趋势, 且DGC-LDP算法得出的效果优于其他3种算法. 这表明隐私保护强度越低、聚类准确性越好. 当$\varepsilon=1$时, DGC-LDP算法的准确性是LF-GDPR-D算法的近4倍, 是LDPGen算法的近5倍. 原因是在有向图中, 使用邻接位向量技术会引入大量噪声边, 降低基于边信息统计节点间相似度的准确性, 进而对聚类质量产生负面影响. 此外, 仅基于度信息聚类无法准确反映节点间的相似性关系.

实验2. 在Fediverse, Government, LastFm三个数据集上, 评估不同参数设置对聚类质量的影响.

实验结果如图3~4所示. 其中, 图3展示了阈值ρ设置对聚簇质量的影响. 在图3实验中, 我们固定隐

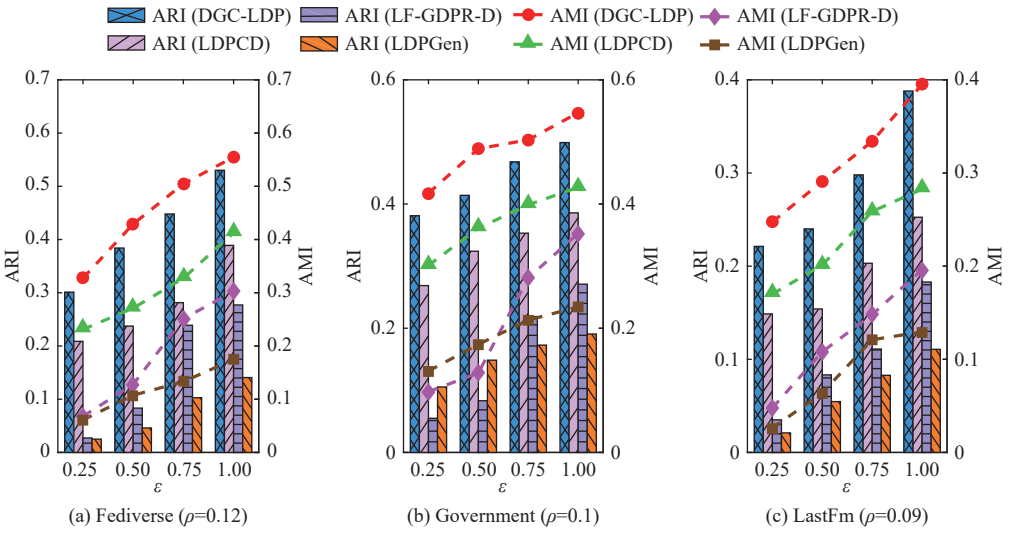


Fig. 2 Effect of privacy budget on cluster quality

图 2 隐私预算对聚簇质量的影响

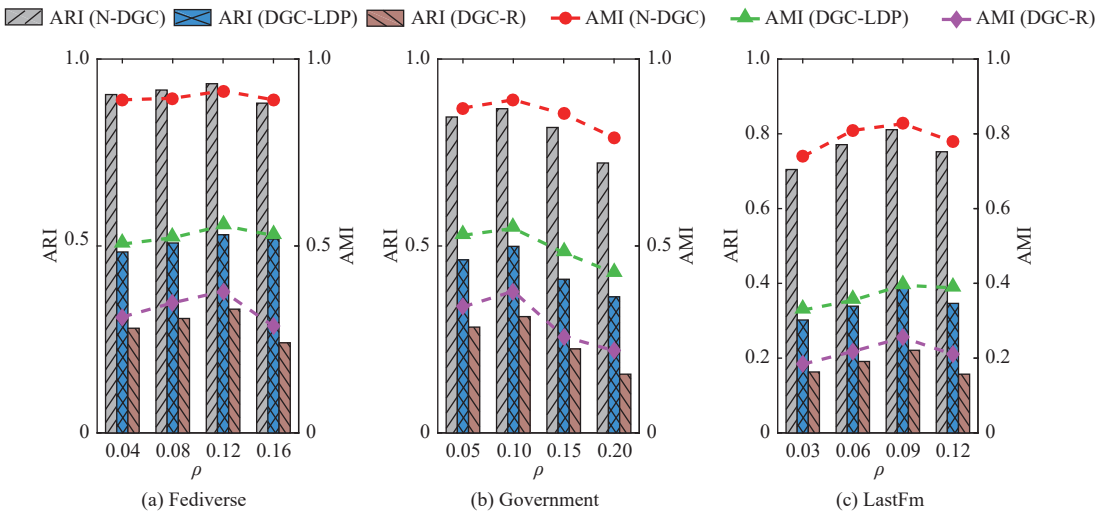


Fig. 3 Effect of threshold ρ on cluster quality

