



文章编号:1001-9081(2019)05-1394-06

DOI:10.11772/j.issn.1001-9081.2018112556

基于 Skyline 计算的社交网络关系数据隐私保护

张书旋¹, 康海燕^{2*}, 闫 涵²

(1. 北京信息科技大学 计算机学院, 北京 100192; 2. 北京信息科技大学 信息管理学院, 北京 100192)

(*通信作者电子邮箱 kanghaiyan@126.com)

摘要:随着社交软件的流行,越来越多的人加入社交网络产生了大量有价值的信息,其中也包含了许多敏感隐私信息。不同的用户有不同的隐私需求,因此需要不同级别的隐私保护。社交网络中用户隐私泄露等级受社交网络图结构和用户自身威胁等级等诸多因素的影响。针对社交网络数据的个性化隐私保护问题及用户隐私泄露等级评价问题,提出基于 Skyline 计算的个性化差分隐私保护策略(PDPS)用以发布社交网络关系数据。首先构建用户的属性向量;接着采用基于 Skyline 计算的方法评定用户的隐私泄露等级,并根据该等级对用户数据集进行分割;然后应用采样机制来实现个性化差分隐私,并对整合后的数据添加噪声;最后对处理后数据进行安全性和实用性的分析并发布数据。在真实数据集上与传统的个性化差分隐私方法(PDP)对比,验证了 PDPS 算法的隐私保护质量和数据的可用性都优于 PDP 算法。

关键词:社交网络;隐私保护;Skyline 计算;个性化差分隐私;基于 Skyline 计算的个性化差分隐私保护算法

中图分类号:TP309 **文献标志码:**A

Privacy preserving for social network relational data based on Skyline computing

ZHANG Shuxuan¹, KANG Haiyan^{2*}, YAN Han²

(1. School of Computer Science, Beijing Information Science and Technology University, Beijing 100192, China;

2. School of Information Management, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract: With the popularity and development of social software, more and more people join the social network, which produces a lot of valuable information, including sensitive private information. Different users have different private requirements and therefore require different levels of privacy protection. The level of user privacy leak in social network is affected by many factors, such as the structure of social network graph and the threat level of the user himself. Aiming at the personalized differential privacy preserving problem and user privacy leak level problem, a Personalized Differential Privacy based on Skyline (PDPS) algorithm was proposed to publish social network relational data. Firstly, user's attribute vector was built. Secondly, the user privacy leak level was calculated by Skyline computation method and the user dataset was segmented according to this level. Thirdly, with the sampling mechanism, the users with different privacy requirements were protected at different levels to realize personalized differential privacy and noise was added to the integrated data. Finally, the processed data were analyzed for security and availability and published. The experimental results demonstrate that compared with the traditional Personalized Differential Privacy (PDP) method on the real data set, PDPS algorithm has better privacy protection quality and data availability.

Key words: social network; privacy preserving; Skyline query; personalized differential privacy; Personalized Differential Privacy based on Skyline (PDPS) algorithm

0 引言

社交网络隐私信息可以分为两种:一种隐私是用户敏感信息隐私,比如用户的手机号码、家庭住址、疾病、收入等;另一种隐私是社交网络关系隐私,即社交网络中人与人之间的连接关系信息,如亲属关系、同学关系。在社交网络中无论是哪种类型隐私信息的披露都可能会使个人的隐私受到威胁,因此,隐私信息的识别和分类是非常必要的,需结合具体的信息类别来采取相应有效的保护策略。本文主要研究社交网络关系型数据的隐私保护与发布。

差分隐私^[1]被公认是一个强大的隐私保护模型,能够为数据提供强大的隐私保证,但是该模型局限于为所有个人提供相同级别的隐私保护;然而并非所有用户都需要相同的隐私级别,为避免对那些不需要太高隐私级别的用户提供过多的隐私保护,需要实现个性化的隐私保护。本文采用采样方法实现个性化差分隐私保护,引入非均匀的不确定性。

采样机制以前为其他目的已经与差分隐私结合过。Li 等^[2]提出了一种满足差分隐私的扰动方法,利用采样的随机性降低隐私成本,证明了均匀随机采样提高了差分隐私保护效果。Kellaris 等^[3]提出了 CS 预处理方法(pre-processes by

收稿日期:2018-12-04;修回日期:2018-12-28;录用日期:2018-12-28。

基金项目:国家自然科学基金资助项目(61370139);北京市社会科学基金资助项目(15JGB099,15ZHA004)。

作者简介:张书旋(1993—),女,辽宁鞍山人,硕士研究生,主要研究方向:信息安全; 康海燕(1971—),男,河北石家庄人,教授,博士,CCF 会员,主要研究方向:网络安全、隐私保护; 闫涵(1994—),女,北京平谷人,硕士研究生,主要研究方向:信息安全。



