doi:10.3772/j.issn.1002-0470.2015.03.003

面向社交网络分析的差分隐私保护研究综述①

王俊丽② 管 敏③ 魏绍臣

(同济大学 CAD 研究中心 上海 201804)

摘 要 阐述了数据的差分隐私保护概念,给出了差分隐私保护模型,从理论上描述了其噪声机制和组合性质,着重进述了差分隐私保护模型在社交网络发布数据隐私保护上的应用及发展,给出了差分隐私保护应用于度分布查询、子图计数、聚类系数计算、边权重计算等社交网络分析技术的实验结果。分析发现,研究差分隐私保护应重点考虑隐私预算和噪声机制,隐私预算决定了隐私保护强度,噪声机制决定了查询准确性;探讨差分隐私保护在社交网络领域的应用,是一个重要的研究方向。

关键词 差分隐私保护, 社交网络分析, 图挖掘, 统计方法

0 引 言

随着互联网的普及,社交网络在世界范围内发展极为迅速:社交网站 Facebook 的全球活跃用户数已超过 10 亿,Twitter 累计注册用户也超过 10 亿。目前,社交网络已成为覆盖用户最多、传播影响最大、商业价值最高的 Web2.0 业务。社交网络服务(social network service,SNS)是帮助人们建立社会性网络的互联网应用服务,为用户提供分享、交流信息的平台。社交网站一般都要求用户使用真实资料注册,包括姓名、邮箱、手机号等,这些个人信息和敏感数据要被社交网络收集和归档,对这些数据的挖掘利用,具有潜在的巨大价值。但是,用户的隐私信息可能会受到严重威胁,因此社交网络的隐私泄露、数据安全问题引起了越来越多的关注。为了保护用户隐私,社交网站通常让用户对自己的信息设置隐私权限。例如在 Facebook、Twitter,用户可以规定哪些人可以申请加为好友,并对浏览个人主页信息等方面进行设置权限。但这些设置的操作步骤可能过于冗余复杂,并不能引起用户警惕,同时不能提供严格的隐私保证。目前常被提及的隐私保护方法有 k-匿名、l- 多样性等,但这些方法缺少严格的攻击模型,不能抵御基于背景知识的攻击。差分隐私保护模型解决了这一问题,它对隐私泄露风险给出了严谨、定量化的表示和证明,极大地保证了数据的可用性。本文对差分隐私保护模型在社交网络领域的应用研究成果进行了综述。

1 数据的隐私保护研究的发展

数据的隐私保护问题最早由统计学家 Dalenius 在 20 世纪 70 年代末提出[1]。他认为,保护数据库中的隐私信息,就是要使任何用户(包括合法用户和潜在用户的攻击者)在访问数据库的过程中无法获取关于任意个体的确切信息[2]。从现有的研究来看,隐私保护技术大体分为三类:数据失真、数据加密和限制发布[3]。

k- 匿名和 l- 多样性是基于限制发布的泛化技术的两种隐私保护方法,比较有代表性。k- 匿名由

① 国家自然科学基金(61105047),港澳台科技合作项目(2013DFM10100),上海市科委项目(14JC1405800)和国家科技支撑计划(2012BAF12B11)资助项目。

② 女,1978 年生,博士,副研究员;研究方向:互联网数据分析研究,隐私保护等;E-mail: junliwang@tongji.edu.cn

③ 通讯作者,E-mail: 5guanmin@tongji.edu.cn

(收稿日期:2014-10-11)

Sweeney 提出,可以保证任意一条记录与另外 k-1 条记录不可区分[4,5]。k- 匿名易受到一致性攻击和背景知识攻击[6],为此,Machanavajjhala 等提出了 l- 多样性原则[7]。若一个数据表满足 k- 匿名,且每个等价类中的敏感属性至少有 l 个值,则其满足 l- 多样性原则。l- 多样性避免了一个等价类中敏感属性取值单一的情况,使得隐私泄露风险不超过 $\frac{1}{l}$。但是,l- 多样性易受到相似性攻击,而且 k- 匿名和 l- 多样性由于缺少严格的攻击模型,不能抵御基于背景知识的攻击。

2006 年 Dwork 提出了差分隐私保护(differential privacy,DP)模型,该模型解决了这一问题。差分隐私保护假设攻击者能够获得数据集中除目标记录外所有其他记录的信息,这些信息的总和可以理解为攻击者所能掌握的最大背景知识,在拥有这一最大背景知识假设下,差分隐私保护能够抵御关于背景知识的攻击。其次,它建立在坚实的数学基础之上,对隐私泄露风险给出了严谨、定量化的表示和证明,极大地保证了数据的可用性。正是由于差分隐私保护方法可以提供可证明、定量的保护力度等诸多优势,它已被广泛应用到各个领域。

本文立足于社交网络领域,对差分隐私保护下的社交网络分析技术如度分布、子图计数(三角形计数、k- 星计数、k- 三角形计数)、聚类系数等的研究现状进行了综述。

2 差分隐私保护模型

差分隐私保护是基于数据失真的隐私保护技术,该技术采用添加噪声机制使敏感数据失真但同时保证数据的有用性。它可以实现在数据集中添加或删除一条数据不会影响到查询输出结果,因此可以保证这一条记录被识别或敏感属性被泄露[8-13]。

