社会网络数据的 k-匿名发布

兰丽辉[1,2] 鞠时光[1] 金 华[1]

(江苏大学计算机科学与通信工程学院 镇江 212013)[1] (吉林师范大学计算机学院 四平 136000)[2]

摘 要 由于科学研究和数据共享等需要,应该发布社会网络数据。但直接发布社会网络数据会侵害个体隐私,在发布数据的同时要进行隐私保护。针对将邻域信息作为背景知识的攻击者进行目标节点识别攻击的场景提出了基于 k-匿名发布的隐私保护方案。根据个体的隐私保护要求设立不同的隐私保护级别,以最大程度地共享数据,提高数据的有效性。设计实现了匿名发布的KNP算法,并在数据集上进行了验证,实验结果表明该算法能够有效抵御邻域攻击。

关键词 社会网络,隐私保护,k-匿名,邻域攻击

中图法分类号 TP309 **文献标识码** A

Social Networks Data Publication Based on k-anonymity

LAN Li-hui[1,2] JU Shi-guang[1] JIN Hua[1]

(School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China)[1]

(School of Computer Science, Jilin Normal University, Siping 136000, China)[2]

Abstract Because of scientific researching and data sharing, social networks data should be released. However, individual privacy will be breached if social networks data will be published directly. Therefore, privacy protection should be carried on while releasing social networks data. A privacy protection method based on k-anonymity was proposed. The method is suitable for the scene that the aggressor with background knowledge of neighborhood information wants to re-identify the target node in published social networks. According to the individual privacy protection requirement, the entities set different levels of privacy protection to share data and improve data utility as possible as. Designed and implemented the KNP algorithm to publish data anonymously and carry on experiment on dataset to validate the algorithm. Experimental results show that the algorithm can effectively resist the neighborhood attack.

Keywords Social networks, Privacy protection, k-anonymity, Neighborhood attack

1 引言

社会网络起源于美国著名社会心理学家米尔格伦于20世纪60年代最先提出的"六度理论",也称"小世界现象"。近年来随着社会网络网站数量的不断增加,有关社会网络的研究也越来越多。由于科学研究和数据共享等需要,社会网络数据要进行发布。但是,直接发布社会网络数据将会侵害个体隐私。因此,为了既不泄露隐私又能保证社会网络数据的效用,需要在社会网络数据的发布中进行隐私保护。

数据发布中的隐私保护最早关注的研究领域是关系数据。针对关系数据发布的隐私保护已经取得了一些研究成果,比较典型的隐私保护模型,如 k-anonymity[1-4],l-diversity[3,5]和 t-closeness[3,6]。但是,关系数据在存储、表示和发布形式上有别于社会网络数据。因此,适用于关系数据的隐私保护方案不能直接应用于社会网络。

社会网络属于复杂网络的研究范畴,符合"小世界网络"模型特征,而且节点的度符合幂律分布[7,8]。社会网络的隐私保护策略,其设计取决于隐私信息的类别、攻击者的背景知识和发布数据的效用。目前,已提出一些社会网络数据的隐私保护方法。文献[9]将社会网络描述为不带有标签的简单无向图,提出了基于节点聚类的匿名方法。文献[10]把社会网络抽象为具有多种类型的边、只有一种类型节点的图,提出了基于边聚类的隐私保护方法。文献[11]采用将节点聚类和边聚类相结合的方法实现社会网络的匿名发布。文献[12]研究了边不带有标签的社会网络的 k-度匿名问题。文献[13]研究了把边的权重作为敏感信息进行隐私保护的匿名方法。文献[14]研究了顶点带有非敏感属性的社会网络的匿名发布。文献[15]采用自同构的方法进行社会网络数据发布的隐私保护。文献[16]提出了对二分图进行匿名发布的(k,l)-安全分组方法。文献[17]应用 l-diversity 模型进行社会网络数据发布的隐私保护。除此之外,还有其他一些文献也针对社会网络数据发布的隐私保护提出了解决方案。

但是,已提出的社会网络数据隐私保护方案都基于相同的隐私保护度而设计。在实际的社会网络应用中,由于个体

到稿日期:2010-12-10 返修日期:2011-02-23 本文受国家自然科学基金项目(60773049),江苏省科技创新资金项目(sbc20080655),江苏大学博士创新计划项目(CX10B_006X)资助。