Grouping and Smoothing, GS), 利用抽样机制对发布数据进行分组, 降低拉普拉斯噪声注入并实现差分隐私。Spiessl 等^[4]设计了一个根据灵敏度采样的采样器, 能自动实现 $(\varepsilon, \delta, \gamma)$ 随机差分隐私。

Skyline 计算的研究分两方面:一是对 Skyline 计算算法的优化,二是将 Skyline 计算算法应用于相关研究领域。目前, 数据的海量性和高维性以及数据环境的多样性和动态性都使 Skyline 计算面临着愈加严峻的挑战。各国学者针对这个问题进行了不少研究, 主要得出以下几种 Skyline 计算算法:块嵌套算法^[5]、最近邻算法^[6]、分支界限算法^[7]等。Skyline 算法在多标准决策系统、城市导航系统、数据库可视化、用户偏好查询等多个研究领域都有着广泛应用, 例如: 在传感器网络应用中, 信俊昌等^[8]提出了基于过滤的 Skyline 节点连续查询算法(Filter-based Skyline node moniToring algorithm, FIST), FIST 算法能有效减少 Skyline 节点连续查询过程中传感器节点的通信代价, 进而降低传感器网络的能量消耗; 多维向量查询方面, 雷婷等^[9]在云环境下提出一种基于超球面投影分区的 Skyline 算法, 通过将空间坐标投影到超球面上转化为超球面投影坐标, 然后使用超球面投影坐标进行分区, 有效提高分区内的数据点的平均减枝力度, 降低 Skyline 的计算代价; 在用户偏好查询方面, Zhang 等^[10]提出了一种使用 Skyline 查询的算法, 通过用户的搜索和查询, 以确定哪种云服务最能满足用户的需求。本文利用 Skyline 计算方法给用户评定隐私泄露等级, 然后根据用户隐私泄露等级进行采样处理, 最后添加噪声实现个性化的差分隐私保护。

本文的主要贡献包括:1) 利用采样方法实现个性化差分隐私, 引入非均匀的不确定性; 2) 构建了用户属性向量, 利用 Skyline 计算方法给用户评定隐私泄露等级; 3) 提出了基于基于 Skyline 计算的个性化差分隐私保护 (Personalized Differential Privacy based on Skyline, PDPS) 算法发布机制的整体流程, 对数据采集、数据处理、数据分析和数据发布各个流程进行了阐述说明; 4) 在真实数据集上与传统个性化差分隐私 (Personalized Differential Privacy, PDP) 算法进行对比, 验证了发布数据安全性和可用性的提升。

1 相关工作

1.1 Skyline 计算

定义 1 Skyline。一个多维数据集的 Skyline, 是指该数据集上不被其他任何数据点支配的点所组成的集合。

已知数据集上两个点 p 和点 q , 如果当且仅当 p 在任一维上的取值都不比 q 差, 且至少在一个维度上比 q 更好, 则称数据点 p 支配点 q 。

定义 2 Skyline 计算^[11], 就是从数据集中快速、准确地找到所有的 Skyline 数据点。

Skyline 计算是一个典型的多目标优化的问题。Skyline 一个经典的例子: 假设去海滩旅游, 想找一个既便宜距离又近的旅馆, 一般情况下越靠近海滩的旅馆价格越高, 所以不能返回一个最好的结果, 只能返回一些用户可能感兴趣的旅馆。列出这些旅馆的评价属性表如表 1, 根据价格和距离两个属性画出散点图如图 1, 在图 1 中找出在价格和距离两个方面都不比其他旅馆差的旅馆, 这些不被支配的旅馆就是 Skyline。

表 1 旅馆的评价属性表

Tab. 1 Evaluation attribute table of hotels

旅馆	距离/km	价格/元	旅馆	距离/km	价格/元
p_1	3	350	p_6	20	160
p_2	6	250	p_7	23	25
p_3	10	400	p_8	26	100
p_4	13	50	p_9	30	300
p_5	16	75	p_{10}	40	130

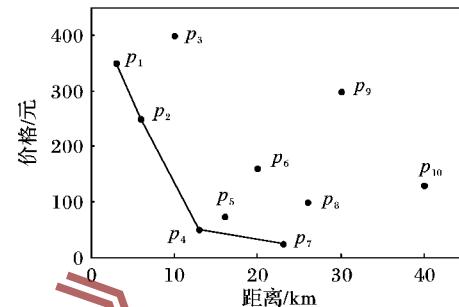


Fig. 1 Scatter diagram of hotel selection using Skyline computation

社交网络图中用户的隐私泄漏等级受多个属性影响。每一个属性的大小都有可能引起隐私的泄露, 所以不能用具体的权重值衡量各属性对用户隐私泄漏等级的影响大小。本文使用 Skyline 计算的方法评定用户的隐私泄漏等级。