2.1 差分隐私保护的定义与相关概念

2.1.1 基本定义

ε- 差分隐私保护是基于“邻近数据集”概念定义的。设两个数据集 D 和 D' 具有相同的属性结构,两者的对称差记作 $D\Delta D'$,$|D\Delta D'|$ 表示 $D\Delta D'$ 中记录的数量。若 $|D\Delta D'|=1$,则称 D 和 D' 为邻近数据集(neighboring database)。

定义 1[14](ε- 差分隐私保护):设有随机算法 M,P_M 为算法 M 所有可能的输出构成的集合。对于任意两个邻近数据集 D 和 D' 以及 P_M 的任何子集 S_M,若算法 M 满足

$$P_r[M(D)\in S_M]\leqslant \exp(\varepsilon)\times P_r[M(D')\in S_M] \tag{1}$$

则算法 M 提供 ε- 差分隐私保护,其中参数 ε 称为隐私保护预算。

2.1.2 相关概念

(1)隐私保护预算

隐私保护预算 ε 体现了算法所能够提供的隐私保护水平。ε 值越小,表示隐私保护水平越高。在实际应用中,ε 通常取很小的值,例如 0.01、0.1 等。

(2)敏感度

差分隐私保护是通过在查询函数的返回值中加入适量的噪声来实现的。加入噪声过多会影响结果的可用性,过少则无法提供所需的安全保证。差分隐私用敏感度来决定加入噪声量的大小。它指删除数据集中任意一条记录对查询结果造成的最大改变。在差分隐私保护模型中定义了两种常用的敏感度:全局敏感度(global sensitivity, GS)和局部敏感度(local sensitivity, LS)。

定义 2[15](全局敏感度):设有函数 $f:D\rightarrow R^d$,输入为一数据集 D,输出为一 d 维实数向量 R^d。对于任意邻近数据集 D 和 D',

$$GS_f=\max_{D,D'}\|f(D)-f(D')\|_1 \tag{2}$$

称为函数 f 的全局敏感度。

函数的全局敏感度由函数本身决定,一些函数具有较小的全局敏感度(例如计数函数,全局敏感度为 1),因此只需加入少量噪声来掩盖因一条记录被添加或删除对查询结果所产生的影响,实现差分隐私保护。但对于某些函数而言,如求平均值、聚类系数等,则往往具有较大的全局敏感度。当全局敏感度较大时,需在函数输出中加入足够大的噪声才能确保隐私安全,但会导致数据可用性差。针对这个问题,Nissim[16]定义了局部敏感度以及相关的其它概念。

定义 3[16](局部敏感度):设有函数 $f:D\rightarrow R^d$,

输入为一数据集 D,输出为一 d 维实数向量 R^d。对于给定的数据集 D 和它的任意邻近数据集 D',则

$$LS_f(D) = \max_{D'} \| f(D) - f(D') \|_1 \quad (3)$$

称为函数 f 在 D 上的局部敏感度。

局部敏感度由函数 f 及给定数据集 D 中的具体数据共同决定,与全局敏感度之间的关系可以表示为

$$GS_f = \max_D(LS_f(D)) \quad (4)$$

从式(4)可以看出,局部敏感度通常要比全局敏感度小。但是,由于它在一定程度上体现了数据集的数据分布特征,如果直接用来计算噪声大小则会泄露数据集中的敏感信息。因此,一般用局部敏感度的平滑上界(smooth upper bound)与其一起确定噪声量的大小。

定义 4[16](平滑上界):给定数据集 D 及其任意邻近数据集 D',函数 f 的局部敏感度为 $LS_f(D)$。对于 $\beta > 0$,若函数 $S: D \to R$ 满足 $S(D) \geqslant LS_f(D)$ 且 $S(D) \leqslant e^{\beta} S(D')$,则称 S 为函数 f 的局部敏感度的 β- 平滑上界。

将局部敏感度代入此函数中则可得到平滑敏感度(smooth sensitivity),即平滑敏感度 $S^*_{f,\beta}(D)$ 是 LS_f 的 β- 平滑上界,进而用于计算噪声大小。

定义 5[16](平滑敏感度):给定数据集 D 及其任意邻近数据集 D',$\beta > 0$,函数 f 的平滑敏感度为

$$S^*_{f,\beta}(D) = \max_{D'}(LS_f(D') \times e^{-\beta | D \Delta D' |}) \quad (5)$$

2.2 噪声机制

拉普拉斯机制(Laplace mechanism)[17]与指数机制(exponential mechanism)[18]是实现差分隐私保护常用的两种噪声机制。其中拉普拉斯机制适用于数值型结果的保护,指数机制适用于非数值型结果的保护。

2.2.1 拉普拉斯机制

拉普拉斯机制通过向确切的查询结果中加入服从拉普拉斯分布的随机噪声来实现 ε- 差分隐私保护。均值为 0,尺度参数为 σ 的拉普拉斯分布为 $Lap(\sigma)$,其概率密度函数为

$$p(x) = \frac{1}{2\sigma} \exp\left(- \frac{| x |}{\sigma} \right) \quad (6)$$

定义 6[17](拉普拉斯机制):给定数据集 D,设有函数 $f: D \to R^d$,其全局敏感度为 GS_f,若算法 M 满足

$$M(D) = f(D) + Lap(GS_f/\varepsilon) \quad (7)$$

则算法 M 提供 ε- 差分隐私保护。$Lap(GS_f/\varepsilon)$ 为随机噪声,服从尺度参数为 GS_f/ε 的拉普拉斯分布。

2.2.2 指数机制

由于拉普拉斯机制仅适用于数值型查询结果,而在许多实际应用中,查询结果为实体对象。对此,McSherry 提出了指数机制。

定义 7[18](指数机制):设随机算法 M,输入为数据集 D,输出为一实体对象 $r \in Range$,$q(D,r)$ 为 r 的可用性函数,用来评估输出值 r 的优劣程度,Δq 为函数 $q(D,r)$ 的敏感度。若算法 M 以正比于 $\exp\left(\frac{\varepsilon q(D,r)}{2\Delta q}\right)$ 的概率从 $Range$ 中选择并输出 r,那么算法 M 提供 ε- 差分隐私保护。