兰丽辉(1976—),女,博士生,讲师,主要研究方向为数据库安全、隐私保护,E-mail:lanlihuicaoyue@163.com;鞠时光(1955—),男,博士,教授,博士生导师,主要研究方向为空间数据库、数据库安全;金 华(1975—),男,博士生,讲师,主要研究方向为数据库安全、隐私保护。

对隐私的认知不同，在对待同一类信息时处理方式也不同，如朋友关系。有人认为发布自己的朋友关系对隐私没有影响，也有人认为发布此类信息将会侵害自己的隐私。而在目前的社会网络数据发布的隐私保护中，都默认参与社会活动的个体具有相同的隐私保护要求，都进行了同等程度的隐私保护。这在某种程度上限制了发布数据的效用，影响了发布数据的质量。为了最大程度保证发布数据的效用，在隐私保护中应根据隐私需求区别对待。

根据个体的隐私保护需求，将隐私保护划分为两个级别：在数据发布中尽量将没有隐私要求的个体信息完整发布；对于涉及隐私的信息进行 k-匿名处理。我们提出的隐私保护方案，即使攻击者在拥有目标对象邻域信息作为背景知识的情况下，也不能以超过 $1/k$ 的概率识别目标节点，而且发布的社会网络数据效用最大。

2 社会网络隐私模型

2.1 社会网络图

社会网络描述社会个体及个体间的交互活动，个体通常指个人。将社会网络抽象为图，图中的节点表示社会个体，边则表示个体间的关系。本文针对节点和边都不带有标签的简单无向图进行研究，将节点的隐私保护划分为两级，用 0、1 表示。“0”表示没有隐私保护要求，“1”表示需要进行隐私保护。要求社会网络中的个体在提交数据时，标记自己的隐私保护级别。

定义 1 社会网络图 $G=(V,E,T)$。其中，节点集 V 代表网络中的社会活动个体；边集 E 表示个体间的关系；T 表示节点的隐私保护级别(由 0、1 构成的序列)，与节点集 V 大小相同。其中，函数 $f:V(G)\rightarrow T(G)$，若节点 $v_i\in V$，则 $f(v_i)=t_i\in T$。

图 1 是表示好友关系的社会网络模型 G_f，图的节点集 $V_f=\{A,B,C,D,E,F,G,H,I,J\}$ 对应的隐私保护级别取值集合 $T_f=\{1,1,0,1,1,1,1,1,0,0\}$。如果将 G_f 直接发布，隐私保护级别为 1 的个体(也称敏感个体)隐私将泄露，要对 G_f 进行隐私保护。最为简单的处理方法是将标识个体的信息隐匿不发布，不改变图的原始结构，发布的匿名图和原始图保持同构。如图 2 所示，用 X 代替了节点的标识信息。

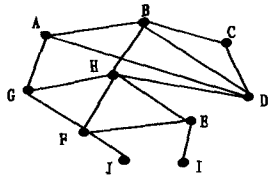

图 1 原始社会网络 G_f

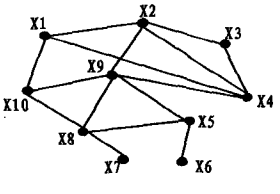

图 2 社会网络 G_f 的简单匿名

2.2 邻域攻击

在简单匿名发布中，如果攻击者不具有任何背景知识，则属于安全发布。但是，在实际的应用场景中，攻击者的背景知识很复杂，通过其他渠道很可能获取关于目标节点的相关信息。因此，不具有任何背景知识的攻击是一种理想情况，实际中不可能存在。简单匿名发布使数据的效用最大，但是泄漏隐私的风险也最大。在攻击者具有一定背景知识的情况下，可能从发布的匿名图中以高置信度识别目标节点。

本文研究的社会网络数据发布场景是：攻击者拥有目标个体的 1-邻域信息作为背景知识，攻击者的目标是识别目标个体在发布社会网络图中对应的节点。如不特殊说明，文中所指邻域即为 1-邻域。

定义 2 社会网络图 $G=(V,E,T)$ 中任意敏感节点 v，如果攻击者获知 v 的邻接点数量和邻接点之间的关系，以此作为背景知识在发布的社会网络图 G_p 中识别目标节点 v 的攻击，称为 1-邻域攻击。

