社会网络数据的三级隐私保护发布算法研究

张 蕊[1,2] 瞿彬彬[1] 张吉昕[2]

(华中科技大学计算机学院 武汉 430070)[1] (武汉理工大学计算机学院 武汉 430070)[2]

摘 要 近年来,社会网络中的隐私保护得到了诸多关注。攻击者可以应用相关背景知识对发布的社会网络进行攻击,从而导致用户的隐私被泄露。已有工作通常只考虑了结点泄露和边泄露的情况,而忽略了攻击者可能通过识别用户的敏感信息来进行攻击。针对现有问题,提出了特征泄露的概念,并进行了理论分析。在此基础上,创造性地提出了三级隐私保护的概念,建立了隐私保护模型 *k-s* 图,并给出了 *k-s* 算法来生成 *k-s* 图。理论分析和实验结果表明,*k-s* 算法是正确有效的。

关键词 数据发布,社会网络,三级隐私保护,特征泄露

中图法分类号 TP309 **文献标识码** A

Three Level Privacy Protection in Social Networks

ZHANG Rui[1,2] QU Bin-bin[1] ZHANG Ji-xin[2]

(School of Computer Science & Technology, Huazhong University of Science and Technology, Wuhan 430070, China)[1]

(School of Computer Science & Technology, Wuhan University of Technology, Wuhan 430070, China)[2]

Abstract Serious concerns on privacy protection in social networks have been raised in recent years. When an adversary uses some types of background knowledge to conduct an attack, an individual's privacy may be threatened. In pioneer work, node disclosure and edge disclosure were studied. However, there is little effort to deal with sensitivity disclosure. To address it, sensitivity disclosure was formally defined and theoretically analyzed. Taking into account all these disclosures, three level privacy protection was proposed, and *k-s* graph was presented as the solution. To generate *k-s* graph, *k-s* algorithm is both theoretically correct and experimentally effective.

Keywords Data publishing, Social networks, Three level privacy protection, Sensitivity disclosure

1 引言

如今,越来越多的人加入不同的在线社会网络,比如 Facebook、Twitter 和新浪微博等。丰富的社会网络数据具有巨大的分析研究价值。与此同时,发布社会网络数据却面临着隐私泄露的危险。这往往导致研究机构和服务提供商不愿公布收集到的数据,而仅仅授权给少部分自己信任的个人或团体进行研究。但这些研究的结果不能复现,其必将阻碍科学研究的发展。

要避免此类数据分享障碍,必须对社会网络数据进行隐私保护下的发布。很容易想到,可以去掉社会网络数据中的姓名、身份证号等标识信息,而将社会网络的结构(图)和其他的结点信息以及边信息发布。这就是简单匿名策略[1-8]。这样可以避免攻击者通过身份信息直接识别目标用户,但仅仅这样无法抵御具有一定背景知识的攻击者[2-6]。比如一个攻击者可能知道目标结点在该社会网络中有 4 个好友,如果此时发布的社会网络中只有一个结点的度数是 4,那么目标就暴露了,甚至攻击者可能注入式攻击[3]。即在社会网络数据发布前,攻击者就加入到该社会网络中,有意识地和一些目标用户联系,并创造出一个独特的子图。如果简单匿名发布,攻击者则可以轻易地识别出这个子图,从而确定目标进行攻击。如果攻击者运用的背景知识是社会网络的结构信息,可称为结构攻击。攻击者可能是识别结点,也可能是识别边,而结构攻击中,最具破坏力的是进行子图匹配攻击,这引起了研究者的关注[7,8]。

许多研究都忽略了一点,即攻击者的背景知识并不仅仅局限于结构知识。即使匿名发布的社会网络能够抵御结构攻击,也可能无法抵御一个同时具有结构背景知识和属性背景知识的攻击者。比如对于目标所在子图,可以匹配多个,但这多个子图中,只有一个子图中的结点符合相关属性,此时攻击者可以立即识别出该用户。另外,并不是只有识别点或者识别边才能进行攻击,攻击者完全可以通过识别边或者结点的敏感属性来进行攻击。比如攻击者并不能确定目标结点是谁,但它确定了目标结点有心脏病,那它很可能据此做出损害目标的事情。

对于攻击者可能具有结点属性背景知识,文献[8]对此有所考虑,但没有考虑边属性背景知识情况。而且其采取的简单分组的办法并不能抵御针对点或边的敏感属性的攻击。本文不仅对攻击者的背景知识给予最强假设,即假定攻击者既具备目标的结构知识,又具备相关属性知识,而且将攻击者可