2.3 差分隐私保护的组合性质

差分隐私保护包含序列组合性与并行组合性两种重要的组合性质。

性质 1[19](序列组合性):设有算法 $M_1, M_2, \cdots, M_n$,其隐私保护预算分别为 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$,那么对于同一数据集 D,由这些算法构成的组合算法 $M(M_1(D), M_2(D), \cdots, M_n(D))$ 提供 $\sum_{i=1}^{n} \varepsilon_i$- 差分隐私保护,其提供的隐私保护水平为全部预算的总和。

性质 2[19](并行组合性):设有算法 $M_1, M_2, \cdots, M_n$,其隐私保护预算分别为 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$,那么对于不相交的数据集 $D_1, D_2, \cdots, D_n$,由这些算法构成的组合算法 $M(M_1(D_1), M_2(D_2), \cdots, M_n(D_n))$ 提供 $(\max\varepsilon_i)$- 差分隐私保护,其提供的隐私保护水平取决于预算最大者。

这两种性质可以用来判断算法是否满足差分隐私以及控制隐私预算的合理分配。

3 面向社交网络的差分隐私保护

社交网络通过使用图结构中的节点和边来进行建模社会关系。节点表示社交网络中的用户个体,边用来记录用户个体间的关系或活动[20]。差分隐私保护的定义是建立在传统数据库上的,它能够使

得在数据集中添加或删除一条数据不会影响到查询输出结果,以此保证这一条记录的敏感属性不会被泄露。本节将介绍差分隐私保护在社交网络图结构中的应用,例如邻近图的定义,一个算法提供差分隐私保护时需满足的条件,以及在差分隐私保护下社交网络分析技术的发展状况。

3.1 图的差分隐私保护

在现有的研究中,图结构一般采取两种常用的差分隐私保护标准[21],分别是节点差分隐私(node-differential privacy)和边差分隐私(edge-differential privacy)。

(1) 节点差分隐私保护

若对所有图 $G_1 = (V_1, E_1)$, 图 $G_2 = (V_2, E_2)$, V_1、V_2、E_1、E_2 分别是图 G_1 和 G_2 的顶点集合和边集合,其中 $V_2 = V_1 - x$, $E_2 = E_1 - \{(v_1, v_2) \mid v_1 = x \vee v_2 = x\}$, $x \in V_1$; 图的查询函数 Q 满足差分隐私保护则称 Q 满足节点差分隐私保护。

在节点差分隐私保护中,如果已知给定的社交网络 G, 则其邻近图 G' 的定义是:从 G 中删除或添加任意一个节点和连接该节点的所有边。攻击者不能确定是否个体节点 x 出现在图中。节点差分隐私完全保证保护了所有个体,但同时对查询强加了很大的限制。在很多情况下,要实现节点差分隐私保护是不可执行的。

(2) 边差分隐私保护

若对所有图 $G_1 = (V_1, E_1)$, 图 $G_2 = (V_2, E_2)$, V_1、V_2、E_1、E_2 分别是图 G_1 和 G_2 的顶点集合和边集合,其中 $V_2 = V_1$, $E_2 = E_1 - E_x$, $\mid E_x \mid = k$; 图的查询函数 Q 满足差分隐私保护则称 Q 满足 k- 边差分隐私保护。

在边差分隐私保护中,如果已知给定的社交网络 G, 则其邻近图 G' 的定义是:从 G 中删除或添加 k 条任意边。它保证了能以高概率确定个体节点 x 和 y 之间是否有关系。当 $k = 1$ 时,是图结构算法的研究中最广泛应用的差分隐私保护标准。边差分隐私保护比节点差分隐私保护提供的保护力度弱,度数高的节点对查询结果仍然有较高的影响,尽管它们个体之间的关系受到了保护。但保护力度对很多应用领域来说已经足够,所以与节点差分隐私保护相比,边差分隐私保护的应用更为广泛。

Task 等提出了出度差分隐私保护的概念(out-link privacy)[22]。其定义如下:若对所有图 $G_1 = (V_1, E_1)$, 图 $G_2 = (V_2, E_2)$, V_1、V_2、E_1、E_2 分别是图 G_1 和 G_2 的顶点集合和边集合,其中 $V_2 = V_1$, $E_2 = E_1 - \{(v_1, v_2) \mid v_1 = x\}$, $x \in V_1$; 图的查询函数 Q 满足差分隐私保护则称 Q 满足出度差分隐私保护。

在出度差分隐私保护中,如果已知给定的社交网络 G, 则其邻近图 G' 的定义是:从 G 中删除或添加任意一个节点及所有从该节点出发的边。同时,定义了 k- 出度差分隐私保护(k- out-link privacy),此时,邻近图 G' 的定义是:从 G 中删除或添加 k 个节点及所有从该 k 个节点出发的边。出度差分隐私的保护力度比节点差分隐私保护弱,但它通过保护节点间的关系削弱了度数高的节点对数据结果的影响,能够简化一些函数敏感度的计算和噪声的添加。