在图 2 中，如果攻击者获知敏感个体 A 的邻域信息(有 3 个好友，且 3 个好友中有 2 个互为好友)，就能够进行唯一匹配识别，判断 $X1$ 就是目标节点。识别出 A 所处位置，攻击者会在该网络中挖掘出更多关于 A 个体的信息，致使其隐私泄露。为了确保敏感个体的隐私安全，在简单匿名的基础上需要采取其他的隐私保护策略，以保证攻击者在拥有一定背景知识的情况下，也不会以较高概率识别目标节点。

3 k-匿名发布

k-匿名模型在关系数据的发布中广泛应用，是经典的隐私保护模型。为了保护发布的社会网络中敏感个体的隐私，要求每个敏感个体都隐藏在 k 个不可区分的网络节点中，使攻击者不能以高于 $1/k$ 的概率识别目标个体。

在本文设定的应用场景中，攻击者的背景知识是目标对象的邻域。通过在发布的社会网络图中构建与目标对象邻域相同的 $k-1$ 个邻域，即实现社会网络的 k-匿名发布，可以保证敏感个体的隐私安全。

3.1 邻域提取

为了实现抵御邻域攻击的 k-匿名发布，首先要获取节点的邻域信息。对于数据发布者而言，节点的邻域信息很容易获取，而较为复杂的是如何在发布的图中寻找与其相匹配的 $k-1$ 个邻域。我们借鉴文献[14]的邻域处理方法，将节点的邻域表示为组件。

定义 3 已知社会网络图 $G=(V,E,T)$，节点 $v\in V$ 的邻域记为 $Neighbor_G(v)$，将 $Neighbor_G(v)$ 中的每个极大连通子图称为 v 的邻域组件。

图 3 中给出的是图 1 中节点 H 和 F 的邻域(图中的虚线是为了突出邻域组件，实际上就是连接节点的边)。$Neighbor_G(H)=\{H_1,H_2\}$，$Neighbor_G(F)=\{F_1,F_2\}$。

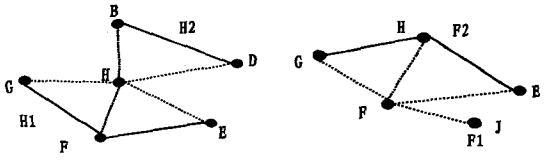

图 3 节点 H 和 F 的邻域

邻域提取算法(Neighborhood Extract Algorithm，NEA)如下：

已知：$G=(V,E,T)$—社会网络图

返回：提取邻域的节点集 W 及其对应邻域组件集合 C_W，任意节点 $w\in W$，$f(w)=1$，$W\subset V$

(1)初始 $W=\phi$，$C_W=\phi$；

(2)对 G 进行深度优先搜索遍历，若 $v_i\in V$ 是敏感节点，提取其邻域信息，将 $Neighbor_G(v_i)\rightarrow C_W$，$v_i\rightarrow W$；

(3)将 W 节点集和组件集 C_W 按邻域组件大小进行排序，邻域组件的比较按文献[14]提出的标准进行；如果节点 v

排在节点 u 之前，则需满足如下条件：$|V(Neighbor_G(u))|<|V(Neighbor_G(v))|$ 或 $|V(Neighbor_G(u))|=|V(Neighbor_G(v))| \wedge |E(Neighbor_G(u))|<|E(Neighbor_G(v))|$；

(4)将排好序的 W 和 C_W 返回。

3.2 邻域匹配

攻击者可以通过获得的邻域信息在发布的社会网络图中构建与目标节点邻域相匹配的候选集。候选集中节点数量越多，目标对象被识别的概率就越低。我们的目标就是对于任意一个敏感节点经隐私保护处理发布后，其候选集中至少有 $k-1$ 个节点，这样攻击者识别目标节点的最大概率不会超过 $1/k$。

定义 4 已知社会网络图 G 的 W 和 C_W 集合，若 $v\in W$，在 C_W 中与 v 的邻域结构相同或最为相似的 $k-1$ 个邻域对应的在 W 中的 $k-1$ 个节点构成的集合称为节点 v 的候选集，记作 $Cand(v)=\{u_1,u_2,\cdots,u_{k-1}\}$，$u_i\in W$。

在 k-匿名发布中，我们关心的是敏感节点，只需要对 W 中的每个节点构建候选集即可。为了抵御邻域攻击，要求每个候选集中的节点具有相同的邻域结构。根据图论的知识，可以获知在社会网络图中存在着同构子图，但是原始图中同构子图的数量不足以用来阻止攻击者的重识别攻击[18-20]，多数节点的邻域同构还需要通过插入边和引入节点的方式实现。