到稿日期:2011-05-17 返修日期:2011-08-04 本文受国家自然科学基金(60803130)资助。

张 蕊(1977-),女,博士生,讲师,主要研究方向为社会网络、数据挖掘,E-mail:zhangrui@whut.edu.cn;瞿彬彬(1970-),女,博士,副教授,主要研究方向为数据挖掘、软件测试,E-mail:bbqu@mail.hust.edu.cn(通信作者);张吉昕(1987-),男,硕士生,主要研究方向为数据挖掘。

能实施的攻击，除了点泄露和边泄露外，扩充到特征泄露，从而提出了三级隐私保护的概念。并且在此基础上进行了理论和实验分析。

2 问题定义

为了更好地理解所讨论的问题，先来看一个例子。另外，为了行文方便，在本文中，“社会网络”、“网络”和“图”可无区别地使用。

对图 1 所示的原始图，去掉每个目标结点信息中的“姓名”，用标识符 a、b、c、d、e 来替代，这就是简单匿名。此时，如果攻击者想攻击 Helen，则不能直接识别出攻击目标，但这不足以抵御具有一定背景知识的攻击者。比如，如果攻击者知道 Helen 的度数为 4，就能识别出结点 e 是 Helen。这是结点泄露。此时，攻击者应用的背景知识是度数，属于结构信息。如果用户通过某些背景知识，已经识别出了结点 e 是 Helen，结点 b 是 Tom，从而确定二者之间有边。这就是边泄露。

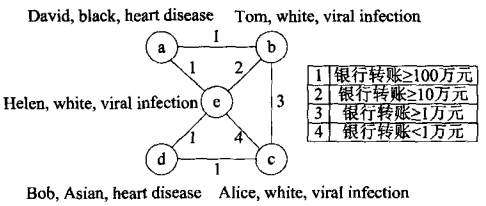


图 1 简单匿名保护

但我们注意到一个事实：即使攻击者不能识别出边或者目标结点，也仍可能获得目标结点的隐私信息。比如即使攻击者不能区分 David 是结点 a 还是结点 d，他也可以判断出该目标结点一定患有心脏病，或者攻击者发现 David 只和白种人交往；或者攻击者不能确定 David 和 Helen 的联系是边 (a,e) 还是 (d,e)，却发现这些边都是 100 万元以上的巨额银行转账等；又或者攻击者不能确定 David 和哪些用户有边，但发现 David 和所有人的联系都是 100 万元以上的巨额银行转账等。而这些信息，都可能被用户视为隐私。

为了进行一般的讨论，将社会网络建模成一个简单无向图 $G=(V,E)$，其中 V 为结点集合，$E=N\times N$ 为边集合。每个结点分别代表一个人，而边就是人与人间的关系和交互。每个结点 $v\in V$ 有一些属性来修饰或者信息来关联，记作 $I(v)$。每条边 $e\in E$ 有一些属性来修饰或者信息来关联，记作 $I(e)$。存在 $S(v)\in I(v)$ 或者 $S(e)\in I(e)$ 是敏感属性，这类属性可能是疾病、收入或联系方式等。以下讨论都建立在简单匿名基础上，即对任意 $v\in V$，只考虑不含标识信息的 $I'(v)$，亦即设 $U(v)$ 为唯一标识信息，$I'(v)=I(v)-U(v)$。

攻击者若能识别目标结点或边，或者确定相关的敏感属性取值，都将导致隐私泄露。其中，对特征泄露定义如下。

定义 1(特征泄露) 给定图 $G=(V,E)$，对任意 $v\in V$，存在结点的关联信息 $I(v)$。而 $U(v)$ 为唯一标识信息，$U(v)\subset I(v)$。$S(v)$ 为结点 v 的敏感信息，$S(v)\in I(v)-U(v)$。对任意 $e\in E$，存在边的关联信息 $I(e)$。$S(e)$ 为敏感信息，$S(e)\in I(e)$。如果攻击者应用背景知识，对敏感系数 s，出现以下 4 种情况之一，即发生特征泄露。

1. 能以大于 $1/s$ 的概率确定 $S(v)\in I(v)$；

2. 对 v 的所有相邻结点 $\cup u$ 构成的关联信息集合 $\cup I(u)$ 和某敏感信息 $S(u)$，能以大于 $1/s$ 的概率确定 $S(u)\in\cup I(u)$；