文献[23]定义了分区差分隐私保护(partition privacy)。分区图 G 的定义: $G = \{g_i\}$, g_i 为不相交的子图。若对所有图 G_1、G_2, 其中 $G_1 = G_2 - g_i$, $g_i \in G_1$, 查询函数 Q 满足差分隐私保护则称 Q 满足分区差分隐私保护。

在分区差分隐私保护中,如果已知给定的社交网络 G, 则其邻近图 G' 的定义是:从 G 中删除或添加一个子图。分区差分隐私保护比节点差分隐私提供的保护力度强,它不是针对单个节点即个体,而是针对某个社会群体提供的保护,能够计算一些无法在边差分隐私、节点差分隐私下进行求解的函数。

3.2 差分隐私保护下的社交网络分析技术

社交网络分析是指利用统计方法、图论等技术对社交网络服务中产生的数据进行定量分析,通过分析个人的网络地图可以挖掘出人们在联络、信息流动与价值交换等互动过程中潜藏的价值。下面将探讨社交网络分析的几种方法,分别是度分布查询(degree distribution)、三角形计数查询(triangle counting)、k- 星计数查询(k- star counting)、k- 三角形计数查询(k- triangle counting)、计算聚类系数(cluster coefficient)、边的权重(edge weight)。

3.2.1 度分布

度分布是被广泛研究的图特性之一,它影响着

图的结构和对图进行操作的整个过程。

定义 8[24](度频繁序列):给定图 G,用 $F(G)$(大小为 $|V(G)|$)表示图 G 的度频繁序列,则 $F(G)$ 中的第 i 个值为

$$\frac{|\{v \in V:\deg(v) = i\}|}{|V|} \tag{8}$$

其中 $\deg(v)$ 表示节点 v 的度。

Dwork 等[17]提出了对一般查询函数提供差分隐私保护的方法。该算法的思想是用户提交查询序列 Q,首先计算出真实的查询结果 $Q(I)$,I 表示要查询的数据集,然后加入一些噪声后再把结果返回给用户。其中噪声的大小取决于查询序列的全局敏感度。

命题 1[17]:令 $\widetilde{Q}$ 表示一随机算法,算法输入的数据集为 I,查询序列为 Q,序列长度为 d,$\varepsilon > 0$,输出结果为

$$\widetilde{Q}(I) = Q(I) + Lap(GS_Q/\varepsilon)^d \tag{9}$$

则称算法 $\widetilde{Q}$ 满足 ε- 差分隐私保护。其中 GS_Q 表示 Q 的全局敏感度,$Lap(\sigma)^d$ 表示从均值为 0、尺度为 σ 的拉普拉斯分布中独立随机抽取的长度为 d 的向量。

文献[25]提出了一种后处理技术,目的是在不牺牲差分隐私保护力度的情况下提高查询结果的精确度。算法的核心思想是对加噪后的查询结果予以一致性约束,找出“最接近”结果 $\widetilde{Q}(I)$ 且满足查询序列约束条件的结果进行返回。其中,“最接近”用 L_2 距离衡量,返回的结果为最小 L_2 解(Minimum L_2 Solution)。例如,当分析一个学生数据库时,需要以下查询函数的结果:学生总数 x_t;成绩为 A、B、C、D、F 的学生数 x_A、x_B、x_C、x_D、x_F;及格的学生数(成绩等于或高于 D)x_p。则查询序列为 $Q = (x_t, x_p, x_A, x_B, x_C, x_D, x_F)$,查询序列的约束为 $x_t = x_p + x_F$ 和 $x_p = x_A + x_B + x_C + x_D$,查询结果应满足此约束条件。

定义 9[25](最小 L_2 解):查询序列 Q,约束条件 γ_Q,I 为输入数据集,$\bar{q} = \widetilde{Q}(I)$,最小 L_2 解 $\bar{q}$ 满足约束条件 γ_Q 且令 $\|\bar{q} - \bar{q}\|_2$ 值最小。

文献[21]用该方法来估计图的度分布,并给出算法来计算度分布查询结果的最小 L_2 解。

3.2.2 子图计数查询

输入图为 I,给定一个查询图 H,如一个三角形、k- 三角形或 k- 星,子图计数查询的目标是返回 I 中边导出的 H 的同构图的数量。k- 三角形由共有一条公共边的 k 个三角形组成。k- 星由一个中心节点连接其余 k 个节点构成。

利用文献[17]提出的对查询函数提供差分隐私保护的方法可以求解该问题。但是和传统的基于表数据的聚集函数如求和、计数不同,子图计数查询这类函数通常有较高的全局敏感度,因此可能导致查询结果的失真较大。针对这一问题,Nissim[16]提出了局部敏感度的概念,用局部敏感度的平滑上界和局部敏感度一起确定噪声量的大小,即将局部敏感度代入 β- 平滑上界的函数中则可得到平滑敏感度,进而用于计算噪声大小。Karwa[26]将此方法扩展应用到 k- 星计数查询,并提出算法用来查询 k- 三角形个数。