在邻域匹配中不可避免地要进行子图的同构比较，而判断两个子图是否同构是 NP-难的。本文采用与文献[14]相似的方法，按深度优先搜索匹配两个邻域组件。优先选取度相同且取值最大的节点对，按深度优先搜索的策略逐一匹配。在匹配的过程中，计算匿名成本。如果匿名成本为 0，说明两个节点的邻域同构，选取匿名成本最低的前 $k-1$ 个节点作为候选集。

在图 1 所示的社会网络图中就存在着邻域同构的节点，节点{A,E,G}拥有相同的邻域结构，{B,D}也拥有相同的邻域结构，如图 4 和图 5 所示。

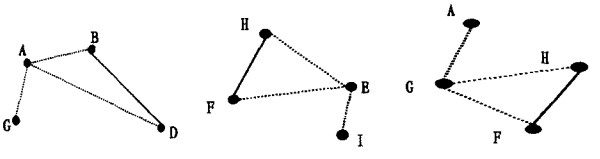

图 4 A,E,G 的同构邻域

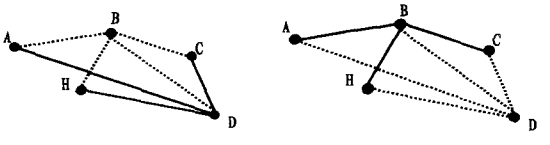

图 5 B,D 的同构邻域

若对图 1 进行 2-匿名发布，即 $k=2$，为了保证数据效用的最大化，只需找到原始图中满足同构的节点，然后对不满足 2-匿名的节点进行邻域同构操作。图 1 中{A,E,G}和{B,D}在原始图中就满足 2-匿名的发布要求，无需处理。

图 1 中，除去满足邻域自同构的节点，还有{C,F,H,I,J}5 个节点。其中，F、H 是敏感节点，需要进行邻域同构操作。通过比较图 3 中 F 和 H 的邻域组件，可知要想实现邻域同构，需要在 F 的邻域中引入一个节点，并且在其和节点 J 之间插入一条边。可以将节点 I 引入 F 的邻域，在 I、J 之间插入边。

经过上述操作，图 1 中的敏感节点都实现了 2-匿名，而非敏感节点 C 未做处理，其邻域结构完整发布。为了实现邻域同构引入的 I 是非敏感节点，插入的边也在非敏感节点间实现。图 1 的 2-匿名发布如图 6 所示，匿名成本为 2(匿名成本计算方式见 3.3 节)。如果采用文献[14]中的算法，因为所有节点都是敏感节点，为实现邻域同构需要递归的执行算法，不断更新节点集的信息，使得发布的匿名成本高，而且算法效率低。

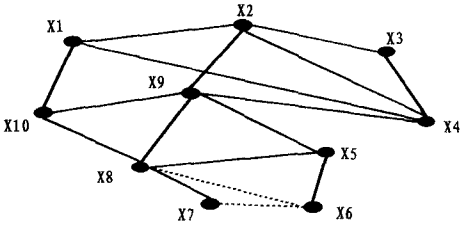

图 6 社会网络 G_J 的 2-匿名发布

在向邻域中引入节点时，应符合的条件是选取度最小的并不在邻域组件中的非敏感节点；若不存在非敏感点，则选取度最小、不在邻域中的未匿名的敏感节点；若不存在未匿名的敏感节点，则选取已匿名的度最小的节点，但同时要将该节点的候选集都标记为未匿名，更新节点集。在邻域中插入边时，也优先选取在邻域中的非敏感节点进行。这样能够保证之前已经实现的同构操作不会受影响，提高匿名效率。

3.3 匿名成本

隐私保护和数据效用是一对矛盾，通常，隐私保护质量越高，数据效用越差，反之亦然。本文通过发布图中插入边的数量和为了实现邻域同构引入邻域中的节点数量来度量匿名发布成本。在发布的社会网络图中，节点的数量并没有改变，在邻域同构中引入节点就等同于插入边。所以，最终匿名发布的成本通过插入边的数量来度量，插入边的数量越大，数据缺损越大。匿名成本按如下公式计算：

$$Cost(G)=\sum_{i=1}^{|W|}Cost(Cand(v_i)) \tag{1}$$

$$Cost(Cand(v_i))=\sum_{j=1,u_j\in Cand(v_i)}^{|Cand(v_i)|}Cost(v_i,u_j) \tag{2}$$

$$Cost(v_i,u_j)=|E(Neighbor_{Gp}(v_i))|+|E(Neighbor_{Gp}(u_j))|-|E(Neighbor_G(v_i))|-|E(Neighbor_G(u_j))| \tag{3}$$