3. 能以大于 $1/s$ 的概率确定 $S(e)\in I(e)$；

4. 对与 v 关联的所有边 $\cup e_v$，能以大于 $1/s$ 的概率确定 $S(e)\in\cup I(e_v)$。

可发现，前面例子里 David 的隐私泄露正好对应了特征泄露定义中的 4 种类型。分析这 4 种类型，可得如下定理。

定理 1 在特征泄露的 4 种类型中，如果不发生类型 1 的特征泄露，就不会发生类型 2 的特征泄露；如果不发生类型 3 的特征泄露，就不会发生类型 4 的特征泄露。

证明：不发生类型 1 的特征泄露，即对任意目标结点 v 和敏感系数 s，攻击者不能以大于 $1/s$ 的概率确定 $S(v)$。也就是说，对于任意目标结点 v，必然至少有 k 个不能区分的结点 $v_1,v_2,\cdots,v_k$ 可以映射到 v，而且 $|S(v_1,v_2,\cdots,v_k)|\geqslant s$，注意 $s\leqslant k$。因此对于结点 v 的任一相邻结点 u 和某敏感信息 $S(u)$，必然至少有 k 个不能区分的结点 $u_1,u_2,\cdots,u_k$ 可以映射到 u，而且 $|S(u_1,u_2,\cdots,u_k)|\geqslant s$。不妨记 v 的所有相邻结点为 $\cup u$，则至少有 $k|\cup u|$ 个不能区分的结点可以映射到 $\cup u$，而且 $|\cup u||S(u_1,u_2,\ldots,u_k)|\geqslant|\cup u|s$。显然 $|\cup u|\geqslant 1$，所以 $|\cup u||S(u_1,u_2,\cdots,u_k)|\geqslant s$。进而对于任意目标结点 v 的所有相邻结点 $\cup u$，攻击者均不能以大于 $1/s$ 的概率确定 $S(u)\in\cup I(u)$。也就是说，如果不发生类型 1 的特征泄露，就不会发生类型 2 的特征泄露。同理可证如果不发生类型 3 的特征泄露，就不会发生类型 4 的特征泄露。

如果发布的社会网络不会发生结点泄露、边泄露和特征泄露，即称为三级隐私保护发布。准确地说，三级隐私保护发布的图是 k-s 安全的。

定义 2(k-s 安全) 给定图 $G=(V,E)$，对任意 $v\in V$，存在目标结点的关联信息 $I(v)$。$S(v)$ 为目标结点的敏感信息，$S(v)\in I(v)$。对任意 $e\in E$，存在边的关联信息 $I(e)$。$S(e)$ 为敏感信息，$S(e)\in I(e)$。假设 $G_{k,s}$ 为匿名发布的图。图 $G_{k,s}$ 是 k-s 安全的，当且仅当 $G_{k,s}$ 满足以下条件：

1)结点安全 对任意 $v\in V$，攻击者不能以大于 $1/k$ 的概率识别出 v；

2)链接安全 对任意结点 v 和 v'，攻击者不能以大于 $1/k$ 的概率判断出 v 和 v' 之间的路径为某值；

3)特征安全 对任意 $v\in V$，攻击者不能以大于 $1/s$ 的概率识别 $S(v)$；或对任意 $e\in E$，攻击者不能以大于 $1/s$ 的概率识别 $S(e)$。

注意，根据定理 1 和定理 2，要保障特征安全，避免第一类和第三类特征泄露即可。

3 k-s 隐私保护模型

为了解决特征泄露，我们采取扰乱的方法。

定义 3(扰乱) 给定图 $G=(V,E)$，$S(v_1),S(v_2),\cdots,S(v_n)$ 分别为图 G 中目标结点 $v_1,v_2,\cdots,v_n$ 的敏感信息。这里 $2\leqslant n\leqslant|V|$，而且 $S(v_1)\neq S(v_2)\neq\cdots\neq S(v_n)$。如果对目标结点 v_1 的敏感信息 $S(v_1)$，引入敏感信息 $S(v_2),\cdots,S(v_n)$，形成一个复合的敏感信息 $\{S(v_1),S(v_2),\cdots,S(v_n)\}$ 来取代 $S(v_1)$，则称 $S(v_1)$ 被 $S(v_2),\cdots,S(v_n)$ 进行了 $n-1$ 次扰乱。特别地，如果目标结点 v_1 的敏感信息 $S(v_1)$ 没有被扰乱，也可以将其称 $S(v_1)$ 进行了 0 次扰乱，也就是说本文中 $S(v_1)=\{S(v_1)\}$。下同。类似地，对边的敏感信息 $S(e_1),S(e_2),\cdots,S(e_n)$，可以定义 $S(e_1)$ 被 $S(e_2),\cdots,S(e_n)$ 进行了 $n-1$ 次扰乱，变成 $\{S(e_1),S(e_2),\cdots,S(e_n)\}$。

