计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20170886
56(3):508-520, 2019

点差分隐私下图数据的度直方图发布方法

张宇轩[1] 魏江宏[1] 李 霁[1] 刘文芬[2] 胡学先[1]
[1](数学工程与先进计算国家重点实验室(中国人民解放军战略支援部队信息工程大学) 郑州 450001)
[2](广西密码学与信息安全重点实验室(桂林电子科技大学) 广西桂林 541004)
(bigzhangq@163.com)

Graph Degree Histogram Publication Method with Node-Differential Privacy

Zhang Yuxuan[1], Wei Jianghong[1], Li Ji[1], Liu Wenfen[2], and Hu Xuexian[1]
[1](*State Key Laboratory of Mathematical Engineering and Advanced Computing* (*PLA Strategic Support Force Information Engineering University*), *Zhengzhou* 450001)
[2](*Guangxi Key Laboratory of Cryptography and Information Security* (*Guilin University of Electronic Technology*), *Guilin*, *Guangxi* 541004)

Abstract The widespread use of various information systems, e.g. social networks, mail systems and recommendation systems, has produced a large amount of graph data. Publishing and sharing these data under the edge or node differential privacy can fully utilize their potential value, meanwhile, the privacy of the involved users can be preserved. Compared with the edge differential privacy, the node differential privacy can effectively prevent users from being re-identified. However, it will lead to a higher sensitivity of the query function at the same time. To conquer this problem, a novel method named sequence edge-removal (SER) is proposed, based on which two graph degree distribution histogram publication mechanisms under node difference privacy are put forward. The experiment results illustrate that the SER method can effectively suppress the sensitivity of the publishing mechanism, and also can retain more edges of the original graph. In addition, it decreases the errors between the published data and the original data. Compared with available works, under the constraint of providing the same level of privacy preservation, the proposed histogram publishing mechanism based on the SER method can describe the degree distribution of the original data more accurately, and thus improves the usability of the published data.

Key words privacy protection; graph data; differential privacy; degree distribution; histogram publishing

摘 要 社交网络、邮件系统、推荐系统等信息系统的广泛使用产生了大规模的图数据,在点或边差分隐私约束下对这些数据进行发布和共享可以充分发挥其潜在价值,同时又能保证数据中所涉及用户的隐私信息不被泄露.针对点差分隐私定义下查询函数敏感度比较大的问题,提出一种基于度排序的边移除方法(sequence edge-removal, SER),并在此基础上进一步给出了2种点差分隐私下图的度分布直方图发布机制.仿真实验表明:SER方法能有效抑制发布机制的敏感度,保留更多原始图中的边,降低了发布数据与真实数据之间的误差.此外,相比于已有工作,基于SER方法的度直方图发布机制在提供同等隐私保护水平的条件下,更好地刻画了真实数据的度分布,提高了发布数据的可用性.

收稿日期:2017-11-24;**修回日期**:2018-07-03
基金项目:国家自然科学基金项目(61702549,61502527,61862011);广西密码学与信息安全重点实验室开放课题(GCIS201704)
This work was supported by the National Natural Science Foundation of China (61702549, 61502527, 61862011) and the Open Funds of the Guangxi Key Laboratory of Cryptography and Information Security (GCIS201704).
通信作者:刘文芬(liuwenfen@guet.edu.cn)

关键词 隐私保护;图数据;差分隐私;度分布;直方图发布

中图法分类号 TP392

随着互联网和信息技术的飞速发展,许多组织机构搜集的个人数据规模急剧增长,随之而来的用户隐私保护问题变得日益重要.对这些搜集的数据进行深入挖掘和分析后所能得到的潜在社会效益和经济价值,如精准医疗、智慧城市、广告投放等,又驱使这些组织机构倾向于发布关于这些搜集数据的统计信息.另一方面,在隐私保护下的数据分析领域,其主要目标是安全地发布关于数据集的统计信息,同时又不会泄露数据集中所涉及用户的隐私信息,如医疗数据中的病人基本信息、位置数据中的用户运动轨迹等.显然,发布数据集中数据分布的有用信息和保护数据集中的用户隐私这2个目标是相冲突的,如何在这两者之间形成折衷就成了一个十分具有挑战性的问题,而差分隐私技术[1]被认为是解决该问题最为理想的方案之一,可以保证相邻数据集上查询结果的分布很接近.

图数据作为一种典型的数据类型,随着社交网络、推荐系统、协作网络等信息系统的广泛使用而变得越发常见,如何在保证用户隐私不被泄露的前提下发布这些数据也逐渐受到了学者的关注.早期通用的方法主要是对图数据进行简单的匿名化处理来实现用户隐私的保护,如用无意义的字符表示图中节点、删除/添加某些边等等.但是,对于按照这种匿名化方法处理后的图数据,利用主动攻击和被动攻击[2]仍能恢复出原始图中的节点信息和节点之间边的关系,进而导致用户敏感信息泄露.2009年Hay等人[3]首次利用差分隐私技术来实现图数据的安全发布,并提出了差分隐私在图数据发布领域的2种变体,即边差分隐私和点差分隐私.

在边差分隐私中,2个相邻图仅相差1条边,而在点差分隐私中,2个相邻图相差1个节点以及与此节点相连的所有边.对于一个节点数目为$|V|=n$的图$G=(V,E)$(其中V是所有节点的集合、E为所有边的集合),删除1条边只影响这条边上2个节点度的变化,而删除1个节点在最坏情况下会导致$n-1$条边被删除.因此,图数据中的点差分隐私比边差分隐私更难满足,但却能提供更高强度的隐私保护.

度分布是图的一种重要统计特性,也是图数据发布过程中的保护重点.如何在点差分隐私约束下实现图的度分布发布在近年来得到了广泛关注,其主要目标是要在满足点差分隐私的条件下给出一种尽可能接近图的度的真实分布的近似分布.目前,解决该问题的一种主要技术是将原始图投影到1个节点度数不超过最大值θ的压缩图,以此来降低度发布过程中的敏感度,而这其中的关键又在于如何在投影过程中尽可能多地保留原始图的信息.

在TCC 2013上,Kasiviswanathan等人[4]提出利用截断(truncation, T)的方法对原始图进行投影,即除去原始图中所有度大于给定门限值θ的节点.这种方法虽可降低敏感度,但却删除了许多本没必要删除的边,损失了原始图中大量的有效信息.Blocki等人[5]考虑了通过边移除(edge-removal, ER)的方法来降低敏感度,即任意选择一个边序列作为删边的次序,当边的2个顶点存在$deg(v)>\theta$的情况时删除此边,否则保留此边,最后对保留的边组成的图的度分布进行差分隐私发布.Raskhodnikova和Smith[6]基于指数机制,利用流图的方法将图的度分布映射到一个直方图,但却仍具有较高的敏感度6θ.在SIGMOD 2016上,Day等人[7]提出了一种加边机制π_θ来降低敏感度,即首先对原始图的边进行稳定排序(如字典序),然后删除原图中所有的边,只保留节点,再按照稳定排序添加边,当所添加边的2顶点存在$deg(v)>\theta$的情况时跳过此边,最后当所有边被遍历完后对所得到的重构图的度分布进行差分隐私发布.但是,边排序的不确定性导致Day等人[7]提出的方法仍不能最大程度地保留原始图的边信息.

为进一步降低点差分隐私约束下图的度分布发布过程中的敏感度,本文在现有几种投影方法的基础上提出了一种新的投影方法,并基于此给出了2种点差分约束下图的度分布的直方图发布机制.具体而言,本文的主要贡献包括3个方面:

1) 提出了一种基于度排序的边移除(sequence edge-removal, SER)投影方法,通过依节点度的大小删除与该节点相连的边,最终将节点度限制在给定的门限值内.相比已有方法,该方法在降低敏感度的同时,保留了原始图中更多的边,减小了投影图和原始图之间的误差.

2) 基于SER投影方法提出了2种度分布的直方图发布机制,即SER-(θ,Ω)-直方图发布和SER-累积直方图发布,并在理论上证明了这2种度分布发布机制满足点差分隐私的定义.

3) 不同数据集上的仿真测试结果表明,相比已有的点差分隐私约束下图的度分布发布机制,本文提出的基于 SER 投影方法的度分布直方图发布机制在提供同等隐私保护水平的条件下,更好地刻画了真实数据的度分布,提高了发布数据的可用性.