1.2 个性化差分隐私

社交网络被定义成一个无向无加权的图 $G = (V, E)$, V 表示社交网络中用户实体的集合, E 表示边。用户的个数记为 N , 边表示用户之间的关系(例如友谊、合作和联系)。 $e(v_i, v_j) \in E$ 表示定点 v_i 和 v_j 的边。令 $|V| = n_0$, $|E|$ 表示边的条数。若将 G 删除或添加一个边得到 G' , 最小改变量 $d(G, G') \leq 1$, 则称两个图 G, G' 相邻, 互为相邻图。 $G \xrightarrow{e} G'$ 表示 G, G' 相邻。

定义 3 ε -差分隐私。机制 $M: G \rightarrow G'$ 满足差分隐私^[12]的条件是对任意 G, G' , 任意可能输出的 $O \in Range(M)$, 都有:

$$\Pr[M(G) \in O] \leq e^\varepsilon \cdot \Pr[M(G') \in O]$$

在这个定义中, $\varepsilon > 0$ 是公开已知的隐私参数, 它控制差分隐私保证的强度: ε 越大隐私强度越弱, ε 越小隐私强度越强。当 ε 足够小时, $e^\varepsilon \approx 1 + \varepsilon$ 。根据差分隐私方法, $f(G)$ 在发布之前应该通过隐私算法的扰动, 使得输出中隐藏关于 G 中边的信息。

差分隐私具有序列组成性。也就是说, 如果有 k 个机制 M_1, M_2, \dots, M_k , 每个机制独立地满足差分隐私, 在输入 G 上依次运行这些机制, 则该序列是差分隐私的, 其中 $\varepsilon' = \sum_{i=1}^k \varepsilon_i$ 。

对于数值型数据来说, 实现差分隐私的最常见的方式是选择适合的随机噪声注入到输出中。本课题研究的是社交网络图, 故采用更适用于非数值型数据的指数机制来实现差分隐私。

定义 4 个性化差分隐私(PDP)^[13]。在隐私要求 P 下, 一个随机机制 M 满足 P -个性化差分隐私(P -PDP), 如果每一



对相邻图 G, G' , 有 $G \xrightarrow{e_{ij}} G'$, $e_{ij} = e(v_i v_j)$, 并且对任意 $O \in Range(M)$ 都有:

$$\Pr[M(G) \in O] \leq e^{\min\{p^{v_i}, p^{v_j}\}} \cdot \Pr[M(G') \in O]$$

PDP 提供了与传统差分隐私提供的类似保护强度的隐私保护,但是 PDP 的隐私保证被个性化地满足每个用户的需求。

定义 5 序列组成性。 Jorgensen 等^[13]提出由传统的差分隐私的组成性自然延伸到 PDP,令 M_1 和 M_2 表示两种机制分别满足 P_1 -PDP 和 P_2 -PDP,则 M_1 和 M_2 组成的序列 $M_3 = (M_1(G), M_2(G))$ 也满足 P -PDP。

定义 6 采样机制。 实现 PDP 的智能通用机制,称为采样机制^[14]。采样机制通过引入两种独立类型的随机性来工作:一种是对边的非均匀随机抽样,另一种是通过使用传统的差分隐私机制对输入采样来添加均匀的随机性。

对于函数 $f(G)$,社交网络 G ,指定阈值 t ,隐私要求 P 。将 $RS(G, P, t)$ 定义为一个算法,该算法独立抽取每个边 $e_{ij} = e(v_i, v_j) \in G$ 的样本,可能性为:

$$\pi(e_{ij}, t) = \begin{cases} \frac{e^{\min\{p^{v_i}, p^{v_j}\}} - 1}{e^t - 1}, & \min\{p^{v_i}, p^{v_j}\} < t \\ 1, & \text{其他} \end{cases}$$

其中 $\min p^v \leq t \leq \max p^v$ 。

采样机制表示为 $S_f(G, P, t) = DP_t(RS(G, P, t))$, 其中 DP_t 是对于函数 f 的任意 t -差分隐私机制。 DP_t 机制可以是拉普拉斯机制,也可以是指数机制,或者是由几种差分隐私算法组合的隐私机制。