定义 4(扰乱代价) 对结点集合 set_v 进行扰乱得到集合 set_v',对任意 $v \in set_v$,对应点 $v' \in set_v'$ 为其扰乱后的结果,那么 $|S(v')-S(v)|$ 为结点 v 被扰乱的次数。因此,该结点集合的扰乱代价为 $|S(set_v')-S(set_v)|$。类似地,可以定义边集合的扰乱代价为 $|S(set_e')-S(set_e)|$。

从该定义可以看到,对于一个有 n 个结点的集合 set_v,对 1 个目标结点进行 $n-1$ 次扰乱,和对 $n-1$ 个目标结点进行 1 次扰乱,其代价是相同的,都是 $n-1$。如果扰乱前 $|S(set_v)|=1$,那么两种办法都可以使得扰乱后 $|S(set_v')|=n$。

定义 5(s-分组) 设图 G 中的多个结点构成一个结点分组 $group_v$,s 为给定的阈值,G 中结点的敏感信息集合记为 $S(V(G))$,要求 $1 \leqslant s \leqslant |S(V(G))|$。记 $group_v$ 中结点的敏感信息集合为 $S(group_v)$,如果 $|S(group_v)| \geqslant s$,则称 $group_v$ 为 s-分组。类似地,可以对边定义 s-分组。

下面证明图的任何点或边的分组经过扰乱,都可以变成 s-分组。

定理 2 对图 G 的任一结点或边的分组,对于给定的 $s(1 \leqslant s \leqslant \min(|S(V(G))|, |S(E(G))|)$,通过代价不超过 $s-1$ 的扰乱,都可以变成 s-分组。

证明:先对结点分组证明如下。设该分组为 $group_v$。如果相应的敏感信息集合 $|S(group_v)| \geqslant s$,则已是 s-分组,此时扰乱代价为 0。如果 $|S(group_v)| < s$,记此时的 $|S(group_v)|=t$,那么任选一目标结点 v,从 $S(V(G))-S(group_v)$ 中选取 $s-t$ 个值对其进行扰乱,则得到 $|S(group_v)|=t+s-t=s$,所以此时已是 s-分组,扰乱代价为 $s-t$。因为 $t \geqslant 1$,所以扰乱代价不超过 $s-1$。类似地,可以对边的分组进行证明。

下面考虑三级隐私保护下发布的图。首先给出图的同构点和同构边的定义。

定义 6(图的同构点与同构边) 图 $G=(V,E)$ 和 $G'=(V',E')$ 是两个图,其中 $|V|=|V'|$。图 G 和 G' 是同构的,如果存在一个双射 $h: V(G) \to V(G')$ 使得 $(u,v) \in E$,当且仅当 $(h(u),h(v)) \in E'$。此时称 u 和 $h(u)$、v 和 $h(v)$ 分别是同构点,(u,v) 和 $(h(u),h(v))$ 是同构边。

在此基础上,定义了 k-s 图。

定义 7(k-s 图) 如果图 G 包含 k 个离散的子图 $g_1, g_2, \cdots, g_k$,其中任意 g_i 和 g_j 是同构的($i \neq j$)。将 $g_1, g_2, \cdots, g_k$ 中的同构点划分成组,分组发布结点信息;将 $g_1, g_2, \cdots, g_k$ 中的同构边划分成组,分组发布边的信息,如果所有的点分组和边分组都是 s-分组,则称该图是 k-s 图。

定理 3(健壮性) 一个 k-s 图是 k-s 安全的。

证明:如果一个图由 k 个离散的连通同构子图组成,那么它是结点安全和链接安全的。即一个 k-s 图一定是结点安全和链接安全的。类似证明可见参考文献[8],此处略。