假设无向图 G、G',节点数为 n,邻接矩阵为 $X = (x_{ij})$,对所有的 $i \in [n]$,$x_{ii} = 0$,a_{ij} 表示特定的一对节点 (i, j) 共享的邻居数,即 $a_{ij} = \sum_{l \in [n]} x_{il} \cdot x_{lj}$,$b_{ij}$ 表示仅和节点 i、j 中的一个连接的节点数,即 $b_{ij} = \sum_{l \in [n]} x_{il} XOR x_{lj} \circ N_{ij}$ 是节点 i 和 j 的共同邻居的集合,$N_{ij} = \{l \in [n] \mid x_{il} \cdot x_{lj} = 1\}$,则 $|N_{ij}| = a_{ij} \circ d(G, G')$ 表示图 G、G' 之间的距离,即图 G、G' 之间边不同的数量。用 f_Δ、f_{k*}、$f_{k\Delta}$ 分别表示查询三角形、k- 星、k- 三角形个数的函数,LS_Δ、LS_{k*}、$LS_{k\Delta}$ 表示查询函数 f_Δ、f_{k*}、$f_{k\Delta}$ 的局部敏感度,$S^*_{\Delta,\beta}$、$S^*_{k*,\beta}$、$S^*_{k\Delta,\beta}$ 表示查询函数 f_Δ、f_{k*}、$f_{k\Delta}$ 的平滑敏感度。

(1) 三角形计数查询:文献[16]给出了 f_Δ 的局部敏感度公式:$LS_\Delta(G) = \max_{i,j \in [n]} a_{ij}$,当距离为 t 时,$LS^{(t)}_\Delta(G) = \max_{i \neq j; i,j \in [n]} c_{ij}(t)$,其中 $c_{ij}(t) = \min\left(a_{ij} + \left\lfloor \frac{t + \min(t, b_{ij})}{2} \right\rfloor, n - 2\right)$。文献[25]证明了如果 $LS_\Delta(G) \geqslant \frac{1}{\beta}$,则 $S^*_{\Delta,\beta}(G) = LS_\Delta(G)$。

(2) k- 星计数查询[26]:$f_{k*}(G) = \sum_{i \in [n]} \binom{d_i}{k}$,$d_i$ 是节点 i 的度。$LS_{k*}(G) =$

$\max_{i\neq j;i,j\in[n]}\left(\binom{d_i-x_{ij}}{k-1}+\binom{d_j-x_{ij}}{k-1}\right)$。$LS_{k*}^{(t)}(G)=\max_{(i,j):d_i\geq d_j}LS_{ij}^{(t)}(G)$，其中 $LS_{ij}^{(t)}(G)$ 是距离为 t 时经过边 (i,j) 的 f_{k*} 的局部敏感度。Karwa 等[26]给出了 $LS_{ij}^{(t)}(G)$ 的计算方法，并证明了图中节点最大度为 $d_{\max}$，若 $d_{\max}\geq\max\left\{k,(k-1)\left(\frac{1-\beta}{\beta}\right)\right\}$，则 $S_{k*,\beta}^{*}(G)=LS_{k*}(G)$。

（3）k-三角形计数查询[26]：$LS_{k\Delta}(G)=\max_{i\neq j;i,j\in[n]}LS_{ij}(G)$，$LS_{ij}(G)=\binom{a_{ij}}{k}+\sum_{l\in N_{ij}}\left(\binom{a_{il}-x_{ij}}{k-1}+\binom{a_{lj}-x_{ij}}{k-1}\right)$，根据 $LS_{k\Delta}(G)$ 来计算拉普拉斯噪声大小，最终释放 $f_{k\Delta}$ 的查询结果。

3.2.3 聚类系数

聚类系数是表示一个图形中节点聚集程度的系数。在现实的网络中，尤其在特定的网络中，由于相对高密度连接点的关系，节点总是趋向于建立一组严密的组织关系。因此它也是反映社交网络图结构的一个很重要的性质。图中一个节点的聚类系数表示了它的相邻节点形成一个完全图的紧密程度，计算公式如下：

$$C_i=\frac{N_\Delta(i)}{N_3(i)}=\frac{N_\Delta(i)}{d_i(d_i-1)/2} \tag{10}$$

其中，C_i 表示节点的聚类系数，$N_\Delta(i)$ 表示涉及到节点的三角形个数，$N_3(i)$ 表示以节点 i 为中心的连接的三元组个数，d_i 表示节点 i 的度。

Wang 等[27]采用“分治法（divide and conquer，D&C）”的思想来实现差分隐私保护的查询。以求聚类系数为例，算法的思想是对于给定的图 G，图的目标统计函数 f，隐私参数 (ε,δ)，先把函数 f 分解成多个低复杂度的单元计算函数 $f_1,\cdots,f_m$，通过基本的数学方法如加法、减法、乘法、除法连接这些函数，给每个 f_i 的输出结果加上拉普拉斯噪声（根据函数 f_i 的敏感度和分配的预算值 ε_i 推导出），最后把这些已修改的 f_i 用上述数学方法结合起来作为最终的输出结果即函数 f 的输出结果。

3.2.4 边的权重

在社交网络中，依靠节点和边建模社会关系。因此，一条边可能会揭露个体节点之间的各种敏感信息。Costea 等在文献[28]中指出在不考虑边信息和目标节点为敏感数据的情况下，将差分隐私保护应用在边的权重上，并利用迪杰斯特拉算法求解最短路径来评估其保护的质量。