2 基于 PDPS 的社交网络数据发布机制

2.1 相关定义

定义 7 CFP(Connection FingerPrint)值。 $CFPi$ 表示用户在 i 跳内连接的用户数量。由社交网络图可以得到用户在各跳内的连接信息,即在各跳内连接到的用户的数量。 $CFP1$ 为第一跳 CFP 值, $CFP2$ 为 2 跳内的连接点数。以图 2 中的社交网络图为例,得到用户的 CFP 连接信息如表 2。表 2 中, v_1 的 $CFP1 = 3$, $CFP2 = 3$ 。 CFP 的数量是社交网络用户属性中最重要的属性之一。通过这些统计数据可以了解用户的社会影响,研究通过媒体传播的方式,制定合理的用户推广广告等。

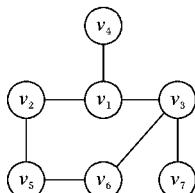


图 2 社交网络示例

Fig. 2 Example of social network

定义 8 隐私要求 P (Privacy Preferences)。 隐私要求是一个用户的个性化的隐私偏好, P 值越小表示隐私要求越高,要求的隐私保护级别越高。本文设定每个用户可以设置自己的隐私要求,以确保每个用户都能得到精确的隐私保护。

定义 9 邻接点威胁等级 T (Threat level of connection points)。 邻接点威胁等级是指一个用户通过邻接点泄漏隐私信息的可能程度。一个用户具有越多的连接点,隐私要求越

低,这个用户就越容易泄露相邻点的隐私信息。

定义 10 用户属性向量 (User properties vectors)。 通过对输入的原始社交网络图的分析,可以提取出用户的 $CFP1$ 值、 $CFP2$ 值、隐私要求 P 、邻接点威胁等级 T 。将这几个属性记录为向量,第 i 个用户的属性向量为 $\{CFP1[i], CFP2[i], P[i], T[i]\}$ 。

定义 11 隐私泄露等级 L (Privacy leak level)。 第一条 Skyline 上的用户隐私泄漏的可能性最小,这些用户的隐私泄露可能性记为 1。删除第一条 Skyline 上的点,从剩下的点中计算出第二条 Skyline,第二条 Skyline 上用户的隐私泄露可能性记为 2,依此类推。

定义 12 采样阈值 S (Sampling threshold)。 设定一个采样阈值 S ,将隐私泄漏等级与采样阈值比较:如果 $L \leq S$,则该用户可以被采样;如果 $L > S$,则不能输出该用户。

表 2 用户的 CFP 连接信息

Tab. 2 CFP connection information of users

用户	CFP 连接点
v_1	$\{v_2, v_3, v_4\}_1, \{v_5, v_6, v_7\}_2$
v_2	$\{v_1, v_5\}_1, \{v_3, v_4, v_6\}_2, \{v_7\}_3$
v_3	$\{v_1, v_6, v_7\}_1, \{v_2, v_4, v_5\}_2$
v_4	$\{v_1\}_1, \{v_2, v_3\}_2, \{v_5, v_6, v_7\}_3$
v_5	$\{v_2, v_6\}_1, \{v_1, v_3\}_2, \{v_4, v_7\}_3$
v_6	$\{v_3, v_5\}_1, \{v_1, v_2, v_7\}_2, \{v_4\}_3$
v_7	$\{v_3\}_1, \{v_1, v_6\}_2, \{v_2, v_4, v_5\}_3$

2.2 PDPS 发布机制

本文提出基于 Skyline 计算的个性化差分隐私保护策略 (Personalized Differential Privacy based on Skyline, PDPS) 的社交网络数据发布机制,发布流程如图 3。

该机制分为以下三个模块。

第一模块 数据采集层。

本文从斯坦福大学大规模数据平台获取社交网络数据。其中包括微信、微博、Facebook 等社交网络平台的数据集。数据集中包括用户的连接关系及相关属性。

第二模块 方法层。

方法层中使用 PDPS 策略进行处理。PDPS 策略具体分为以下几个步骤:

第一步 输入原始社交网络图数据集。

网络图数据集包括节点集和边集。节点集中的每一个节点代表每一个用户,每个用户都有其属性值,初始输入应存有用户的隐私要求 P 值。边集中的用 0 和 1 分别表示两节点无连接和有连接。

第二步 构建用户属性向量集。

首先计算每个用户的第一跳连接点数量 $CFP1$ 值,第二跳连接点数量 $CFP2$ 值和邻接点威胁等级。然后将这三个属性值和用户的隐私要求 P 值记录为该用户的属性向量,第 i 个用户的属性向量为 $\{CFP1[i], CFP2[i], P[i], T[i]\}$ 。最后将所有用户的属性向量构建为用户属性向量集,表示为 $N * 4$ 的矩阵。以 U_1 到 U_{10} 这 10 个用户举例,可列出用户属性表如表 3 所示。