其中，式(3)计算的是两个节点邻域同构中增加边的数量。

3.4 KNP 算法

定义 5 社会网络图 $G=(V,E,T)$ 的匿名发布 $G_p=(V_p,E_p)$，若 $v\in V$ 且 $v\in V_p$，$f(v)=1$，$|Cand(v)|\geqslant k-1$，则称 G_p 是 G 的 k-匿名发布。

我们设计了 KNP(K-anonymity Against Neighborhood Attack Publication)算法实现社会网络的 k-匿名发布。KNP 算法如下：

输入：原始的社会网络图 G 和参数 k

输出：G 的匿名发布 G_p

步骤：

(1)初始化图 G，隐匿节点标识信息；

(2)调用 KNM 算法提取邻域，得到 W 和 C_W；将 W 中的节点都标记为未匿名，$Marked(w_i)=\text{False}$，$w_i\in W$；

(3)对 W 中的节点进行邻域自同构匹配,将自同构匹配的节点标记为匿名节点,将其放入候选集,并标记为自同构节点;

(4)更新经自同构匹配后的节点集 W,若对于节点 $w\in W$,且 $Cand(w)\geqslant k$,则满足匿名发布,将候选集中节点标记为已匿名;若 $Cand(w)<k$,则在 W 中寻求与其邻域匹配成本最低的 $k-|Cand(w)|$ 个未匿名节点加入候选集,更新 W;

(5)通过引入节点和插入边对候选集的节点进行同构操作,按 3.2 节提出的标准引入节点、插入边;若 W 中未匿名节点的数量小于 k,则通过引入非敏感节点实现 k-匿名;若 W 中没有未匿名的节点,则匿名结束;

(6)用同构后的邻域取代原始图中节点的邻域,得到图 G 的 k-匿名发布 G_p;

(7)返回 G_p。

4 实验结果及分析

4.1 实验环境

我们采用 Pajek 软件生成社会网络图 FriendNet(友谊网),测试 KNP 算法。FriendNet 包含节点的数量为 500,节点的平均度为 5,实验环境采用 Windows XP Professional 操作系统,CPU 2.70GHz,内存 2GB,编程语言为 C++,运行平台为 Visual C++6.0。

4.2 实验方案

我们设计了 3 种实验方案。方案 1:针对同一数据集,按两种方法进行发布:一种是对社会网络中的所有个体采用相同的隐私保护策略;另一种按照 3∶7 的比例对个体的隐私保护级别进行 0、1 的设置。为了区别,称实现第一种方法的算法为 SKNP,第二种方法采用 KNP 算法。SKNP 和 KNP 采用相同的邻域提取和匹配方案,区别在于 SKNP 算法对所有节点都同等对待。我们对采用两种方法发布的社会网络中的节点信息完整性和算法的执行效率进行测试。实验结果如图 7 所示。结果表明,采用 KNP 算法的执行效率和节点的信息的完整率要高于 SKNP。

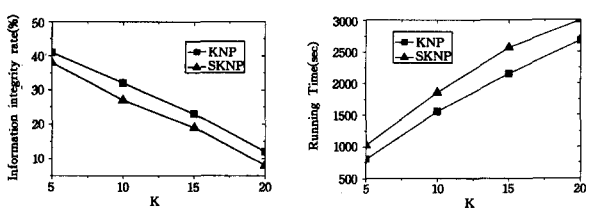

图 7 方案 1 实验结果

方案 2:根据 k 的不同取值,测试社会网络数据经匿名发布后发布网络图中插入边的数量,即匿名成本。实验结果如图 8 所示。实验表明,随着 k 值的不断增大,发布网络中插入边的数量越多,为实现邻域同构的代价越大。

方案 3:通过查询发布社会网络图,测试查询结果的错误比率。随着 k 取值的不同,随机选取 100 个节点对,测试节点间的平均最短路径。平均最短路径的查询错误率按文献[14]提出的公式计算:错误率 $r=\frac{d-d'}{d}$,d 和 d' 分别为原始网络和发布网络中计算出的平均最短路径长度。实验结果如图 9 所示。实验表明,对发布的社会网络图进行聚集查询,有着较高的准确度。随着 k 值的不断增加,错误率并没有大幅上升。