算法分为三个过程：（1）使用 Erdos-Renyi 模型生成图 $G=(V,E)$；（2）利用拉普拉斯噪声机制对图 G 的边权重加以差分隐私保护从而得到新的图 $G'=(V,E')$；（3）利用迪杰斯特拉算法计算图 G 和 G' 中的最短路径 P、P'，对其进行误差测量。测量方法为

$$err=c(P',G)-c(P,G) \tag{11}$$

其中 $c(P,G)$ 为计算 P 时所需的开销代价。

4 实验结果对比分析

4.1 实验数据来源

实验数据一般从真实社交网络，例如 Flickr、YouTube、GrQc 等和合成的图获取。常用的三种合成的图模型是 ER 随机图、WS 小世界模型、BA 无标度网络[29]。

构造 ER 随机图：由 n 个节点构成，以概率 p 连接每组节点。

构造 WS 小世界模型：（1）从规则图开始，考虑一个含有 n 个点的最近邻耦合网络，它们围成一个环，其中每个节点都与它左右相邻的 $\frac{K}{2}$ 个节点相连，K 是偶数；（2）随机化重连：以概率 p 随机地重新连接网络中的每个边，即将边的一个端点保持不变，而另一个端点取为网络中随机选择的一个节点。其中规定，任意两个不同的节点之间至多只能有一条边，并且每一个节点都不能有边与自身相连。

构造 BA 无标度网络：（1）增长：从一个具有 m_0 个节点的网络开始，每次引入一个新的节点，m_1 条新的边被增加到新节点和一些原有节点之间；（2）连接新边的原有节点的选择遵循优先连接规则：新的节点更倾向于与那些具有高连接度的节点相连接；一个新节点与一个已存在的节点 i 相连接的概率 $\prod$ i 与节点 i 的度 d_i 之间满足 $\prod^{i}=\frac{d_i}{\sum_j d_j}$。

满足差分隐私保护的算法需要在保护隐私的同时,兼顾保护后数据的可用性以及隐私预算 ε 分配的合理性。通常从算法的可伸缩性(scalability)、有用性(utility)、隐私预算 ε 、算法误差等角度进行度量。

4.2 度分布

Hay 等[21]从两方面进行了实验,分别是:(1)评估文中提出的算法的可扩展性,即根据"一致性约束"条件提高返回查询结果的精确度;(2)推断隐私保护强度与有用性之间的平衡关系。实验数据来源:真实数据集 Flickr(≈1.8M 节点), LiveJournal(≈5.3M 节点),Orkut(≈3.1M 节点), YouTube(≈1.1M 节点);合成的数据集 Random(ER 随机图,服从泊松分布 $\lambda = 10$),Power(ER 随机图,服从幂律分布 $\alpha = 1.5$)。

(1) 可扩展性

实验结果表明文献[21]的算法的时间复杂度较低,运行速度较快。处理一个节点数 200 百万的图运行时间小于 6s,相反,文献[25]的算法处理节点数 1 百万的图大约 20min,节点数 2 百万的图超过 1h。因此,前者更适用于较大的图,可扩展性更强。

(2) 有用性

用两种方法来评估精确度:Kolmogorov-Smirnoff(KS)统计和 Mallows 距离。从"偏置(Bias)和方差(variance)、准确度(accuracy)和保护力度(privacy)、准确度(accuracy)和大小(size)"三个角度来看,实验表明用"一致性约束"条件限制返回的查询结果更接近真实的查询结果,更为准确,是真实查询结果的无偏估计或接近无偏估计。

4.3 子图计数

文献[26]有两个实验目标:(1)比较实例依赖(instance-dependent)算法(即求三角形计数[16]和 k-星计数、k- 三角形计数[26])与 RHMS[30]中提出的算法;(2)评价在一定规模的图上实例依赖算法的性能。算法的精确度用平均绝对误差(median absolute error)来衡量。实验数据来源:真实数据集 GrQc、HepTh、CondMat、HepPh、AstrpPh、Enron;合成的数据集 ER 随机图($p = \log n/n, n \in [100,1000]$ 以步长 100 增加)、ER 随机图($p \in [100,1000]$ 以步长 0.1 增加, $n \in [100,1000]$ 以步长 100 增加)。

实验结果表明,Karwa 等[26]提出的三种计数查询的算法求解的噪声大小更接近于目标值 LS_f/ε,准确性较高,性能优于 RHMS[30]中的方法,后者的使用范围是有限的。实验还发现在不同类型(密集或稀疏)的图中三种计数查询的敏感度不同。k- 星计数算法的适用性较强,三角形、k- 三角形计数算法在稀疏图上的适用性较低。

4.4 聚类系数

Wang[27]比较了直接计算查询结果方法(direct approach,DA),即在查询结果上直接加入校准后的拉普拉斯噪声和 *D&C* 方法的有用性。其中为了测试不同的分解策略如何影响最终输出结果的精确度, *D&C* 方法采用两种分解策略: $D\&C_{N_3}$ 和 $D\&C_D$,两者分别将 $(N_\Delta(i), N_3(i))$ 和 $(N_\Delta(i), d_i)$ 作为参数求解聚类系数 C_i。实验数据来源:真实数据集 GrQc、Enron、Polbooks、Polblogs、YesIWell;合成的数据集 ER、WS、BA(节点数 $n = 1000$)。

实验结果表明, *D&C* 方法的有用性较高,且 $D\&C_D$ 分解策略更好;在相同的隐私阈值情况下,一次查询所有顶点的聚类系数比一个个地查询顶点的聚类系数效果更好;当图稀疏时, $D\&C_D$ 方法的优势会减小。