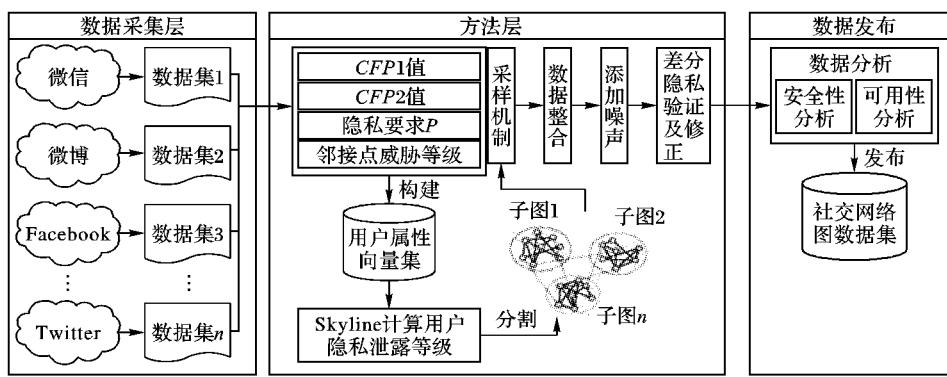


图3 基于PDPS的社交网络数据发布流程

Fig. 3 Social network data publishing process based on PDPS

表3 用户属性表

Tab. 3 User properties table

用户 U	第1跳连接点数量 $CFP1$	第2跳连接点数量 $CFP2$	隐私要求等级 P	邻接点威胁等级 T
U_1	2	3	0.9	0.1
U_2	4	5	0.7	0.4
U_3	5	6	1.0	0.2
U_4	6	10	0.2	0.6
U_5	8	14	0.3	0.8
U_6	11	16	0.6	0.5
U_7	13	19	0.1	0.8
U_8	14	20	0.4	0.7
U_9	15	21	0.8	0.6
U_{10}	20	24	0.5	0.4

第三步 Skyline 计算用户隐私泄露等级。

根据所有用户的属性向量计算第1条 Skyline, 该 Skyline 上的点隐私泄露等级定义为 $L = 1$, 若以 $CFP1$ 值和隐私要求 P 为 Skyline 的决策标准可得到第一条 Skyline 如图4 所示, 然后去掉这些节点, 计算第2条 Skyline, 该 Skyline 上的点隐私泄露等级定义为 $L = 2$; 以此类推。与选择旅店的例子类似, $CFP1$ 值越小即第一跳连接用户越少, P 值越小即隐私要求越高的用户, 为隐私越不容易泄漏的用户。则 L 值越小, 隐私越不容易被泄漏。设共分了 m 个等级, 根据数据集规模设定分割系数 k , 将数据集分割, 每 m/k 个等级为一个子数据集, L_1 至 $L(m/k)$ 的用户存入子数据集 1。

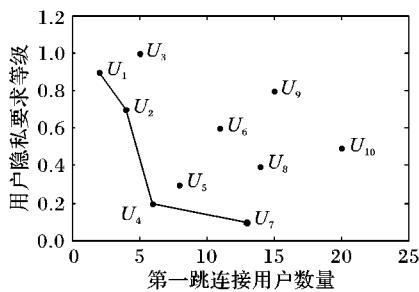


图4 Skyline 计算用户威胁等级散点图

Fig. 4 Scatter diagram of Skyline calculating user threat level

第四步 采样。

采样机制是实现差分隐私个性化的常用有效机制。首先根据用户的隐私要求 P , 计算每一条边的可能性 e_{ij} , 然后为每一子数据集设定采样阈值 $S[i]$ ($i = 1, 2, \dots, k$), 子数据集 1 中的用户隐私泄露等级较低, 则设定较大的采样阈值。根据

采样机制去掉隐私泄露等级较高的节点。最后将各子数据集整合, 子数据集间原有的连接关系基本不变, 输出 PDPS 处理后的社交网络数据集。

第五步 添加噪声。

为满足差分隐私, 需要在发布前对发布图添加噪声。针对社交网络图数据, 本文采用的加噪方法为添加虚拟边和添加虚拟节点, 然后验证是否满足差分隐私, 如果不满足则修改噪声参数。

第三模块 数据发布层。

在发布前应验证发布图的安全性和可用性。抵抗隐私攻击的能力能够反映数据的安全性, 本文使用隐私攻击的方法来验证数据的安全性。设定攻击者具有一定的背景知识, 结合发布的社交网络图进行链接攻击, 得出攻击结果的匹配度。将图数据结构特征参数中的平均最短路径及平均聚类系数与隐私保护之前原始数据集进行比较, 验证社交网络图的可用性。