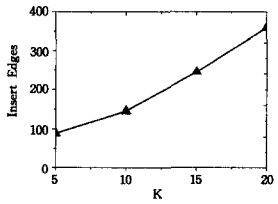

图 8 方案 2 实验结果

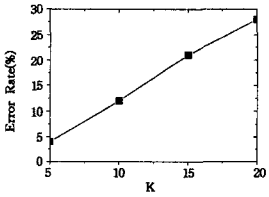

图 9 方案 3 实验结果

4.3 结果分析

经过学者们对实际社会网络的分析,可知社会网络中节点间的连接具有对称性且节点的度符合幂律分布;社会网络是由度最高的 10%的节点支撑,路径的长度很短[7]。度高的节点在网络中的地位相对比较重要,通常要进行隐私保护,即所谓的敏感个体。同时,网络中存在着相当数量度较低的节点,挖掘它们的隐私意义不大,因此这些节点的隐私要求较低,即所谓的非敏感节点。所以,本文中提出的社会网络数据发布场景实际可用。

通过对上述 3 种方案的实验测试表明,由于社会网络中个体对隐私认知程度的不同,导致社会网络的发布中存在一定数量无需进行隐私处理的节点,充分利用这类节点帮助敏感节点实现隐私保护,可以降低匿名成本,提高算法效率,同时使得发布的社会网络数据效用最大。

结束语 本文研究了社会网络数据的 k-匿名发布问题。根据社会活动个体对隐私保护的不同要求,设立了两个保护级别。采用 k-匿名发布抵御拥有邻域信息作为背景知识的攻击者进行节点识别攻击,设计实现了 KNP 算法。我们采用的隐私保护方法最大程度地使非敏感节点的信息完整发布,同时通过邻域的同构保护了敏感节点的隐私信息,提高了数据效用。而现有的社会网络隐私保护方法对所有社会个体都采用了相同的隐私保护策略,降低了发布的社会网络数据效用。

在本文中,我们主要针对 1-邻域的结构攻击进行了研究,实际上算法也可以抵御其他类型的邻域攻击。在实际应用中,攻击者获得目标节点的邻域信息通常都不完整。因此,我们提出的 k-匿名发布能够有效抵御邻域攻击。今后,我们将针对实际的社会网络数据集进行实验测试,并将继续对不同应用场景下社会网络数据的发布进行隐私保护研究。

参 考 文 献

[1] Sweeney L. k-anonymity: A model for protecting privacy[J]. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570

[2] 韩建民,岑婷婷,虞慧群. 数据表 k-匿名化的微聚集算法研究[J]. 电子学报, 2008, 36(11): 2021-2029

[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[J]. 计算机学报, 2009, 32(5): 847-860

[4] 杨晓春,刘向宇,王斌,等. 支持多约束的 K-匿名化方法[J]. 软件学报, 2006, 17(5): 1222-1231

[5] 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): 24-35

[6] Li N, Li T. t-closeness: Privacy beyond k-anonymity and l-diversity[C]// Proc. of the 23rd International Conference on Data Engineering. Istanbul: IEEE, 2007: 106-115

[7] Mislove A, Marcon M, Krishna P, et al. Measurement and Analysis of Online Social Networks[C]// Proc. of IMC'07. New York: ACM, 2007: 29-41

[8] Albert R, Jeong H, Barabasi A L. Error and attack tolerance of complex networks[J]. Nature, 2000, 406: 378-382

[9] Hay M, Miklau G, Jensen D, et al. Anonymizing social networks [R]. 07-19. University of Massachusetts Amherst, 2007

[10] Zheleva E, Getoor L. Preserving the privacy of sensitive relationships in graph data[C]// Proc. of the 1st ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD. Berlin: ACM, 2007: 153-171

[11] Campan A, Truta T M. A clustering approach for data and structural anonymity in social networks[C]// Proc. of the 2nd ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD. Las Vegas: ACM, 2008: 93-104

[12] Liu L, Wang J, Liu J, et al. Privacy preserving in social networks against sensitive edge disclosure[R]. CMIDA-HiPSCCS 006-08. Department of Computer Science, University of Kentucky, 2008

[13] Liu K, Terzi E. Towards identity anonymization on graphs[C]// Proc. of the 2008 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008: 93-106