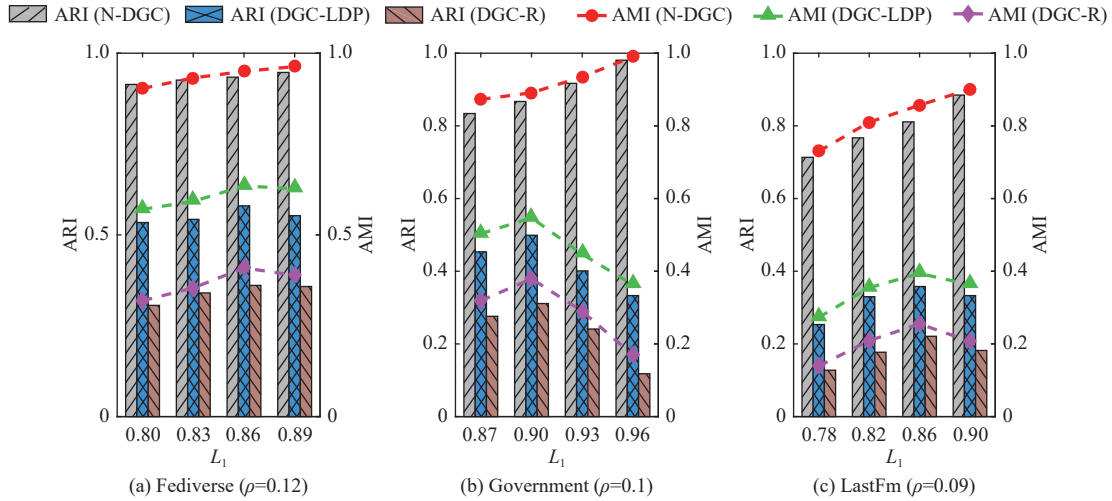
图 3 阈值ρ对聚簇质量的影响

私预算和截断参数 L_1，以准确观察阈值ρ对簇质量的影响. 在图 2 实验中，随着ρ的逐渐增加，3 种算法的 ARI 和 AMI 都是先增大后逐渐减小. 原因是，较小的阈值ρ将生成大量簇，导致属于同一簇的节点被划分为不同的簇，降低了簇的准确性. 同样，较大的阈值ρ将导致原始不属于同一簇中的节点划归在一起，影响簇可用性.

图 4 描述了截断参数 L_1 的设置对聚簇质量的影响. 随着 L_1 取值增大，N-DGC 算法的 ARI 一直呈上升趋势，而其他算法的 ARI 则呈先上升后下降的趋势，并在 L_1=0.86 附近达到峰值. 此外，随着 L_1 的增加，AMI 遵循类似的模式. 原因是，当 L_1 较小时，由于截断会丢失更多信息，从而导致聚类质量降低. 相反，当 L_1 较大时，尽管丢失的信息较少，但聚类需要更多迭代，从而导致大量错误累积，影响聚类准确性. 对于 N-DGC，由于未考虑隐私问题，L_1 长度越长，丢失的信息越少，簇构建得越好.

实验 3. 在 Fediverse, Government, LastFm 三个数据集上，评估计算开销.