1 相关工作

差分隐私使用扰动技术对数据集或者查询结果添加随机噪声,使得数据或查询结果失真,但又保持了数据整体良好的统计特性和可用性,以便后续的数据挖掘和发布. 具体而言,令 X 是需要发布统计信息的数据集,f 是查询函数,则 ε-差分隐私机制就给出一个扰动的输出 $f(X)+Y$,而不是直接输出真实答案 $f(X)$,这里面 Y 就是所添加的噪声.

自 2006 年 Dwork[1] 提出差分隐私的概念以来,由于其严谨的数学理论保证和可证明隐私安全性,这种隐私保护技术得到了广泛的认可和关注. 相较传统的隐私保护模型(如 k-anonymity[8], l-diversity[9] 等)需要对攻击者的背景知识和攻击模型给出特定的假设,差分隐私不依赖攻击者的特定背景知识,并且可以提供可量化的隐私保证. 近年来,关于差分隐私技术的理论研究[10-13]取得了长足进展,已逐步应用到了数据分析和挖掘的各领域.

1.1 差分隐私约束下的图数据发布

差分隐私技术在图数据发布领域的应用有 2 种变体,即边差分隐私和点差分隐私.

在图的边差分隐私中,2 个相邻图仅相差 1 条边. Nism 等人[14]为降低查询输出中的噪声,提出了局部敏感度的概念和抽取聚合框架,并应用其解决了边差分隐私约束下图中三角形的计数问题; Rastogi 等人[15]通过引入 Bayesian 攻击者,将原有边差分隐私的定义进行了削弱,提出了边差分隐私定义的变体——边对抗隐私(edge adversarial privacy),并基于此给出了可对任何子图进行计数查询的通用算法; Karwa 等人[16]进一步对 Nism 等人[14]的工作进行了扩展,即对图中的 k-stars 子图计数实现了 ε-差分隐私,对 k-triangles 子图计数实现了 (ε,δ)-差分隐私; Zhang 等人[17]提出了用指数机制实现边差分隐私约束下的子图计数的阶梯框架,并在其中给出了定义查询函数含噪输出的概率分布的方法.

在图的点差分隐私中,2 个相邻图相差 1 个节点以及与此节点相连的所有边. 鉴于许多针对图数据的查询函数在(最大)度 θ 较小的图 G^θ 上具有较低的敏感度,因此一般通过将图 G 转换为度上界为 θ 的图 G^θ 来实现节点差分隐私[4-5],即将图 G 中度高于 θ 的节点删除. Chen 和 Zhou[18]针对节点差分隐私提出了一种迭代机制. 给定图 G 和实值函数 f,他们定义了一个具有递归单调性质的实值函数序列 $0=f_0(G)\leqslant f_1(G)\leqslant\cdots\leqslant f_m(G)=f(G)$,再利用这种递归方法实现节点差分隐私约束下的任何类型子图的计数. 但是,构造这样的函数序列 $f_i(G)$ 通常是 NP 困难的,因此如何高效地实现这种机制仍然是一个挑战.

1.2 差分隐私约束下的直方图发布

将数据以直方图的形式表示后能够显著降低查询函数的敏感度,为数据发布带来很大便利[19]. 对于一个具有 N 条记录的直方图,差分隐私机制可用来保护每条记录的频率.

为减少直方图中添加过多噪声所带来的误差, Xu 等人[20]利用最小化查询函数集的均方差的方法给出了 2 种分割策略. Qardaji 等人[21]考虑了直方图上的范围查询函数,以一种分层的方式分割属性值,并将属性值范围分配到一个树上. Hay 等人[22]定义了约束推理来调整查询函数的输出,并提出了 2 种具体的一致性约束. Lin 和 Kifer[23]将有序直方图集合看作 1 个 Markov 链,基于 Bayesian 决策理论给出了一个新的隐私保护的排序直方图问题的估计算法. 张啸剑等人[24]利用 Markov 链蒙特卡洛方法中 Metropolis-Hastings 技术与指数机制,提出了一种有效的排序方法,并对排序后的直方图,给出了一种基于懒散分组下界的自适应贪心聚类方法,进一步减少了发布后直方图中的误差. Lee 等人[25]将直方图中对加噪数据的后处理过程抽象为带约束的极大似然估计问题,并给出了一种快速、通用的解决方法.

2 基础知识

本文主要考虑边和节点均不带标签和权重的无向图 $G=(V,E)$,并用 $deg(i)$ 表示节点 i 的度, $hist(G)$ 表示图 G 的度直方图.

2.1 差分隐私基本概念

差分隐私技术的思想源自一个很简单的观察:对于给定的包含个体 U 的数据集 S,进行任何查询操作 f(例如计数、求和、平均值等)后所得到的结果为 $f(S)$,如果将 U 的信息从 S 中移除,查询结果仍

然为 $f(S)$或与 $f(S)$十分接近,就可以认为 S 没有额外地泄漏 U 的信息.而差分隐私技术就是要保证无论个体数据在或不在数据集中,对最终的查询结果都没有显著影响.基于这种思想,ε-差分隐私的严格定义如下:

定义 1[1]. ε-差分隐私.若随机算法 K 对任意一对相邻数据集 D, D'及任意输出 $S\subseteq range(K)$均满足:

$$Pr[K(D)\in S]\leqslant \exp(\varepsilon)\times Pr[K(D')\in S],$$

则称算法 K 满足ε-差分隐私.

在定义 1 中,相邻数据集 D 和D'是指这 2 个数据集仅相差 1 条数据记录,即 $|D/D'|=1$,并用 $D\simeq D'$表示这种相邻关系.可以看出,差分隐私的严格数学定义保证了无论数据样本是否存在于数据集 D 中,算法输出的概率分布几乎不变,而隐私系数 ε 的大小则反映了隐私保护程度的强弱,即 ε 的值越小,算法在相邻数据集上的输出的概率分布就越相近,提供更高强度的隐私保护,同时算法输出的可用性也会越低.

2.2 噪声机制

噪声机制是实现差分隐私的主要技术,即在算法输出中加入一定量的噪声,而所添加噪声的多少依赖于具体的噪声机制和查询函数的全局敏感度.Dwork 等人[26]提出的 Laplace 机制是目前应用最广泛的差分隐私机制,其通过向算法输出中添加 Laplace 噪声实现差分隐私.

定义 2[26]. 全局敏感度.对于任意 1 个实值查询函数 f 和相邻数据集 D,D',查询函数 f 的全局敏感度定义为

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

其中,$\|f(D)-f(D')\|_1$ 为查询输出 $f(D)$和 $f(D')$之间的 1-阶范数距离.

定义 3[26]. Laplace 机制.对于给定的数据集 D 和实值查询函数 f,令 Δf 为 f 在数据集 D 上的全局敏感度,则随机算法 K:$K(D)=f(D)+Y$ 满足ε-差分隐私,其中 $Y\sim Lap(\Delta f/\varepsilon)$是加入的随机噪声量,服从尺度参数值为 $b=\Delta f/\varepsilon$ 的 Laplace 分布.

在定义 3 中,Laplace 机制的概率密度函数为

$$Lap(x,b)=\frac{1}{2b}\exp\left(\frac{-|x|}{b}\right).$$

Laplace 机制通过添加 Laplace 噪声实现差分隐私,要求查询函数 f 的输出必须是实数,这在一定程度上限制了其应用.为此,Mcsherry 和 Talwar[27]提出了指数机制,采用满足特定分布的随机抽样来代替添加噪声,以此来实现差分隐私,从而使得指数机制具有更加广泛的应用范围.

粗略来讲,指数机制定义了效用函数 q,使得每一种输出方案都对应着 1 个由 q 刻画的效用函数值,而效用函数值最大的输出方案有更大的概率被采用.