4.5 边的权重

Costea 等[28]从图的大小、边的权重、隐私预算值三个角度进行了多次实验,图的大小节点数范围为[128,2048],每个图都将产生 100 个节点对来计算每条最短路径的相对误差。

实验结果表明,差分隐私保护算法的精确度依赖于权重和预算值,拉普拉斯噪声与初始的边权重之间的比例,与图的大小没有显著的关系。当添加的拉普拉斯噪声和边的权重近似时,误差增加;当噪声值远小于实际的权重时,误差会显著减小。

4.6 实验结果分析

本文 4.2 至 4.5 节分别讲述了度分布、子图计数、聚类系数、边的权重上的差分隐私保护方法应用的实验结果,也简单分析了其优势与不足。下面从算法的可伸缩性、结果的有用性以及适用范围的角度,用表 1 对以上各方法的结果进行了对比分析。

表 1 实验结果分析

应用	文献	可伸缩性	有用性	图类型（密集/稀疏）
度分布	[21]	200 百万/6 秒	无偏估计	
	[25]	1 百万/20 分钟	准确度低	
子图计数	[16]		准确度较低	
	[26]	k-星计数算法的适用性强	噪声大小更接近目标值	三角形、k-三角形在稀疏图上优势降低
	[30]	可扩展性较弱		
聚类系数	[27]/DA			无法求解复杂函数
	[27]/D&C	适用性较广	准确度、效率更高	适用于图密集
边的权重	[28]		权重的赋值对结果的有用性影响最大	

5 未来发展趋势

社交网络数据的隐私保护大致分为两个方向：(1)发布数据的隐私保护；(2)挖掘数据的隐私保护。发布数据的差分隐私保护关注的是在给定的隐私保护预算下，发布查询结果的精确性。数据挖掘差分隐私保护关注的则是在实现隐私保护的前提下，挖掘模型的分类/预测准确性。针对不同的需求，出现了相应的社交网络隐私保护技术。本文着重讲述了差分隐私保护模型在发布数据隐私保护上的应用及发展状况，从研究中可以发现差分隐私保护的重点是隐私预算的分配和噪声机制的大小。隐私预算的大小决定了隐私保护的强度。预算值越小，保护力度越大。噪声机制的大小影响着查询结果的精确性，在提供足够的保护程度的情况下要使得噪声尽可能地小。噪声的大小由敏感度和隐私预算来推导出。当全局敏感度较大时，可以采用平滑敏感度来代替，但对于其他复杂的社交网络分析技术，无法有效地或准确地求出平滑敏感度，通过计算局部敏感度的平滑上界可近似取值。而利用单元计算函数间的相关性[27]与 k- 匿名相结合[31]能够进一步降低噪声大小和提高精确度，但仍需寻求最佳分配策略和最佳隐私预算分配。

此外，对于差分隐私保护模型在社交网络领域的应用，如图挖掘[32-34]、图发布[35-37]，尤其在数据相关的情况下的应用是一个重要的研究方向。文献[24]利用四叉树将原图的邻接矩阵进行分区从而重构出加入差分隐私保护后的待发布图。文献[38]利用 dk - 图模型构造匿名后的图以保持结构一致性从而提高发布图的精确度。近两年也有人在差分隐私保护的基础上提出一种新的隐私保护模型——零知识保护（zero-knowledge privacy，ZKP），并将其应用在图领域中[39,40]。

6 结 论

本文从社交网络发布数据的隐私保护角度出发，对差分隐私保护模型在该领域的研究成果进行了综述。阐述了差分隐私保护模型的基本知识如定义、性质等，介绍了社交网络分析技术，如求解度分布、子图计数、聚类系数以及边的权重如何加以差分隐私保护，并从算法的可伸缩性、有用性、误差等方面对实验数据进行了分析，最后探讨了差分隐私保护在社交网络领域的发展方向。差分隐私保护是一个能够提供定量化、强有力的保护模型，但在提高数据有用性、效率及扩展模型应用领域方面仍需进一步深入研究。

参考文献

[1] Dalenius T. Towards a methodology for statistical disclosure control. *StatistikTidskrift*, 1977, 15:222-449

[2] 熊平，朱天清，王晓峰. 差分隐私保护及其应用. 计算机学报，2014，37(1)：101-122

[3] 周水庚，李丰，陶宇飞等. 面向数据库应用的隐私保护研究综述. 计算机学报，2009，32(5)：847-861

[4] Sweeney L. k-anonymity: a model for protecting privacy.

International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5):557-570

[5] Sweeney L. Achieving *k*-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 2002, 10(5):571-588

[6] Li N H, Li T C, Venkatasubramanian S. *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. In: Proceedings of the IEEE International Conference on Data Engineering, Istanbul, Turkey, 2007. 106-115

[7] Machanavajjhala A, Gehrke J, Kifer D, et al. *l*-diversity: privacy beyond *k*-anonymity. In: Proceedings of the 22nd International Conference on Data Engineering. Atlanta, USA, 2006. 24-35

[8] Dwork C. Theory and Applications of Models of Computation. Springer Berlin Heidelberg, 2008. 1-19

[9] Dwork C. Theory of Cryptography. Springer Berlin Heidelberg, 2009. 496-502

[10] McSherry F. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In: Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. Providence, USA, 2009. 19-30

[11] Dwork C. Differential Privacy in New Settings. In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms - SODA. Austin, USA, 2010. 174-183

[12] Dwork C. The promise of differential privacy: A tutorial on algorithmic techniques. In: Proceedings of the 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science. Palm Springs, USA, 2011. 1-2