2.3 PDPS 发布算法

PDPS 发布算法的伪代码如下:

输入 原始社交网络图 G , 隐私要求 P , 采样参数 $S[i]$, 分割参数 k , 总用户数 n ;

输出 保护后的社交网络图 G' 。

```

1) count  $N1[i], N2[i], T[i]$ ;
2)  $V[i] = \{N1[i], N2[i], P[i], T[i]\}$ ; // 构建用户属性向量
3)  $t = n/k$ ;
4) for  $i = 1$  to  $n$ ,  $j = 1$  to  $n$ ,  $m = 1$ ; // 计算用户隐私泄露等级
5) if  $V[i]$  root  $V[j]$ 
6) insert  $V[i]$  to  $\text{Skyline}[m]$ ;
7) remove  $V[i]$  from  $G$ ;
8) end for
9) insert  $\text{Skyline}[i] - \text{Skyline}[i + m/k]$  to  $G[i]$ ; // 根据 k 值分组
10) for each  $G[j]$ 
11) if  $P_V[i] \leq S[j]$ 
12) sample  $Node_V[i]$ ;
13)  $Node_V[i]$  to  $G[j]'$ ;
14) end for
15) put each  $G_i'$  to  $G'$ ;
16) add noise;
17) if  $G_{output}$  dissatisfies the unit-differential privacy
18) correct it;
19) publish the  $G_{output}$ ;
其中: 1) ~ 9) 表示计算用户隐私泄露等级, 根据泄露等

```



级分割;10)~15)表示采样机制及重组;16)~19)表示添加噪声、验证及发布。

3 实验评估

本文的实验环境为 macOS 10.13.1 操作系统,开发环境为 IntelliJ IDEA,编程语言为 Java,实验数据分析采用 Matlab。

实验数据为 SNAP 数据平台中 ego-Facebook 数据集(4039个用户,88234条边,平均聚类系数为0.6055,平均最短路径为1.4734)。采用随机数生成器生成区间在[0,1]的随机小数(步长为0.1)作为用户的隐私要求。比较在采样参数 $S[1]=0.9$ 和 $S[1]=0.7$ 时,各实验结果的变化情况。设定 $S[i]=S[1]-0.05*(i-1)$ 。对于不同规模的数据集,分割系数 k 的选取也不同。经过多次实验得出,当 $10 < k < 100$ 时,PDPS 算法处理后的数据可用性较高。

实验内容为验证 PDPS 算法的隐私保护质量和数据可用性两部分。

3.1 隐私保护质量

本文在 Facebook 数据集中验证本文提出的 PDPS 算法的隐私保护质量。验证方法为将隐私保护后的发布图作为子图,查看子图攻击下得到的发布结果与真实数据集的匹配度,匹配度越低,说明隐私保护效果越好。首先,假定攻击者拥有攻击对象的背景知识和不完整的子图信息;然后,进行节点识别攻击;最后,分别将一般的个性化差分隐私保护 PDP 算法和本文提出的 PDPS 算法下发布的社交网络图作为子图,比较两种背景知识下的攻击结果。实验结果为随着分割参数 k 的增加,即子图中节点数减小,攻击结果的匹配度,如图 5 所示。

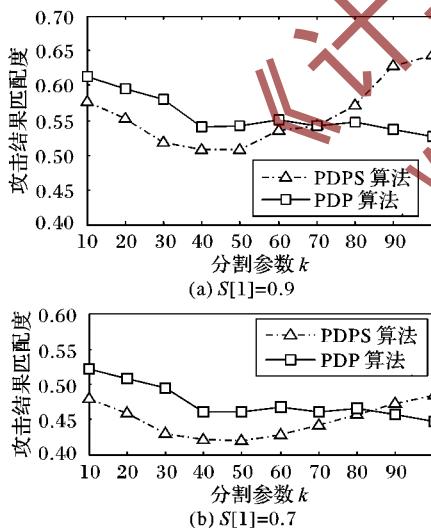


图 5 2 种算法的攻击结果匹配度对比

Fig. 5 Comparison of attack matching ratio between two algorithms

由图 5 可以看出:1)PDPS 算法攻击下数据的匹配度明显比 PDP 算法要小,即隐私保护质量高。2)随着分割系数的增加,匹配度逐渐降低,当 $k=50$ 时,达到最小值,说明 $k=50$ 时隐私保护效果最好。3)随着分割系数的进一步增大,攻击结果匹配度升高,且当分割系数过高时,由于 PDPS 算法发布的子图所包含的节点过少,比 PDP 算法下攻击得到的数据匹配度低,隐私保护质量变差。4)采样参数减小,数据量降低,算法执行时间降低。