定义 4[27]. 指数机制.对于给定的数据集 D,令 q 是评估数据集 D 上所有输出方案的效用函数,如果算法 K 满足输出为 r 的概率与 $\exp(\varepsilon q(D,r)/2\Delta q)$呈线性关系,则算法 K 满足ε-差分隐私,其中 Δq 为效用函数 q 的敏感度:

$$\Delta q=\max_{D\simeq D',r}\|q(D,r)-q(D',r)\|_1.$$

2.3 差分隐私的组合性质

对于一个复杂且困难的隐私保护问题,应根据数据集自身特性多次、分步骤地使用差分隐私机制,进而使得问题得到有效解决,但是总的隐私预算 ε 却是有限的.因此,为了将整体隐私预算控制在 ε 内,需要对整个过程的每步合理地分配隐私预算,而这就需要使用差分隐私机制的 2 个组合性质:

引理 1[28]. 序列组合性.给定 n 个随机算法 $\{A_i\}_{1\leqslant i\leqslant n}$,其中,$A_i$ 满足 ε_i-差分隐私,则 $\{A_i\}_{1\leqslant i\leqslant n}$ 按指定顺序组合后的算法满足 $\sum_{i=1}^{n}\varepsilon_i$- 差分隐私.

引理 2[28]. 并行组合性.给定 n 个随机算法 $\{A_i\}_{1\leqslant i\leqslant n}$,其中,$A_i$ 满足 ε_i-差分隐私,且任意 2 个算法的输入数据集是不相交的,则由这些算法所构成的新算法$(A_1(D_1),A_2(D_2),\cdots,A_n(D_n))$满足 $\max_{1\leqslant i\leqslant n}\{\varepsilon_i\}$-差分隐私.

2.4 度量方法

为评估不同图的度分布发布机制的性能,本文采用文献[4,6]中所用的 L_1 误差和文献[3]中所用的 KS-距离(Kolmogorov-Smirnov 距离)2 个量化指标.具体来讲,对于 2 个长度为 M 的度分布 D 和 D',它们之间的 L_1 误差可由 $L_1=\sum_{i=0}^{M-1}|D_i-D'_i|$ 得到,而对于长度比 M 小的度分布,为便于比较,将其用 0 填充到长度为 M.

2 个度分布之间的 KS-距离从另一个角度刻画了它们的接近度.对于度分布 D 和 D',它们之间的 KS-距离定义为

$$KS(D,D')=\max_i|CDF_D(i)-CDF_{D'}(i)|,$$

其中,$CDF_D(i)$为分布 D 上度为 i 的累积分布函数值.

3 基于度排序的边移除投影方法

为降低度分布发布过程中的敏感度,提高发布后数据的可用性,本节提出一种新的图投影方法,即基于度排序的边移除投影方法(SER). 粗略来讲,该方法按照度的大小依次删除图 $G=(V,E)$ 中与度数较大的节点相连的边,最终将图中每个节点的度限制到给定的门限值 θ 之内,同时又使得 G 中原有的边能最大程度地保留,为差分隐私机制在压缩图中的应用提供基础. *SER* 投影方法的细节以算法的形式给出:

算法 1. 基于度排序的边移除算法.

输入:图 $G(V,E)$、度限制 θ;

输出:限制图 $SER_\theta(G)$.

① *sorted_l*; /* 计算 $deg(i)$,按从大到小排序 $[i,deg(i)]$ */

② while max$deg(i)>\theta$ do

③ *sort_list*; /* 找 i 的所有相邻节点 j 按 $deg(j)$从大到小排序$[j, deg(j)]$ */

④ while $deg(i)>\theta$ do

⑤ $deg(i)= deg(i)-1$;

⑥ $deg(j)=deg(j)-1$; /* j 遍历 *Sort_list*,直到 $deg(i) =\theta$ */

⑦ end while

⑧ *sorted_l*; /* 对 *sorted_l* 按 $deg(i)$从大到小重排序 */

⑨ end while

⑩ return $G^\theta(V,E^\theta)$.

该算法首先计算图 $G=(V,E)$ 中所有节点的度并按照从大到小排序;然后找出度最大的节点 i,将其相邻节点按度从大到小进行排序,并按照此序列对与节点 i 相连的边进行删边处理,直到 $deg(i)=\theta$ 时结束本次运算;对所有节点按照度从大到小进行重新排序,重复进行上述操作,直至所有节点都满足条件,结束算法.

为了更直观地说明算法 1 的流程,图 1 给出了相关 3 种图投影方法的例子,门限值 $\theta=2$,其中 ER 使用的边排序为随机序列,我们设边排序为:10,9,8,7,6,5,4,3,2,1;π_θ 使用的边序列为字典序:1,2,3,4,5,6,7,8,9,10. 可以看出,使用文献[5]中的 ER 方法时,我们按照假设的随机边序列进行减边,当存在构成边的顶点的度大于 θ 时,我们删除此边,遍历边序列,我们可知最后能保留 4 条边;使用文献[7]中的 π_θ 方法时,我们首先删除所有的边只保留节点,然后按照边序列进行加边,当存在构成边的顶点的度大于 θ 时,我们跳过此边,遍历边序列,我们可知最后能保留 5 条边;而使用我们提出的 SER 投影方法时,首先对节点按照度大小进行排序,找出度最大的节点 d,然后再对 d 的相邻节点进行排序,当 $deg(d)>\theta$ 时,对 d 按照相邻节点序列进行删边,直到 $deg(d)=\theta$ 时,结束此次计算,并对节点按照度大小再次进行排序,依照上述方法计算,直到所有节点的度都小于等于 θ 时,结束算法. 通过 SER 投影方法可知,最后能保留 6 条边.

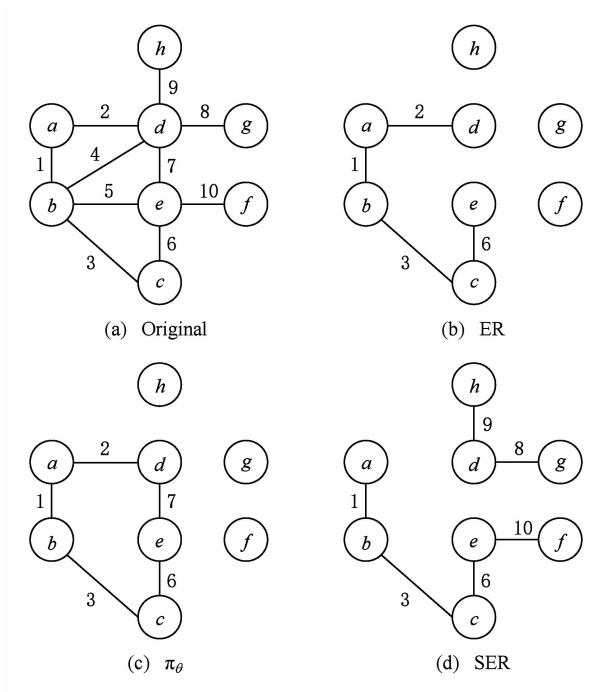


Fig. 1 ER, π_θ, SER method example

图 1 ER, π_θ, SER 方法举例

从图 1 简单例子可以看出,ER 方法对要删除的边进行随机排序,这虽然保证了算法的运行效率,但却损失了很多原始图的边信息;π_θ 这种方法从形式上看似是在增加边,但其本质仍还是删除多余的边,进而把图中节点的度限制在给定的门限值内,同样没有对要删除的边做一定规则的排序;而我们提出的 SER 投影方法,规定了边的排序规则,在限制度的前提下,更多的保留了原始图中的边,减小了投影图和原始图之间的误差,从而提高了差分隐私保护后的数据可用性.

事实上,无论采取何种投影方法,最终目的都是要使得图 1 中所有节点的度小于给定的门限值 θ. 基于此,我们可以将图数据中的节点简单地分成 2 类,即节点度 $deg>\theta$ 和 $deg\leqslant\theta$. 图 2 给出了这 2 类节

点在图数据中的所有连接方式,在图 2 中边依据其顶点度与 θ 之间的大小关系也相应地分为Ⅰ,Ⅱ,Ⅲ三类.为了尽可能多地保留原始图中的边,我们希望在降低节点度的过程中所删除的全都是Ⅰ类边.但是,实际应用中的图数据很难满足这种理想情况,此时我们就需要去删除Ⅱ类边.在这种情形下,如果不对要删除的边进行排序,就会在没有删除完Ⅰ类边的情况下去删除Ⅱ类边,造成不必要的信息损失.而本文所设计的算法 1,就是在尽可能地把Ⅰ类边删除完的情况下再去删除Ⅱ类边,从而实现尽可能多地保留原始图中边的目的.