现在证明一个 k-s 图是特征安全的。取任意一个目标结点 v,根据 k-s 图的定义,v 必属于某个结点分组 $group_v$。显然,$|group_v| \geqslant k$ 且该组中结点互为同构点。也就是说,$group_v$ 中结点都可以映射为 v。而 $group_v$ 为 s-分组,即 $|S(group_v)| \geqslant s$,所以攻击者不能以大于 $1/s$ 的概率识别 $S(v)$。故 k-s 图一定是结点特征安全的。同理可证,k-s 图一定是边特征安全的。所以 k-s 图一定是特征安全的。

现在来考虑匿名代价,即如何衡量从原始图到匿名发布的图之间的信息损失。本方法可能导致的信息损失可分为两类。第一类是在将原始图划分成 k 个离散的连通同构子图时,因为增减边而产生的信息损失。第二类是对相应的结点或边进行扰乱而产生的信息损失。匿名代价用修改距离定义如下。

定义 8(修改距离) 图 G 和图 $G_{k\text{-}s}$ 的匿名代价 $ED(G, G_{k\text{-}s})=||E(G_{k\text{-}s})|-|E(G)||+|S(V(G_{k\text{-}s}))-S(V(G))|+|S(E(G_{k\text{-}s}))-S(E(G))|$。

显然,应当尽可能地使匿名代价最小。这样得到的匿名发布的图和原始图之间的信息损失才最小。

4 k-s 算法

4.1 算法框架

k-s 算法如图 2 所示。第一步得到的由 k 个离散的连通同构子图构成的图的示例如图 3 所示。这 3 个子图的同构点构成的结点映射表如图 4 所示,对应的边映射表如图 5 所示。对应形成未处理的点分组和边分组分别如图 6、图 7 所示。

输入:原始图 G,参数 k(k≥2),s(1≤s≤k)

输出:三级隐私保护的匿名图 $G_{k\text{-}s}$,相关的结点分组和边分组

1. 由图 G 和参数 k,得到由 k 个离散的连通同构子图 $g_1, g_2, \cdots, g_k$ 组成的图 $G_{k\text{-}s}$、结点映射表 VM 和边映射表 EM;
2. for VM 中的每一行 do
3. 形成结点分组 $group_v$;
4. if diff=s−|S($group_v$)|>0
5. 从 S(V($G_{k\text{-}s}$))−S($group_v$)挑选 diff 个对 $group_v$ 进行扰乱;
6. end if
7. end for
8. for EM 中的每一行 do
9. 形成边分组 $group_e$;
10. if diff=s−|S($group_e$)|>0
11. 从 S(E($G_{k\text{-}s}$))−S($group_e$)挑选 diff 个对 $group_e$ 进行扰乱;
12. end if
13. end for
14. 返回 $G_{k\text{-}s}$、$\bigcup group_v$ 和 $\bigcup group_e$;

图 2 k-s 算法

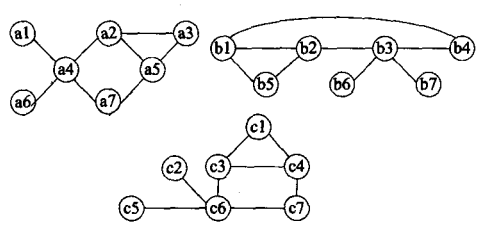


图 3 由 k 个离散的连通同构子图构成的图 $G_{k\text{-}s}$

g_1	g_2	g_3
a1	b6	c2
a2	b2	c3
a3	b5	c1
a4	b3	c6
a5	b1	c4
a6	b7	c5
a7	b4	c7

图 4 结点映射表

g_1	g_2	g_3
(a1,a4)	(b3,b6)	(c2,c6)
(a2,a3)	(b2,b5)	(c1,c3)
(a2,a4)	(b2,b3)	(c3,c6)
(a2,a5)	(b1,b2)	(c3,c4)
(a3,a5)	(b1,b5)	(c1,c4)
(a4,a6)	(b3,b7)	(c5,c6)
(a4,a7)	(b3,b4)	(c6,c7)
(a5,a7)	(b1,b4)	(c4,c7)

图 5 边映射表

结点分组	结点信息
{a1,b6,c2}	{I(a1),I(b6),I(c2)}
{a2,b2,c3}	{I(a2),I(b2),I(c3)}
{a3,b5,c1}	{I(a3),I(b5),I(c1)}
{a4,b3,c6}	{I(a4),I(b3),I(c6)}
{a5,b1,c4}	{I(a5),I(b1),I(c4)}
{a6,b7,c5}	{I(a6),I(b7),I(c5)}
{a7,b4,c7}	{I(a7),I(b4),I(c7)}