[13] 张啸剑,孟小峰. 面向数据发布和分析的差分隐私保护. 计算机学报, 2014,37(4):927-949

[14] Dwork C. A firm foundation for private data analysis. *Communications of the ACM*, 2011, 54(1):86-95

[15] Dwork C. Differential privacy. In: Proceedings of the 33rd International Colloquium on Automata, Languages and Programming. Venice, Italy, 2006. 1-12

[16] Nissim K, Raskhodnikova S, Smith A. Smooth sensitivity and sampling in private data analysis. In: Proceedings of the 39th Annual ACM Symposium on Theory of Computing. SanDiego, USA, 2007:75-84

[17] Dwork C, McSherry F, Nissim K, et al. Calibrating noise to sensitivity in private data analysis. In: Proceedings of the 3rd Conference on Theory of Cryptography. New York, USA, 2006. 265-284

[18] McSherry F, Talwar K. Mechanism design via differential privacy. In: Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science. Providence, USA, 2007. 94-103

[19] McSherry F. Privacy integrated queries: An extensible platform for privacy-preserving data analysis. *Communications of the ACM*, 2010, 53(9):89-97

[20] 刘向宇,王斌,杨晓春. 社会网络数据发布隐私保护技术综述. 软件学报, 2014, 25(3):576-590

[21] Hay M, Li C, Miklau G, et al. Accurate estimation of the degree distribution of private networks. In: Proceedings of the 9th IEEE International Conference on Data Mining, Miami, USA, 2009. 169-178

[22] Task C, Clifton C. A Guide to differential privacy theory in social network analysis. In: Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Los Alamitos, USA, 2012. 411-417

[23] Task C, Clifton C. State of the Art Applications of Social Network Analysis. Springer International Publishing, 2014. 139-161

[24] Chen R, Fung B C M, Philip S Y, et al. Correlated network data publication via differential privacy. *The VLDB Journal*, 2014, 23(4):653-676

[25] Hay M, Rastogi V, Miklau G, et al. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 2010, 3(1-2): 1021-1032

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

[27] Wang Y, Wu X, Zhu J, et al. On learning cluster coefficient of private networks. *Social network analysis and mining*, 2013, 3(4): 925-938

[28] Costea S, Barbu M, Rughinis R. Qualitative analysis of differential privacy applied over graph structures. In: Proceedings of the 11th International RoEduNet Conference, Iasi, 2013. 1-4

[29] Costa L F, Rodrigues F A, Travieso G, et al. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 2007, 56(1): 167-242

[30] Rastogi V, Hay M, Miklau G, et al. Relationship privacy: output perturbation for queries with joins. In: Proceedings of the 28th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of database systems, Providence, USA, 2009. 107-116

[31] Soria-Comas J, Domingo-Ferrer J, Sánchez D, et al. Enhancing data utility in differential privacy via microaggregation-based *k*-anonymity. *The VLDB Journal*, 2014, 23(5):1-24

[32] Clifton C, Tassa T. On syntactic anonymity and differential privacy. In: Proceedings of IEEE 29th International Conference on Data Engineering Workshops, Los Alamitos, USA, 2013. 88-93

[33] Soria-Comas J, Domingo-Ferrer J, Sánchez D, et al. Improving the utility of differentially private data releases via *k*-anonymity. In: Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, Washington, DC, USA, 2013. 372-379

[34] Yuan G, Zhang Z, Winslett M, et al. Low-rank mechanism: optimizing batch queries under differential privacy. *Proceedings of the VLDB Endowment*, 2012, 5(11): 1352-1363

[35] Zhang J, Zhang Z, Xiao X, et al. Functional mechanism: regression analysis under differential privacy. *Proceedings of the VLDB Endowment*, 2012, 5(11): 1364-1375

[36] Mohan P, Thakurta A, Shi E, et al. GUPT: privacy preserving data analysis made easy. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, New York, USA, 2012. 349-360

[37] Xu J, Zhang Z, Xiao X, et al. Differentially private histogram publication. *The VLDB Journal*, 2013, 22(6): 797-822

[38] Sala A, Zhao X, Wilson C, et al. Sharing graphs using differentially private graph models. In: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference, New York, USA, 2011. 81-98

[39] Gehrke J, Liu E, Pass R. Theory of Cryptograph Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2011. 432-449

[40] Shoaran M, Thomo A, Jens H. Weber-Jahnke. Zero-knowledge private graph summarization. In: Proceedings of 2013 IEEE International Conference on Big Data. Santa Clara, USA, 2013. 597-605

A survey on differential privacy research for social network analysis

Wang Junli, Guan Min, Wei Shaochen

(CAD Center Group, Tongji University, Shanghai 201804)

Abstract

The concept of differential privacy protection of data is interpreted. The differential privacy model, and its noising mechanism and combination properties, are theoretically described and anlayzed. The application of the differential privacy model to social network data's privacy protection and its development are emphatically reviewed with a rigorous, quantitative representation, and the experimental results of differential privacy applications to the social network analysis techniques of degree distribution inquiry, Subgraph counting, clustering coefficient computation and edge weight computing are given. It is concluded from analysis that the privacy budget and the noising mechanism are the main factors to differential privacy (the former determines the privacy protection intensity, while the latter determines the inquiring accuracy), and exploring the application of the differential privacy protection to the social network field is the main future research direction.

Key words: differential privacy, social network analysis, graph mining, statistical method