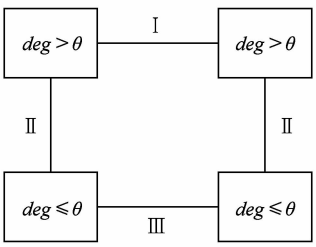


Fig. 2 Two nodes of connection in graph

图 2 2 类节点在图中的连接方式

从上述分析可以看出,本文所提出的基于度排序的边移除算法已经最大程度上逼近了原始图中所能保留边的最大数目.此外,该算法通过不断地更新度排序,在算法运行效率和运行结果之间实现了较好地平衡.下面我们通过定理 1 说明,图 $G=(V,E)$ 在 SER 投影方法处理后,所得到压缩图 $SER(G)$ 的直方图 $hist(SER(G))$ 的全局敏感度 Δ_{hist} 被限定在 $2\theta+1$ 的范围内.

定理 1. 对任意 2 个仅相差 1 个节点的图 G 和 G'(相邻图),在经过 SER 投影方法所生成的度直方图的敏感度为

$$\|hist(SER_\theta(G))-hist(SER_\theta(G'))\|_1\leqslant 2\theta+1.$$

证明. 假设图 $G'(V',E')$ 比图 $G(V,E)$ 多了 1 个节点 $\hat{v}$,即 $V'=V\cup\{\hat{v}\}$,$E'=E\cup\hat{E}$,其中 $\hat{E}$ 为 E' 中和 $\hat{v}$ 相连的边的集合.$SER_\theta(G')$ 表示用 SER 投影方法对图 G' 进行投影后得到的压缩图.从度大小的角度考虑,节点 $\hat{v}$ 存在 3 种情况,即 $deg(\hat{v})>\theta$,$deg(\hat{v})=\theta$ 和 $deg(\hat{v})<\theta$.如图 2 所示,由 SER 投影方法的排序规则可知,在度直方图中当 $deg(\hat{v})\leqslant\theta$ 时,$\hat{v}$ 只作为相邻节点进行排序,对敏感度的影响没有当 $deg(\hat{v})>\theta$ 时大,故只需考虑 $deg(\hat{v})>\theta$ 的情况即可.

由算法 1 可知,我们考虑 2 种情况,即 $\hat{v}$ 是图 G' 中度最大的节点和 $\hat{v}$ 不是图 G' 中度最大的节点.当 $\hat{v}$ 是图 G' 中度最大的节点时,首先对 $\hat{v}$ 进行删边处理,直到 $deg(\hat{v})=\theta$ 时停止对 $\hat{v}$ 的操作,进行下一个点的删边处理,直到算法结束;当 $\hat{v}$ 不是图 G' 中度最大的节点时,首先对度排序在 $\hat{v}$ 前面的节点进行删边处理,此时可能会删除和 $\hat{v}$ 相连的边,若此时 $\hat{v}$ 已经满足 $deg(\hat{v})=\theta$,则后续就不再对 $\hat{v}$ 进行处理,后之则按照算法 1 把 $\hat{v}$ 处理为 $deg(\hat{v})=\theta$,直到算法结束.由上述可知在 $SER_\theta(G')$ 中 $\hat{v}$ 至多和 θ 个节点相连,则和 $SER_\theta(G)$ 相比有 θ 个节点存在不同的度,这些节点在直方图中引起的误差至多为 2θ,再加上 $\hat{v}$ 本身在直方图中引起的误差至多为 1,由定义 2 可知,$SER_\theta(G)$ 和 $SER_\theta(G')$ 的度直方图全局敏感度至多为 $2\theta+1$. 证毕.

4 基于 SER 的点差分隐私度直方图发布

在文献[29]中,最早提出了把聚合的思想应用到直方图中来减少误差;Day 等人[7]通过结合增边的图投影方法 π_θ 和聚合的思想,提出了 2 种高效的点差分隐私度直方图发布方法,即 (θ,Ω)-直方图发布和累计度直方图发布.在本节,我们基于第 3 节所提出的 SER 投影方法和文献[7]中的直方图发布机制,给出 2 种新的点差分隐私下度分布发布机制,并证明其全局敏感度的大小和隐私安全性.

4.1 SER-(θ,Ω)-直方图

在 SER-(θ,Ω)-直方图发布方法中,先用预先设置好的聚合规则 Ω 和 Θ 值得出一个由 (θ_i,Ω_j) 组成的候选集 T,并计算质量函数,然后通过指数机制选择 (θ_i,Ω_j),再运用提出的 SER 投影方法和度门限值 θ_i 对原始图进行压缩,最后用选择的聚合规则 Ω_j 对压缩图的度直方图进行聚合并加噪,并做尾部处理.在我们的仿真实验中,Θ 的值为[1,100]中的整数,聚合规则 Ω 在文献[7]中给出.

在第 3 节我们已经证明了 SER 投影方法在直方图中的全局敏感度为 $2\theta+1$,则其质量函数为

$$q_\varepsilon^{\text{h}}(G,\theta_i,\Omega_j)=-e_{\text{h}}(G,\theta_i,\Omega_j)-e_{\text{p}}(G,\theta_i),$$

其中,e_{h} 为创建直方图过程中引入的误差:

$$e_{\text{h}}(G,\theta_i,\Omega_j)=\sum_{k_m\in\Omega_j}\sum_{d\in k_m}\left(\left|s_d-\sum_{d\in k_m}s_d/|k_m|\right|+\frac{2\theta_i+1}{\varepsilon|k_m|}\right),$$

其中,$\left|s_d-\sum_{d\in k_m}s_d/|k_m|\right|$ 为各个桶取平均值所造成的误差,s_d 为分组 k_m 中每一个桶的值,$(2\theta_i+1)/\varepsilon|k_m|$ 为添加拉普拉斯噪声造成的误差.此外,e_{p} 为投影过程所造成的误差:

$$e_{\text{p}}(G,\theta_i)=2\times|\{v|v\in V,deg_{SER_\Theta(G)}(v)>\theta_i\}|,$$

$deg_{SER_\Theta(G)}(v)$ 为节点 v 在 $SER_\Theta(G)$ 中的度.

在上述参数约束下,所用指数机制中的全局敏感度为 $6\Theta+4$[7]. 进而,基于 SER 投影方法的点差分隐私(θ,Ω)-直方图发布机制可通过算法 2 给出,其中 $\varepsilon_1,\varepsilon_2$ 为进行差分隐私机制中所需隐私预算.

算法 2. SER-(θ,Ω)-直方图发布算法.

输入:图 $G(V,E)$、隐私预算 $\varepsilon_1,\varepsilon_2$、候选集 T;

输出:加噪后的度分布 D.

① 计算每个 $\theta_i\in T$ 的质量函数的值 $q_\varepsilon^{\mathrm{h}}(G,\theta_i,\Omega_j,\varepsilon_2)$;

② 通过指数机制 $\exp(\varepsilon_1 q_\varepsilon^{\mathrm{h}}(G,\theta_i,\Omega_j,\varepsilon_2)/2\Delta_{q_\varepsilon^{\mathrm{h}}})$ 选择(θ^*,Ω^*);

③ 通过算法 1 来投影图 G,得到 G^{θ^*};

④ $h=(hist(G^{\theta^*}))_{\Omega^*}+Lap\left(\frac{\Delta_{\mathrm{hist}}}{\varepsilon_2}\right)^{|\Omega^*|}$;

/* 对聚合后的直方图加拉普拉斯噪声 */

⑤ 通过算法 5 对 h 做尾部处理,得到 h';

⑥ return $D=h'/\|h'\|_1$.

该算法中我们首先计算每个候选组(θ_i,Ω_j)所对应的质量函数,其中 θ_i 遍历[1,100]中的整数. 然后通过指数机制来选择候选组,并用候选组中的 θ_i 对输入图 G 进行压缩得到 G^{θ^*}. 最后,用聚合规则 Ω^* 对 G^{θ^*} 的度直方图进行合并,用拉普拉斯机制对合并后的直方图加噪. 由于算法 1 的原因,直方图在靠近 θ_i 的部分会明显高于相邻桶,不符合一般度分布的规律,所以我们在第 5 步设计了尾部处理(这一部分将在 4.3 节中详细介绍).