图 6　未处理的结点分组

边分组	边信息
{(a1,a4),(b3,b6),(c2,c6)}	{I((a1,a4)),I((b3,b6)),I((c2,c6))}
{(a2,a3),(b2,b5),(c1,c3)}	{I((a2,a3)),I((b2,b5)),I((c1,c3))}
{(a2,a4),(b2,b3),(c3,c6)}	{I((a2,a4)),I((b2,b3)),I((c3,c6))}
{(a2,a5),(b1,b2),(c3,c4)}	{I((a2,a5)),I((b1,b2)),I((c3,c4))}
{(a3,a5),(b1,b5),(c1,c4)}	{I((a3,a5)),I((b1,b5)),I((c1,c4))}
{(a4,a6),(b3,b7),(c5,c6)}	{I((a4,a6)),I((b3,b7)),I((c5,c6))}
{(a4,a7),(b3,b4),(c6,c7)}	{I((a4,a7)),I((b3,b4)),I((c6,c7))}
{(a5,a7),(b1,b4),(c4,c7)}	{I((a5,a7)),I((b1,b4)),I((c4,c7))}

图 7　未处理的边分组

4.2　分析和讨论

首先分析算法的正确性。

定理 4　设 G 为 k-s 算法的输入，$G_{k,s}$ 为 k-s 算法的输出。那么，$G_{k,s}$ 是 k-s 安全的。

证明：因为 k-s 算法的输出图 $G_{k,s}$ 由 k 个离散的同构子图构成，并且对应的结点分组和边分组都是 s-分组，所以该图是 k-s 图。而根据定理 3，一个 k-s 图是 k-s 安全的，故 $G_{k,s}$ 是 k-s 安全的。

下面讨论算法的计算代价。k-s 算法是一个算法框架，实际上任何可以将图 G 划分成 k 个离散的连通同构子图的方法都可以应用其中。而 k-s 算法的计算代价可以分为两部分，一部分是将图 G 划分成 k 个离散的连通同构子图的计算代价，另一部分是扰乱的计算代价。前者参考文献[8]进行了详尽分析，本文不再赘述。现重点讨论扰乱的时间复杂度。

定理 5（k-s 算法的扰乱代价）　给定输入图 G，k-s 算法的扰乱代价最多为 $(s-1)(\lfloor |V(G)|/k \rfloor -1)+(s-1)(\lfloor |E(G)|/k \rfloor -1)$，其中 $s=\min(k,|S(V(G)|,|S(E(G)|)$。

证明：根据 k-s 算法，在图 G 作为输入的情况下，算法产生的结点分组有 $\lfloor |V(G)|/k \rfloor$ 个，边分组有 $\lfloor |E(G)|/k \rfloor$ 个。最坏的情况下，取 $s=\min(k,|S(V(G)|,|S(E(G)|)$，只有一个组的敏感属性取值为全集，其他组的敏感属性取值均为同一个值。此时要对 $\lfloor |V(G)|/k \rfloor -1$ 个分组进行扰乱，每组的扰乱代价为 $s-1$。此时所有结点分组的扰乱代价为 $(s-1)(\lfloor |V(G)|/k \rfloor -1)$。同理，边分组的扰乱代价为 $(s-1)(\lfloor |E(G)|/k \rfloor -1)$。因此，$k$-$s$ 算法的扰乱代价最多为 $(s-1)(\lfloor |V(G)|/k \rfloor -1)+(s-1)(\lfloor |E(G)|/k \rfloor -1)$。

因此最坏情况下，k-s 算法的扰乱部分的时间复杂度为 $O(\max(|V(G)|,|E(G)|))$。

5　实验

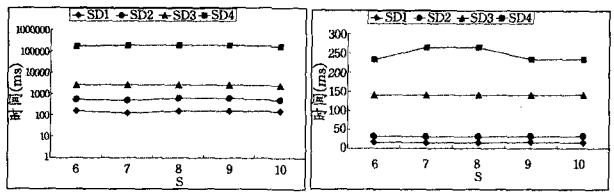
对于本文提出的 k-s 算法，用 C++实现。所有程序均在双核 2.27GHz CPU、4G 内存的机器上运行。实验机器的操作系统是 Windows Server 2008 企业版。开发环境是 minGW Developer Studio。