3.2 数据可用性

实验选取社交网络中重要的两个结构特征:平均聚类系数、平均最短路径;并计算数据损失率,根据这三个指标对数据的可用性进行评估。同时观察分割系数对数据可用性的影响。随着分割系数的增加,PDPS 算法和 PDP 算法所发布的社交网络图的平均最短路径如图 6,平均聚类系数如图 7,数据的缺失率如图 8。

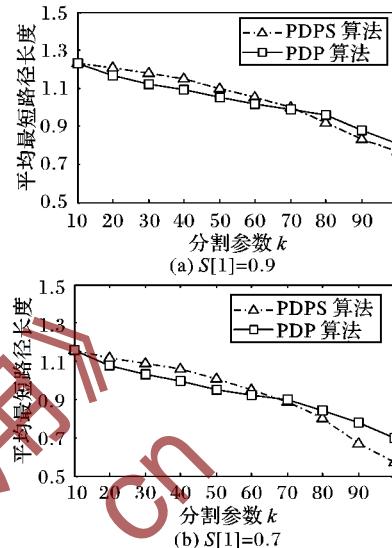


图 6 2 种算法的平均最短路径长度对比

Fig. 6 Comparison of average shortest path between two algorithms

从图 6 可以看出:1)PDPS 算法下发布数据集的平均最短路径比 PDP 算法要大。2)随着 k 值的增加,发布图的平均最短路径逐步减小。当 k 大于 80 时,图结构破坏严重,PDPS 算法下发布数据集的平均最短路径比 PDP 算法小。3)采样参数减小,发布的数据量少,图结构被破坏,所以平均最短路径减小。

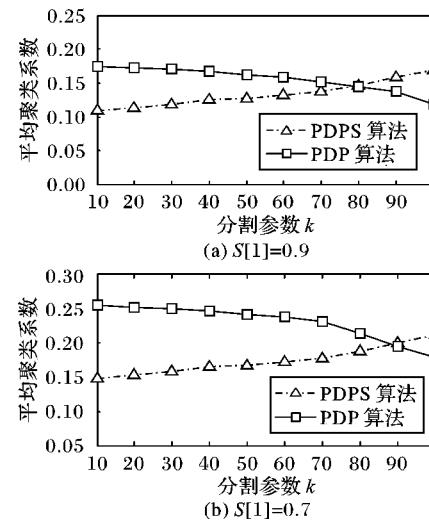


图 7 2 种算法的平均聚类系数对比

Fig. 7 Comparison of average clustering coefficient between two algorithms

从图 7 可以看出:1)随着 k 取值的增加,应用 PDP 算法发布的社交网络图的平均聚类系数逐渐下降,而 PDPS 算法的平均聚类系数逐渐增加。整体来看两种算法的性能比较接近, k 值的变化对平均聚类系数的影响并不显著,但也有一定影响。2)当 k 取值较小时,PDPS 算法的平均聚类系数较小,



即性能优于 PDP 算法;当 k 取值较大时,PDPS 算法发布的社交网络图与原始网络图偏离较大,此时发布图的平均聚类系数增大,性能相对 PDP 算法较差。3)采样参数减小,PDPS 算法在分割后的子图中进行采样,且按聚类系数进行采样,所以发布图的结构变化没有 PDP 算法大。因此,采样参数的变化对 PDPS 算法影响比 PDP 算法的影响较小。

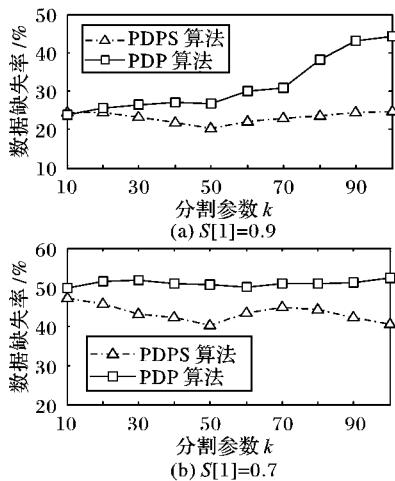


图 8 2 种算法的数据缺失率对比

Fig. 8 Comparison of data loss ratio between two algorithms

由图 8 可以看出:1) $S[1] = 0.9$ 时, PDPS 算法发布数据的损失率不超过 25%,且随着分割系数的增加,损失率有明显降低,当 $k = 50$ 时,数据损失率低至 20%,所以要根据数据集的大小及需要的隐私保护程度来合理选取分割系数的值。2) PDPS 算法的数据损失率均低于 PDP 发布算法,表明 PDPS 算法的数据可用性相对较好。3) 采样参数减小,发布的数据量减少,两算法的数据缺失率升高了,数据可用性降低。