4.2 SER-累积度直方图

在 SER-累计度直方图发布机制中,第 k 个桶的计数为 $cumhist_k=|v:deg(v)\leqslant k|$,即图中度不超过 k 的节点的个数. 下面我们通过定理 2 证明在累积度直方图发布机制中,SER 投影方法的全局敏感度 $\Delta_{\mathrm{cumhist}}$ 为 $\theta+1$.

定理 2. 对任意 2 个仅相差一个节点的图 G 和 G'(相邻图),在经过 SER 后所生成的累积度直方图的敏感度为

$$\|cumhist(SER_\theta(G))-cumhist(SER_\theta(G'))\|_1\leqslant\theta+1.$$

证明. 沿用定理 1 证明中的符号. 对节点 v 分为 2 种情况进行讨论,即 $deg(v)\geqslant\theta$ 和 $deg(v)<\theta$. 在图 G'中,当 $deg(v)\geqslant\theta$ 时,$SER_\theta(G')$中 v 至多和 θ 个节点相连,而节点 v 本身只能影响累积度直方图中最后一个桶计数,和 $cumhist(SER_\theta(G))$相比只差 1;相邻的 θ 个节点对累积直方图的影响最大为 θ,根据定义 2 可知,在累积直方图中的 $SER_\theta(G)$和 $SER_\theta(G')$的全局敏感度为 $\theta+1$. 当 $deg(v)<\theta$ 时,令 $t=deg(v)$,$SER_\theta(G')$中 v 至多和 t 个节点相连,故在 $cumhist(SER_\theta(G'))$中 t 到 θ 的 $\theta-t+1$ 个桶中,每个桶和 $cumhist(SER_\theta(G))$相比都相差 1,则最多相差 $\theta-t+1$;相邻的 t 个节点对累积直方图的影响最大为 t,根据定义 2 可知,在累积度直方图中 $SER_\theta(G)$和 $SER_\theta(G')$的全局敏感度为 $\theta+1$. 综上所述,累积度直方图中 $SER_\theta(G)$和 $SER_\theta(G')$的全局敏感度为 $\theta+1$. 证毕.

在界定 SER 投影方法在累积直方图发布机制中的全局敏感度之后,则满足累积度直方图的质量函数可定义为

$$q_\varepsilon^{\mathrm{c}}=-2\times|\{v|v\in V,deg_{SER_\Theta(G)}(v)>\theta_i\}|-\sqrt{\theta}\times\frac{\theta+1}{\varepsilon},$$

而用指数机制选择参数时的全局敏感度为 $2\Theta+2$. 因此,基于 SER 投影方法的累积度直方图发布机制如算法 3 所示:

算法 3. SER-累积度直方图发布算法.

输入:图 $G(V,E)$、隐私预算 $\varepsilon_1,\varepsilon_2$、候选集 T;

输出:加噪后的度分布 D.

① 计算每个 $\theta_i\in T$ 的质量函数的值 $q_\varepsilon^{\mathrm{c}}(G,\theta_i,\varepsilon_2)$;

② 通过指数机制 $\exp\left(\frac{\varepsilon_1 q_\varepsilon^{\mathrm{c}}(G,\theta_i,\varepsilon_2)}{2\Delta_{q_\varepsilon^{\mathrm{c}}}}\right)$选择 θ^*

③ 通过算法 1 来投影图 G,得到 G^{θ^*};

④ $ch=cumhist(G^{\theta^*})+Lap\left(\frac{\Delta_{\mathrm{cumhist}}}{\varepsilon_2}\right)^{\theta^*}$;

/* 对累积度直方图加拉普拉斯噪声 */

⑤ 通过算法 4 对 ch 做拆分处理,得到 h;

⑥ 通过算法 5 对 h 做尾部处理;

⑦ return $D=h'/\|h'\|_1$.

该算法中我们首先计算每个候选组 θ_i 所对应的质量函数,其中 θ_i 遍历[1,100]中的整数. 然后通过指数机制来选择候选组,并用算法 1 对输入图 G 进行压缩得到 G^{θ^*}. 最后,对 G^{θ^*} 的累积直方图用拉普拉斯机制加噪. 由于我们最后得到是累积度直方图,为了方便实验对比,我们还要用算法 4(下面会详细介绍)对直方图进行拆分. 同算法 2,对拆分后的直方图进行尾部处理.

累积度直方图除了具有添加噪声少的优点外,同时还具有单调性,即直方图中桶的值是递增的. 基于此,我们把累积直方图转化为普通直方图,设计了一种校准直方图的算法,对发布结果进行调整,同时也方便进行对照实验. 在算法 4 中,如果前一个桶比

后一个桶小,则直接用差值作为当前桶的计数.但是,由于噪声的破坏,有可能会出现前一个桶比后一个桶大的情况,这时就需要在直方图桶 i 到 θ 中找到 1 个大于桶 i 的桶 j,在桶 i 到桶 j 中均匀地分配计数(行⑤~⑭).

算法 4. 提取累积直方图.

输入:界限为 θ 噪声累积直方图 ch;

输出:界限为 θ 度直方图 h.

① $ch_0=0, i=1$;
② if $ch_1<0$ then
③ $ch_1=0$;
④ end if
⑤ while $i\leqslant\theta$ do
⑥ if $ch_i<ch_{i+1}$ then
⑦ $h_i=ch_i-ch_{i-1}$;
⑧ $i=i+1$;
⑨ else
⑩ $j\in[i,\theta]$,从 i 到 θ 遍历集合找到第 1 个使得 $ch_i<ch_j$ 的 j;
⑪ $h_k=\frac{ch_j-ch_{i-1}}{j-i+1}(i\leqslant k\leqslant j)$;
⑫ $i=j+1$;
⑬ end if
⑭ end while
⑮ return h.

4.3 直方图尾部处理

通过观察原始图可以发现,度分布一般遵循长尾分布,低度节点的计数通常较大,高度节点的计数通常较小且作出的直方图类似于长尾.但是,经过投影后的图的度分布却与此不符:在度为 θ 的节点周围的计数很大.这就导致最后发布的度分布与原始分布有较大的差异,并且当噪声不够大时,很有可能造成隐私泄露.事实上,这是由于所设计投影算法把度大于 θ 的节点全部投影在了度等于 θ 的节点周围,进而导致 θ 周围桶的计数过大.针对此类问题,一般采用基于线性回归的尾部处理方案,即通过直方图中除去 θ 的后半部分进行学习,得出线性回归的斜率 k 和截距 b,然后对加噪后的分布进行处理.

结合所提 SER 投影方法的细节,下面给出适用于基于 SER 投影方法的 2 种度分布发布机制的尾部处理方法:

算法 5. 基于线性回归的尾部处理.

输入:直方图 $H=\{h_1,h_2,\cdots,h_\theta\}, n=|V|$;

输出:处理过的直方图 h.

① $H'=\{h_{\theta/2},h_{\theta/2+1},\cdots,h_{\theta-1},h_\theta\}$;
② 根据 H' 拟合出二次函数 F,并找出二次函数的拐点 j;
③ $budget=sum([h_{j+1},h_{j+2}\cdots,h_\theta])$; /* 设直方图中度从拐点 j 开始到 θ 的计数为预算 */
④ $H=\{h_{\theta/2},h_{\theta/2+1},\cdots,h_j\}$; /* 取直方图中除去预算的部分作为回归学习的样本 */
⑤ $c=\frac{2}{\theta}\sum_{k=2/\theta}^{j}h_k$;
⑥ 根据 H 学习出线性回归的斜率 k 和截距 b.
⑦ for $i=j$ to n
⑧ if $k<0$ then
⑨ $h_i=k\times i+b$;
⑩ else
⑪ $h_i=c$;
⑫ end if
⑬ $budget=budget-h_i$; /* 一般得出的斜率应为正值,可由于加噪的缘故,得出的可能为负,这时我们就取后半部分平均值作为这个桶的计数 */
⑭ if $budget<0$ then
⑮ break
⑯ end if
⑰ end for
⑱ return h.

算法 5 的主要思想根据前半段符合长尾分布的直方图学习出线性回归的斜率和截距,然后把靠近 θ 的异常桶计数按照学习出的参数进行分配.此过程扩展了直方图的横轴,使得经过差分隐私后的度直方图更加符合原始图的分布.

4.4 隐私安全性证明

我们需要说明基于 SER 投影方法的 (θ,Ω)-直方图和累积直方图的隐私安全性并通过定理给出.