实验是在 Speed Dating 数据集上进行的。经过数据清理后，Speed Dating 有 552 个结点和 8388 条边。本图的平均度数是 15.2。本实验中，每个结点有 21 个属性，代表着 4 类信息，包括 age，race，field，hobbies 和 income。其中，hobbies 包括 17 个属性。指定 income 作为结点敏感属性。每条边代表两人之间的一个“date”。边的标号为 match，1 代表两人都对约会满意，0 则表示至少一方不满意。int_corr 为两人兴趣的相关系数。整个数据集中，int_corr 最大为 0.91，最小为 −0.83。指定 int_corr 为边的敏感属性。此时为无向图，共 4194 条边。实验中，将 income 分成 11 个区间，分别用 0 到 10 作为区间标号。将 int_corr 分成 10 个区间，分别用 0 到 9 作为区间标号。

为了更好地对比结点数、边数和平均度数对实验结果的影响，对 Speed Dating 数据集进行抽取，形成了 4 个数据集，分别命名为 SD1，SD2，SD3 和 SD4。其中，SD1 和 SD2 结点相同，SD1 随机抽取了部分边；SD3 和 SD4 结点相同，SD3 随机抽取了部分边。它们的特征如表 1 所列。

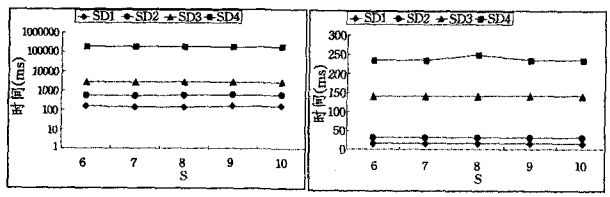
表 1　不同数据集的特征

数据集	平均度数	结点数	边数
SD1	7.65	75	287
SD2	13.44	75	504
SD3	8.23	552	2272
SD4	15.20	552	4194



(a)总运行时间　　(b)扰乱部分的运行时间

图 8　$k=12$ 时，不同 s 取值下的运行时间



(a) 总运行时间　　(b)扰乱部分的运行时间

图 9　$k=20$ 时，不同 s 取值下的运行时间

图 8(a)和图 9(a)分别显示了 $k=12$ 和 $k=20$ 时各数据集上的运行时间。可以发现，k 和 s 的改变对运行时间影响很小，而且 SD1，SD2，SD3 和 SD4 的运行时间依次递增。特别是 SD4 的运行时间比 SD3 高了一个数量级。考虑到 SD3 和 SD4 结点数相同，可见平均度数对运行时间的影响远大于结点数。单独考察算法扰乱部分的时间，从图 8(b)和图 9(b)可以看出，仍然是 SD1＜SD2＜SD3＜SD4，但不存在数量级上的差异。考虑到 SD2 和 SD4 的平均度数接近，可以发现扰乱

（下转第 188 页）

因而在工程实际中具有良好的应用前景。本算法不足之处在于，对一些“平底锅”型函数，无法保证其稳定性。其原因一方面在于增量 SVR 响应面拟合时尚无法做到十分精确，另一方面在于样本采样不能自适应建模情况。

未来的任务是在模型的样本集合中使用结合本算法特点的采样方式，构建一种自适应的响应面搜索方法。

参 考 文 献

[1] Younis A, Gu J C, Dong Z M, et al. Trends, features, and tests of common and recently introduced global optimization methods [C]// The 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference. Victoria, Canada, 2008: 691-718

[2] 吴义忠，陈立平. 多领域物理系统的仿真优化方法[M]. 北京：科学出版社，2011：248-249

[3] An Jin-long, Yang Qing-xin, Ma Zhen-ping, et al. A global optimization algorithm based on Support Vector Machines for electromagnetic inverse problem[C]// Automation Congress, 2008. WAC 2008. World, 2008: 1-5

[4] Chapelle O, Haffner P, Vapnik V N. Support Vector Machines for Histogram-based Image Classification[J]. IEEE Trans. on Neural Networks, 1999, 10(5): 1055-1064

[5] 吕勇波. 基于支持向量机与遗传算法的结构优化研究[J]. 华中科技大学，2007

[6] Gunn S R. Support vector machines for classification and regression[M]. Faculty of Engineering and Applied Science Department of Electronics and Computer Science, 1998: 63