[14] Zhou B, Pei J. Preserving privacy in social networks against neighborhood attacks[C]// Proc. of the 24th IEEE International Conference on Data Engineering. Washington D C: IEEE, 2008: 506-515

[15] Zou Lei, Chen Lei, Özsu M T. K-Automorphism: General Framework for Privacy Reserving Network Publication[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 946-957

[16] Cormode G, Srivastava D, Yu T, et al. Anonymizing bipartite graph data using safe groupings[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 833-844

[17] Panda G K, Mitra A, Prasad A, et al. Applying l-Diversity in anonymizing collaborative social network[J]. International Journal of Computer Science and Information Security, 2010, 8(2): 324-329

[18] 谭建荣，岳小莉，陆国栋. 图形相似的基本原理、方法及其在结构模式识别中的应用[J]. 计算机学报，2002，25(9)：59-967

[19] He P R, Zhang W J, Li Q. Some further development on the eigen system approach for graph isomorphism detection [J]. Journal of Franklin Institute, 2005, 342(6): 657- 673

[20] Hay M, Miklau G, Jensen D, et al. Resisting Structural Reidentification in Anonymized Social Networks [C] // Proc. of the VLDB Endowment. New Zealand, Auckland: VLDB Endowment, 2008: 102-114

(上接第 155 页)

据序列，是当前一项重要且有现实意义的课题，它能够帮助用户发现数据中隐藏着的丰富知识，揭示数据变化规律及预测序列的发展趋势。本文提出了一种基于函数的时间序列分段线性表示方法，与 PAA 和 RPAA 方法相比，本文的方法不仅考虑到了时间序列的时间特性，而且能够实时划分时间序列，结果表明 FPAA 方法能够比较准确地表示原始时间序列的特性，进行有效的在线查询。但对于不同的数据，如何找到更加合适的函数影响因子来表现原始时间序列的时间特性，是一项有待研究的课题。

参 考 文 献

[1] Fu T C, Chung F L, Ng V. et al. Pattern Discovery from Stock Time Series Using Self-Organizing Maps[C]// Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Minning. San Francisco: ACM Press, 2001: 23-27

[2] Gavrilov M, Anguelov D, Indyk P, et al. Mining the Stock Market: Which Measure is Best? [C]// Proceedings of the 6th ACM International Conference on Knowledge Discovery and Data Mining. Boston: ACM Press, 2000: 487-496

[3] Mantegna R N. Hierarchical Structure in Financial Markets[J]. European Physical Journal, 1999(11): 193-197

[4] Bar-joseph Z, Gerber G, Gifford D, et al. A new Approach to Analyzing Gene Expression Time Series Data[C]// Proceedings of the 6th Annual International Conference on Research in Computational Molecular Biology. New York, USA: ACM Press, 2002: 39-48

[5] Koski A, Juhola M, Meriste M. Syntactic Recognition of ECG Signals by Attributes Finite Automata[J]. Pattern Recognition, 1995, 28(12): 1927-1940

[6] Harm S K, Reichenbach S, Goddard S E, et al. Data Minning in a Geospatial Decision Support System for Drought Risk Management[C]// Proceeding of the 1st National Conference in Digital Government. Los Angeles, CA, 2002: 9-16

[7] Keogh E, Kasetty S. A Survey and Empirical Demonstration[C]// Proceeding of the 8th ACM SIGKDD International Confe-rence on Knowledge Discovery and Data Mining. Ganada: Edmonton, Alberta, 2002: 102-111

[8] Keogh E, Chu S, Hart D, et al. An online algorithm for segmenting time series[C]// Proceedings of IEEE International Conference on Data Mining. Los Alamitos: IEEE Computer Society Press, 2001: 289-296

[9] Keogh E, Chakrabarti K, Pazzani M J, et al. Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases[J]. Knowledge and Information Systems, 2001, 3(3): 263-268

[10] Yi B K, Faloustsos C. Fast Time Sequence Indexing for Arbitrary Lp Norms[C]// Proceeding of the 26th International Conference on Very Large Databases. San Fracisco: Morgan Kaufmann Publishers Inc, 2000: 385-394

[11] 王元珍，李俊奎，曹忠升. 一种基于时间特性的时间序列建模表示[J]. 计算机科学，2007，34(3)：83-86

[12] Keogh E, Pazzani M. A Simple dimensionality reduction technique for fast similarity search in large time series databases[C]// Proceeding of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining. Kyoto, Japan: Berlin, 2000: 122-133