定理 3. 基于 SER 投影方法的 (θ,Ω)-直方图发布机制满足 $(\varepsilon_1+\varepsilon_2)$-点差分隐私.

证明. 在算法 2 中,行①②计算质量函数并通过指数机制来选择参数.相当于 n 个查询,这 n 个查询记为 Q_1.由 4.1 节可知,删除图 G 中的 1 个节点,最多影响图 G 中的 $6\Theta+4$ 个节点的度,则 $\Delta Q_1=6\Theta+4$.根据定义 4 可知行①②满足 ε_1-差分隐私.行③④对压缩图的聚合直方图利用拉普拉斯机制添加噪声.由于聚合组的大小已知,相当于查询每个组的计数总和,该类查询记为 Q_2.由定理 1 可知,$\Delta Q_1=2\theta+1$.根据定义 3 可知行③④满足 ε_2-差分隐私.根

据引理 1 及 1.1 节中对图的点差分隐私描述可知，基于 SER 投影方法的(θ,Ω)-直方图发布机制满足$(\varepsilon_1+\varepsilon_2)$-差分隐私. 证毕.

同理可知基于 SER 投影方法的累积直方图也满足$(\varepsilon_1+\varepsilon_2)$-点差分隐私，其隐私安全性证明与定理 3 类似，不再赘述.

5 实验与结果

为评估本文所提 SER 投影方法以及基于该算法的直方图度发布机制的性能，本节将 SER 投影方法和已有的 3 种图投影方法 Truncation, ER, π_θ[4-5,7]在不同数据集上的运行效果做一对比. 仿真实验所用的数据集包括社交网络(Facebook, Twitter)、选举投票(Wiki-Vote)、电子邮件(Email-Enron)、协作网络(Ca-HepPh, DBLP)6 个现实世界中的真实数据集，均来自 Stanford Large Network Dataset Collection 网站[30]. 表 1 给出了这 6 个数据集的部分特征，其中 $deg_{\max}$表示图中节点的最大度，deg_{avg}表示图中节点的平均度. 实验平台采用 Intel® Core™ i5-7400 CPU、8GB 内存主机.

Table 1 Information About Dataset

表 1 数据集信息

Dataset	$\|V\|$	$\|E\|$	$deg_{\max}$	deg_{avg}
Facebook	4 039	88 234	1 045	43.69
Wiki-Vote	7 115	100 736	1 065	28.32
Ca-HepPh	12 008	118 521	491	21.00
Email-Enron	36 692	183831	1 383	10.73
Twitter	81 306	1 342 310	3 383	33.02
DBLP	317 080	1 049 866	343	6.62

5.1 实验设置

由于把节点度限制到门限值 θ 时，大量度大于 θ 的节点投影成了度小于等于 θ 的节点，进而导致了度等于 θ 的节点的计数增加，从而影响 L_1 误差的计算结果，掩盖了投影方法的其他特性. 因此，在表 2 的对比结果中，为了更好地反映出投影方法的优劣，我们在计算 L_1 误差时删除了度等于 θ 的节点计数.

在点差分隐私约束下直方图度发布算法的对比实验中，由于存在拉普拉斯噪声，为了更好地反映算法的优势，我们对每个 ε 取值计算 30 次，最后取平均值作为输出. 同时，取候选集的大小为 100，即 $\Theta\in[1,100]$.

Table 2 The Results on Different Datasets

表 2 Truncation, ER, π_θ, SER 方法在 5 个数据集上的实验结果

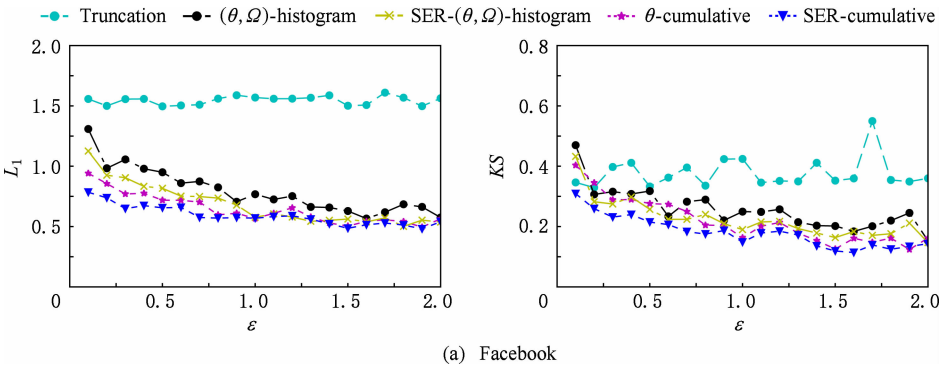
Dataset	Metrics	$\theta=16$				$\theta=64$				$\theta=128$			
		T	ER	π_θ	SER	T	ER	π_θ	SER	T	ER	π_θ	SER
Facebook	E'	2 653	19 083	23 440	25 478	24 695	55 473	58 585	61 653	50 320	76 193	77 256	79 021
	L_1	3 920	4 804	3 236	3 283	1 897	1 905	1 419	1 378	954	1 069	913	923
Wiki-Vote	E'	1 102	12 906	19 422	24 858	11 136	45 591	51 733	57 343	35 274	71 092	73 499	76 515
	L_1	5 875	1 286	3 322	3 269	3 607	2 130	1 793	1 675	3 207	1 300	1 150	929
Ca-HepPh	E'	11 874	35 091	39 725	43 037	26 107	71 455	75 412	79 611	65 231	90 468	92 566	96 272
	L_1	4 923	5 287	3 708	5 057	4 069	1 990	1 651	1 422	3 897	1 443	1 280	724
Email-Enron	E'	23 932	54 338	63 788	70 271	62 537	104 835	110 816	115 617	90 126	134 332	136 840	139 333
	L_1	14 750	11 814	12 478	8 053	7 297	7 905	7 583	4 008	6 978	5 751	5 650	3 050
Twitter	E'	40 325	244 421	324 639	385 269	296 378	703 189	761 571	816 162	631 038	971 842	999 936	1029546
	L_1	69 993	65 122	47 159	52 915	38 978	19 922	17 419	16 007	19 453	11 399	9 860	7 684
DBLP	E'	451 642	729 033	756 476	786 445	703 510	1 008 403	1 010 410	1 014 090	1 018 473	1 044 235	1 044 375	1 044 975
	L_1	83 630	59 289	50 646	35 351	36 530	5 673	4 756	2 923	3 065	1 152	1 184	1 152

5.2 结果分析

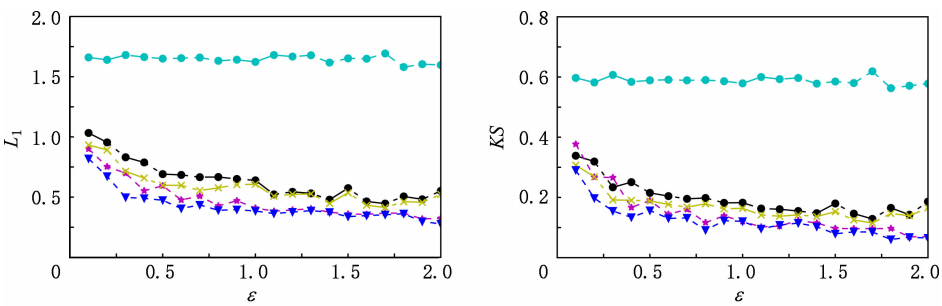
表 2 给出了 Truncation[4], ER[5], π_θ[7] 3 种图投影算法和本文所提出的 SER 投影方法在 6 个不同数据集上，θ 取 16, 64, 128 时的实验结果，其中 E' 为经过投影后保留的原始图的边个数，L_1 为定义的 L_1 误差，越小表示数据可用性越好.