4 结语

本文针对社交网络用户隐私泄露等级评定和差分隐私保护的个性化这两个问题,提出 PDPS 算法用来发布社交网络关系数据。将 Skyline 计算用于用户隐私泄露等级的评定,并用采样机制来实现了个性化的差分隐私保护。在真实数据集上对该算法的隐私保护效果和数据的效用进行了验证,证明了 PDPS 算法能够提升社交网络数据发布的安全性和可用性。今后可对分割和采样系数的选取规则、带权重的社交网络图的发布算法等方向进行下一步研究。

参考文献 (References)

- [1] 王豪,徐正全.面向轨迹聚类的差分隐私保护方法[J].华中科技大学学报(自然科学版),2018,46(1):32–36.(WANG H, XU Z Q. Differential privacy preserving method for trajectory clustering[J]. Journal of Huazhong University of Science & Technology (Natural Science Edition), 2018, 46(1):32–36.)
- [2] LI N, QARDAJI W, DONG S. On sampling, anonymization, and differential privacy or, k -anonymization meets differential privacy [C]// Proceedings of the 2012 ACM Symposium on Information, Computer and Communications Security. New York: ACM, 2012: 32–33.
- [3] KELLARIS G, PAPADOPOULOS S. Practical differential privacy via grouping and smoothing[J]. Proceedings of the VLDB Endowment, 2013, 6(5):301–312.
- [4] SPIESSL S M, BECKER D A. Sensitivity analysis of a final repository model with quasi-discrete behaviour using quasi-random sampling and a metamodel approach in comparison to other variance-based techniques [J]. Reliability Engineering & System Safety, 2015, 134: 287–296.
- [5] SARMA A D, LALL A, NANONGKAI D, et al. Randomized multi-pass streaming Skyline algorithms[J]. Proceedings of the VLDB Endowment, 2009, 2(1):85–96.
- [6] KANJ S, ABDALLAH F, DENEUX T, et al. Editing training data for multi-label classification with the k -nearest neighbor rule[J]. Pattern Analysis & Applications, 2016, 19(1):145–161.
- [7] ZHENG J, CHEN J, WANG H. Efficient geometric pruning strategies for continuous Skyline queries[J]. ISPRS International Journal of Geo-Information, 2017, 6(3):91.
- [8] 信俊昌,王国仁.无线传感器网络中 Skyline 节点连续查询算法[J].计算机学报,2012,35(11):2415–2430.(XIN J C, WANG G R. Continuous Skyline nodes query processing over wireless sensor networks[J]. Chinese Journal of Computers, 2012, 35(11):2415–2430.)
- [9] 雷婷,王涛,曲武,等.云环境下基于超球面投影分区的 Skyline 计算[J].计算机科学,2013,40(6):164–171.(LEI T, WANG T, QU W, et al. Distributed Skyline processing based on hypersphere projection partitioning on cloud environments[J]. Computer Science, 2013, 40(6): 164–171.)
- [10] ZHANG B, ZHOU S, GUAN J. Adapting Skyline computation to the MapReduce framework: algorithms and experiments[C]// Proceedings of the 16th International Conference on Database Systems for Advanced Applications. Berlin: Springer-Verlag, 2011: 403–414.
- [11] GULZAR Y, ALWAN A A, SALLEH N, et al. A model for Skyline query processing in a partially complete database[J]. Advanced Science Letters, 2018, 24(2):400–407.
- [12] 康海燕,马跃雷.差分隐私保护在数据挖掘中应用综述[J].山东大学学报(理学版),2017,52(3):16–23.(KANG H Y, MA Y L. Survey on application of data mining via differential privacy[J]. Journal of Shandong University (Natural Science), 2017, 52(3): 16–23.)
- [13] JORGENSEN Z, YU T, CORMODE G. Conservative or liberal? Personalized differential privacy [C]// Proceedings of the 2015 IEEE 31st International Conference on Data Engineering. Piscataway, NJ: IEEE, 2015: 1023–1034.
- [14] WANG Y, ZHENG B. Preserving privacy in social networks against connection fingerprint attacks [C]// Proceedings of the 2015 IEEE 31st International Conference on Data Engineering. Piscataway, NJ: IEEE, 2015: 54–65.

This work is partially supported by the National Natural Science Foundation of China (61370139), the Beijing Social Science Foundation (15JGB099, 15ZHA004).

ZHANG Shuxuan, born in 1993, M. S. candidate. Her research interests include information security.

KANG Haiyan, born in 1971, Ph. D., professor. His research interests include network security, privacy preserving.

YAN Han, born in 1994, M. S. candidate. Her research interests include information security.