为了比较 4 种算法的效率，我们在表 2 中列出了各个算法的运行时间. 可以看出，与 LF-GDPR-D 以及 LDPGen 相比，DGC-LDP 效率确实较低；而与 LDPCD 相比，效率相差很小，甚至在 LastFm 数据集上更优. 这是因为 LF-GDPR-D 和 LDPGen 采用的是单轮模式，而 LDPCD 和 DGC-LDP 采用多轮交互模式. 多轮交互模式需要牺牲效率来换取高准确性. 最



(a) Fediverse (ρ=0.12) (b) Government (ρ=0.1) (c) LastFm (ρ=0.09)

Fig. 4 Effect of L_1 on cluster quality

图 4 L_1 对聚簇质量的影响

后在同样模式下, 我们在效率相差不大情况下获得了更好的挖掘质量, 实现了效率和精度的兼顾.

Table 2 Computation Cost

表 2 计算代价 min

算法	数据集		
	Fediverse	Government	LastFm
LF-GDPR-D	2.88	3.49	15.22
LDPGen	2.06	2.77	5.77
LDPCD	6.17	4.84	28.49
DGC-LDP (本文)	6.88	5.32	26.77

5 结 论

本文针对无属性有向图的隐私保护聚类问题, 提出了一种满足 Edge-LDP 的聚类算法 DGC-LDP, 实现了隐私安全与聚簇质量的兼顾. 采用直接编码方式取代位向量编码方法, 以减少在隐私保护处理时的数据量. 提出基于星型图结构的动态扰动机制, 通过自适应添加噪声来平衡隐私性和统计效用, 并在此基础上, 在终端和收集者间构建迭代机制. 收集者依据终端上传的个体信息, 提取节点间的相似性信息, 并设计基于轮廓系数测量模型的节点聚合算法, 通过迭代机制不断地优化节点聚合形式以形成高质量簇. 理论分析了 DGC-LDP 的隐私性和可用性. 下一步, 将对动态图以及属性图分别进行研究, 在不同场景下实现隐私保护聚类分析.

作者贡献声明: 付楠提出了算法思路和实验方案; 倪巍伟撰写论文; 姜泽鹏完成实验; 侯立贺、张东月、张如玉提出指导意见并修改论文.

参 考 文 献

[1] Duchi J C, Jordan M I, Wainwright M J. Local privacy and statistical minimax rates[C] // Proc of the 54th IEEE Annual Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2013: 429−438

[2] Ye Qingqing, Meng Xiaofeng, Zhu Minjie, et al. Survey on local differential privacy[J]. Journal of Software, 2017, 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, 2017, 29(7): 1981−2005)

[3] Li Weiqing, Chen Hongyu, Zhao Ruifeng, et al. A federated recommendation system based on local differential privacy clustering[C] // Proc of the 7th IEEE Smart World, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Int of People and Smart City Innovation. Piscataway, NJ: IEEE, 2021: 364−369

[4] Liu Gaoyuan, Tang Peng, Hu Chengyu, et al. Multi-dimensional data publishing with local differential privacy[C] //Proc of the 26th Int Conf on Extending Database Technology. Berlin: Springer, 2023: 183−194

[5] Zhang Shaobo, Yuan Liujie, Mao Xinjun, et al. Privacy protection method for K-modes clustering data with local differential privacy[J]. Journal of Acta Electronica Sinica, 2022, 50(9): 2181−2188 (in Chinese)
(张少波, 原刘杰, 毛新军, 等. 基于本地差分隐私的 K-modes 聚类数据隐私保护方法[J]. 电子学报, 2022, 50(9): 2181−2188)

[6] Zhu Suxia, Wang Lei, Sun Guanglu. A perturbation mechanism for classified transformation satisfying local differential privacy [J]. Journal of Computer Research and Development, 2022, 59(2): 430−439 (in Chinese)

（朱素霞，王蕾，孙广路. 满足本地差分隐私的分类变换扰动机制[J]. 计算机研究与发展, 2022, 59(2): 430-439）

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

[8] Ye Qingqing, Hu Haibo, Au M H, et al. Towards locally differentially private generic graph metric estimation[C] // Proc of the 36th IEEE Int Conf on Data Engineering (ICDE). Piscataway, NJ: IEEE, 2020: 1922−1925

[9] Qin Zhan, Yu Ting, Yang Yin, 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

[10] Zhang Yuxuan, Wei Jianghong, Zhang Xiaojian, et al. A two-phase algorithm for generating synthetic graph under local differential privacy[C] //Proc of the 8th Int Conf on Communication and Network Security. New York: ACM, 2018: 84−89