从表 2 中可以看出，随着度门限值 θ 的增加，这 4 种算法中保留边的数目都在不断增加，同时 L_1 不断减小. 但是，与其他 3 种已有算法相比，本文所提 SER 投影方法能在保留边最多的情况下，同时也能保持较好的 L_1 误差. 这说明 SER 方法在投影后能更好的保证原始图的度分布的形状，使其更加接近

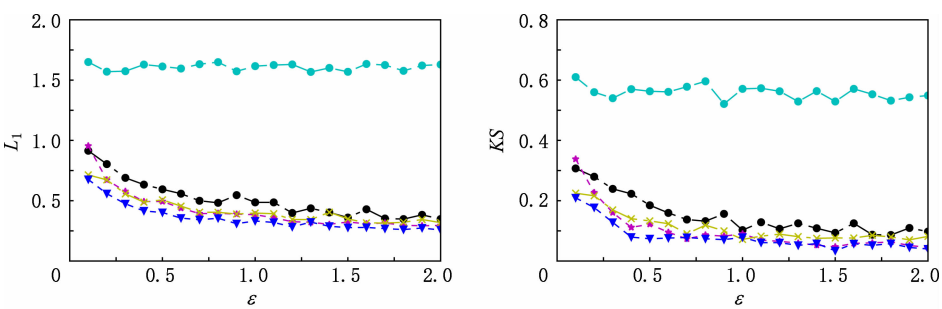
真实分布,为后续的数据分析和处理奠定基础.图 3 比较了在 L_1,KS 这 2 种不同的度量指标下,本文提出的方法与文献[4]、文献[7]提出的方法在刻画节点度分布时的差异.其中图 3(a)~(f)左半部分是用 L_1 度量的结果,图 3(a)~(f)右半部分为用 KS 度量的结果.从图 3 可以看出,随着数据集的变化,截断算法下的 L_1,KS 一直是所有方法里面最大的,说明截断算法的效果最差.这是由于算法本身删除了许多本没必要删除的边,损失了原始图中的大量有效信息,造成了度分布的误差过大.(θ,Ω)-直方图发布方法和累积度直方图发布方法的结果表明:对不同数据集,随着规模的增大,2 种方法的误差呈现减小趋势;对于相同的数据集,随着隐私预算的增大,2 种方法的误差呈现减小趋势,这符合我们一般的规律.



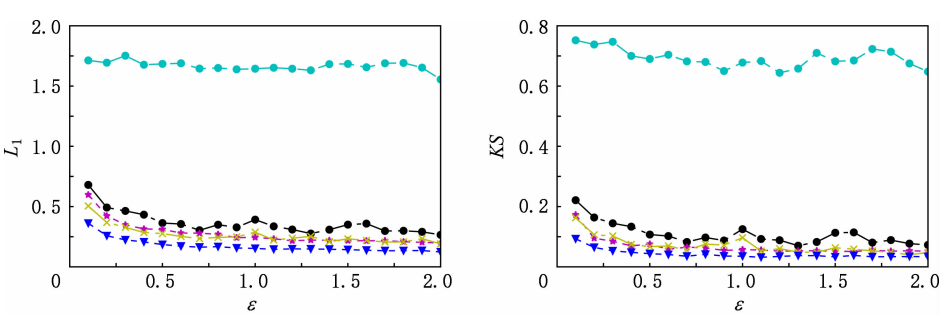
(a) Facebook



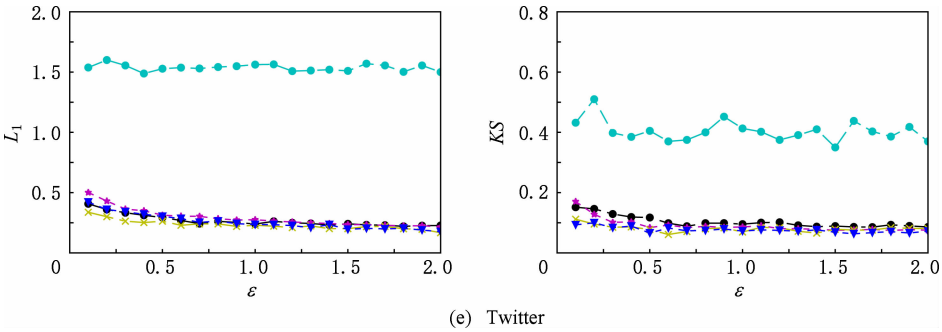
(b) Wiki-Vote



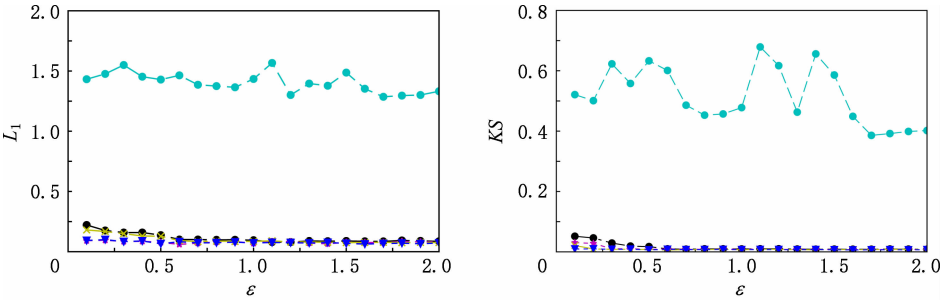
(c) Ca-HepPh



(d) Email-Enron



(e) Twitter



(f) DBLP

Fig. 3 Comparison of two histogram publishing methods

图 3 2 种直方图发布方法实验

从表 1 可以看出,Twitter 和 DBLP 这 2 个数据集的边个数差距不大,但是 DBLP 有更多的节点,Twitter 有更大的最大度和平均度,这说明 DBLP 所构成的图相较 Twitter 来说更加分散. 反映到图 3 中就是在 2 种直方图发布方法下,DBLP 相较 Twitter 的误差更小,这说明了分散图的度分布更好去刻画. 这是由于分散图存在大量的节点 p,其中 $deg(p)\leqslant\theta$. 而我们提出的方法能更好地保留度小于 θ 的节点所连接的边,所以出现这种情况. 同时为了方便对比,我们把同一类型的纵坐标取到相同的范围,这使得最后 2 个数据集的效果不能很好地区分,不过从原始数据来看,我们提出的 2 种基于 SER 投影方法的直方图发布方法效果要好于文献[7]中的方法,这也符合前 4 种数据集的实验结果.

总的来说,累积直方图发布方法要优于(θ,Ω)-直方图发布方法,基于 SER 投影方法的 2 种直方图发布方法在不同数据集上的效果优于基于 π_θ 的直方图发布算法和截断算法. 特别地,当隐私预算 $\varepsilon\leqslant 1$ 时,这种优势更为明显. 这说明本文所提点差分隐私约束下度分布发布算法适用于对隐私预算控制很严格的情况,更符合隐私保护的相关要求.

6 总 结

随着社交网络、推荐系统等新型信息系统的广泛使用,图数据在大数据处理领域变得越发常见. 大数据分析技术在图数据上的广泛应用可以带来巨大的经济价值和社会效益,这又促使各种组织机构开始搜集和发布大规模的图数据集. 另一方面,所搜集的图数据中又包含用户的敏感信息,如何在保护用户隐私的前提下发布这些数据变得尤为迫切.

本文重点关注点差分隐私约束下的图的度分布发布机制的设计,首先提出了一种基于度排序的边移除(SER)算法,通过将原始图投影到一个压缩图来降低发布机制中的全局敏感度. 进一步地,基于所提 SER 投影方法给出了 2 种满足点差分隐私的度直方图发布机制,并证明了其隐私安全性. 仿真实验表明:相比已有方法,在相同的约束条件下,本文所提 SER 投影方法能最大程度地保留原始图中的边信息,为后续的数据处理奠定了良好的基础. 与已有度分布发布方法相比,基于 SER 投影方法的度直方图发布机制在 L_1 误差和 KS-距离这 2 个评估指标上均具有优势,使得发布后的度分布更接近原始图的度分布,可用性也越高.