[7] Jones D R. DIRECT Global optimization algorithm[J]. Encyclopedia of Optimization. Dordrecht: Kluwer Academic Publications, 2001: 25

[8] Myers R H, Montgomery D C. Response Surface Methodology-Process and Product Optimization Using Designed Experiments [J]. New York: Wiley Press, 2002: 308

[9] Jones D R, Perttunen C D, Stuckman B E. Lipschitz optimization without Lipschitz Constant [J]. Journal of Optimization Theory and Applications, 1993, 79(1): 201-210

(上接第 167 页)

部分的时间和结点数与边数存在正相关关系，但是平均度数不是重要影响因素。这与我们对 k-s 算法扰乱部分的时间复杂度的分析是一致的。

可以注意到，扰乱部分的运行时间要比将图 G 划分成 k 个离散的连通子图的时间快一个或者多个数量级。也就是说，整个算法的运行时间主要取决于后者。这提醒我们，如果能高效地将图 G 划分成 k 个离散的连通子图，将极大地改善整个算法的性能。

显然地，总运行时间减去扰乱部分的运行时间，就是将原图划分成 k 个离散的子图的运行时间。我们没有单独给出分析，是因为这部分选用的是参考文献[8]的算法，虽然实验的数据集不同，但结论是一致的。

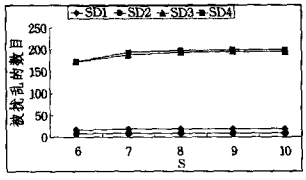


图 10 $k=12$ 时，不同 s 取值时扰乱数目(点数和边数)

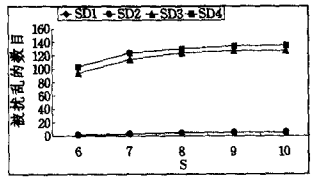


图 11 $k=20$ 时，不同 s 取值时扰乱数目(点数和边数)

另外，从图 10 和图 11 可以看出，s 改变，被扰乱的点和边的数量没有明显改变。k 变大，被扰乱的点和边的数量有减少趋势。这是因为 k 越大，点分组和边分组中敏感属性的取值个数可能会增加。可以看见相同 k 和 s 取值的情况下，SD1 和 SD2、SD3 和 SD4 被扰乱的点和边的数量分别比较接近。这里没有发现平均度数的影响。

结束语 本文提出了防止结点泄露、边泄露和特征泄露的三级隐私保护的概念。首次正式提出了社会网络中特征泄露的概念，并对此进行了理论探讨。在此基础上，提出了 k-s 算法作为解决方案，该算法框架具有一定的灵活性。最后，我们在真实数据集 Speed Dating 上进行了实验，实验结果证明了本算法有效。在后面的工作中，可以将其他找 k 个离散的连通同构子图的高性能算法引入，也可将点或边的敏感属性个数从 1 个扩展到多个。

参 考 文 献

[1] 兰丽辉，鞠时光，金华. 社会网络数据发布中的隐私保护研究进展[J]. 小型微型计算机系统，2010，31(12)：2318-2323

[2] Hay M, Miklau G, Jensen D, et al. Resisting structural re-identification in anonymized social networks[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 102-114

[3] Backstrom L, Dwork C, Kleinberg J M. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography[C]// Proceedings of the Sixteenth International World Wide Web Conference, WWW, 2007. Piscataway: IEEE Computer Society, 2007: 181-190

[4] Zhou Bin, Pei Jian, Luk W-S. A brief survey on anonymization techniques for privacy preserving publishing of social network data[C]// Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008. New York: ACM Press, 2008: 12-22

[5] Liu Kun, Terzi E. Towards identity anonymization on graphs[C]// Proceedings of the 2008 ACM SIGMOD international conference on Management of data, 2008. New York: ACM Press, 2008: 93-106

[6] Zhou Bin, Pei Jian. Preserving privacy in social networks against neighborhood attacks[C]// Proceedings of the 24th International Conference on Data Engineering, 2008. Piscataway: IEEE Computer Society, 2008: 506-515

[7] Zou Lei, Chen Lei, Özsu M T. K-automorphism: A general framework for privacy preserving network publication[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 946-957

[8] Cheng J, Fu A W-C, Liu Jia. K-isomorphism: privacy preserving network publication against structural attacks[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data, 2010. New York: ACM Press, 2010: 459-470