[11] 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

[12] Yang Jingyu, Ma Xuebin, Bai Xiangyu, et al. Graph publishing with local differential privacy for hierarchical social networks[C] // Proc of the 10th IEEE Int Conf on Electronics Information and Emergency Communication (ICEIEC). Piscataway, NJ: IEEE, 2020: 123−126

[13] 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

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

[15] Liu Zichun, Huang Liusheng, Xu Hongli, et al. PrivAG: Analyzing attributed graph data with local differential privacy[C] // Proc of the 26th IEEE Int Conf on Parallel and Distributed Systems (ICPADS). Piscataway, NJ: IEEE, 2020: 422−429

[16] Zhang Zhejian. LDPCD: A novel method for locally differentially private community detection [J]. Computational Intelligence and Neuroscience 2022, 2022: 4080047

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

[19] Shamir A. How to share a secret[J]. Communications of the ACM, 1979, 22(11): 612−613

[20] 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

[21] Wang Teng, Zhang Xuefeng, Feng Jingyu, et al. A comprehensive survey on local differential privacy toward data statistics and analysis[J]. Sensors, 2020, 20(24): 7030

[22] Yang Mengmeng, Lyu Lingjuan, Zhao Jun, et at. Local differential privacy and its applications: A comprehensive survey [J]. arXiv preprint, arXiv: 2008.03686, 2020

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

[24] Erlingsson Ú, Pihur V, Korolova A. Rappor: Randomized aggregatable privacy-preserving ordinal response[C] //Proc of the 21st ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2014: 1054−1067

[25] Zhang Jun, Xiao Xiaokui, Xie Xing. PrivTree: A differentially private algorithm for hierarchical decompositions [C] //Proc of the 2016 ACM Int Conf on Management of Data. New York: ACM, 2016: 155−170

[26] Liu Shang, Cao Yang, Murakami T, et al. A crypto-assisted approach for publishing graph statistics with node local differential privacy [J]. arXiv preprint, arXiv: 2209.02231, 2022

[27] Kairouz P, Bonawitz K, Ramage D. Discrete distribution estimation under local privacy[C] //Proc of the 33rd Int Conf on Machine Learning. New York: ACM, 2016: 2436−2444

[28] Allard A, Serrano M, Boguñá M. Geometric description of clustering in directed networks [J]. arXiv preprint, arXiv: 2302.09055, 2023

[29] Rand W M. Objective criteria for the evaluation of clustering methods [J]. Journal of the American Statistical Association, 1971, 66(336): 846−850

[30] Vinh N X, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, aormalization and correction for chance[J]. Journal of Machine Learning Research, 2010, 11: 2837−2854



Fu Nan, born in 1988. PhD candidate. His main research interests include data mining, privacy protection and machine learning.

付　楠, 1988 年. 博士研究生. 主要研究方向为数据挖掘、隐私保护、机器学习.



Ni Weiwei, born in 1979. PhD, professor, PhD supervisor. His main research interests include database theory, data mining, Web mining, and machine learning.

倪巍伟, 1979 年. 博士, 教授, 博士生导师. 主要研究方向为数据库理论、数据挖掘、Web 挖掘、机器学习.



Jiang Zepeng, born in 2000. Master candidate. His main research interests include data mining, deep learing, and privacy protection.

姜泽鹏, 2000 年. 硕士研究生. 主要研究方向为数据挖掘、深度学习、隐私保护.



Hou Lihe, born in 1996. PhD candidate. His main research interests include data mining, graph analysis, and privacy protection.

侯立贺, 1996 年. 博士研究生. 主要研究方向为数据挖掘、图分析、隐私保护.



Zhang Dongyue, born in 1993. PhD candidate. His main research interests include database theory, data mining, and privacy protection.

张东月, 1993 年. 博士研究生. 主要研究方向为数据库理论、数据挖掘、隐私保护.



Zhang Ruyu, born in 1998. PhD candidate. Her main research interests include privacy protection and machine learning.

张如玉, 1998 年. 博士研究生. 主要研究方向为隐私保护、机器学习.