参 考 文 献

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

[2] Backstrom L, Dwork C, Kleinberg J. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography [C] //Proc of the 16th Int Conf on World Wide Web. New York: ACM, 2007: 181-190

[3] Hay M, Li Chao, Miklau G, et al. Accurate estimation of the degree distribution of private networks [C] //Proc of the 11th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2009: 169-178

[4] Kasiviswanathan S P, Nissim K, Raskhodnikova S, et al. Analyzing graphs with node differential privacy [C] //Proc of the 10th Conf on Theory of Cryptography. Berlin: Springer, 2013: 457-476

[5] Blocki J, Blum A, Datta A, et al. Differentially private data analysis of social networks via restricted sensitivity [C] // Proc of the 4th Conf on Innovations in Theoretical Computer Science. New York: ACM, 2013: 87-96

[6] Raskhodnikova S, Smith A. Lipschitz extensions for node-private graph statistics and the generalized exponential mechanism [C] //Proc of the 57th IEEE Annual Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2016: 495-504

[7] Day W-Y, Li Ninghui, Lyu M. Publishing graph degree distribution with node differential privacy [C] //Proc of the 16th Int Conf on Management of Data. New York: ACM, 2016: 123-138

[8] Samarati P. Protecting respondents identities in microdata release [J]. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(6): 1010-1027

[9] Machanavajjhala A, Gehrke J, Kifer D, et al. *l*-diversity: Privacy beyond *k*-anonymity [J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 1-52

[10] Dwork C, Naor M, Pitassi T, et al. Differential privacy under continual observation [C] //Proc of the 42nd ACM Symp on Theory of Computing. New York: ACM, 2010: 715-724

[11] Sarathy R, Muralidhar K. Some additional insights on applying differential privacy for numeric data [G] //LNCS 6344: Proc of the 7th Int Conf on Privacy in Statistical Databases. Berlin: Springer, 2010: 210-219

[12] Dwork C, Naor M, Pitassi T, et al. Pan-private streaming algorithms [C] //Proc of the 1st Symp on Innovations in Computer Science. Beijing: Tsinghua University Press, 2010: 66-80

[13] Cormode G, Procopiuc M, Srivastava D, et al. Differentially private summaries for sparse data [C] //Proc of the 15th Int Conf on Database Theory. New York: ACM, 2011: 299-311

[14] Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis [C] //Proc of the 39th ACM Symp on Theory of Computing. New York: ACM, 2007: 75-84

[15] Rastogi V, Hay M, Miklau G, et al. Relationship privacy: Output perturbation for queries with joins [C] //Proc of the 28th ACM SIGMOD-SIGACT-SIGART Symp on Principles of Database Systems. New York: ACM, 2009: 107-116

[16] Karwa V, Raskhodnikova S, Smith A, et al. Private analysis of graph structure [J]. Proceedings of the VLDB Endowment, 2011, 4(11): 1146-1157

[17] Zhang Jun, Cormode G, Procopiuc C M, et al. Private release of graph statistics using ladder functions [C] //Proc of the 36th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2015: 731-745

[18] Chen Shixi, Zhou Shuigeng. Recursive mechanism: Towards node differential privacy and unrestricted joins [C] //Proc of the 34th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2013: 653-664

[19] Xiong Ping, Zhu Tianqing, Wang Xiaofeng. A survey on differential privacy and applications [J]. Chinese Journal of Computers, 2014, 37(1): 101-122 (in Chinese)
(熊平,朱天清,王晓峰. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(1): 101-122)

[20] Xu Jia, Zhang Zhenjie, Xiao Xiaokui, et al. Differentially private histogram publication [J]. The VLDB Journal, 2013, 22(6): 797-822

[21] Qardaji W, Yang Weining, Li Ninghui. Understanding hierarchical methods for differentially private histograms [J]. Proceedings of the VLDB Endowment, 2013, 6(14): 1954-1965

[22] Hay M, Rastogi V, Miklau G, et al. Boosting the accuracy of differentially private histograms through consistency [C] // Proc of the 36th Int Conf on Very Large Data Bases. New York: ACM, 2010: 1021-1032

[23] Lin B-R, Kifer D. Information preservation in statistical privacy and bayesian estimation of unattributed histograms [C] //Proc of the 34th ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2013: 677-688

[24] Zhang Xiaojian, Shao Chao, Meng Xiaofeng. Accurate histogram release under differential privacy [J]. Journal of Computer Research and Development, 2016, 53(5): 1106-1117 (in Chinese)
(张啸剑,邵超,孟小峰. 差分隐私下一种精确直方图发布方法[J]. 计算机研究与发展, 2016, 53(5): 1106-1117)

[25] Lee J, Wang Yue, Kifer D. Maximum likelihood postprocessing for differential privacy under consistency constraints [C] //Proc of the 21st ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2015: 635-644

[26] Dwork C, Mcsherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis [C] //Proc of the 3rd Theory of Cryptography Conf. Berlin: Springer, 2006: 265-284

[27] Mcsherry F, Talwar K. Mechanism design via differential privacy [C] //Proc of the 48th Annual IEEE Symp on Foundations of Computer Science. Piscataway, NJ: IEEE, 2007: 94-103

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

[29] Acs G, Chen Rui. Differentially private histogram publishing through lossy compression [C] //Proc of the 11th IEEE Int Conf on Data Mining. Piscataway, NJ: IEEE, 2012: 84-95

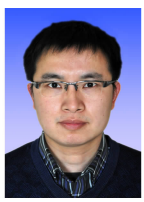
[30] Stanford University. Stanford Large Network Dataset Collection[EB/OL]. (2009-07)[2017-06-08]. http://snap.stanford.edu/data/

Zhang Yuxuan, born in 1994. Master. His main research interests include big data security and privacy protection

Wei Jianghong, born in 1987. PhD, lecturer. His main research interests include authentication protocol, big data security and privacy protection.

Li Ji, born in 1992. Master. His main research interests include big data security and privacy protection.

Liu Wenfen, born in 1965. PhD, professor and PhD supervisor. Her main research interests include cryptography and information security.



Hu Xuexian, born in 1982. PhD, lecturer. His main research interests include security protocol and big data security.

2019 年《计算机研究与发展》专题(正刊)征文通知

——密码学与智能安全研究

依托于云计算、物联网、大数据技术的发展,自动驾驶、人脸识别、智能家居等人工智能技术快速进入了人们的视野,目前已经成为先进科技社会化应用的代表和社会热点.但是,安全问题却为这些技术的广泛应用提出了严峻挑战,没有强大的自主可控的安全技术的支撑,人工智能带来的也许不仅仅是便利,更可能是灾难.安全问题可以说是人工智能走向大规模应用的瓶颈和一个关键问题.而作为解决安全问题的核心技术——密码学,如何适应人工智能安全的需要是另一个关键问题.

为推动我国学者在智能安全领域的研究,为人工智能的现实应用提供理论与技术支撑,及时报道我国学者在智能安全理论与技术方面的最新研究成果,《计算机研究与发展》将于 2019 年 10 月出版网络与信息安全专题——密码学与智能安全研究,主要聚焦人工智能系统及其基础性环境安全所涉及的密码学、系统安全理论与技术,欢迎相关领域的专家学者和科研人员踊跃投稿.

征文范围 本专辑包括(但不限于)下列主题:

- 用于人工智能运行环境的密码算法与密码协议;
- 群智安全服务与安全交易;
- 网络系统安全、物联网安全与外包数据安全;
- 人工智能安全与基于人工智能的密码分析;
- 移动智能计算安全;
- 大数据安全与隐私保护下的数据处理.

投稿要求

1) 论文应属于作者的科研成果,数据真实可靠,具有重要的学术价值或推广应用价值,未在国内外公开发行的刊物或会议上发表过,不存在一稿多投问题.作者在投稿时,需向编辑部提交版权转让协议.
2) 论文一律用 Word 格式排版,格式体例参考近期出版的《计算机研究与发展》文章的要求(http://crad.ict.ac.cn).
3) 论文请通过《计算机研究与发展》期刊网站(http://crad.ict.ac.cn)进行投稿,投稿时提供作者的联系方式,并在作者留言中注明"网络与信息安全 2019 专题"(否则按自由来搞处理).

重要日期

征文截止日期: 2019 年 6 月 11 日
录用通知日期: 2019 年 7 月 20 日
作者修改稿提交日期:2019 年 8 月 6 日
出版日期:2019 年 10 月

特邀编委

徐秋亮 教授 山东大学 xql@sdu.edu.cn
张玉清 教授 中国科学院大学、国家计算机网络入侵防范中心 zhangyq@ucas.ac.cn
董晓蕾 教授 华东师范大学 dongxiaolei@sei.ecnu.edu.cn

联系方式

联 系 人:徐秋亮 xql@sdu.edu.cn
编 辑 部:crad@ict.ac.cn,010-62620696,010-62600350
通信地址:北京 2704 信箱《计算机研究与发展》编辑部 邮政编码:100190