软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(9):2323−2333 [doi: 10.13328/j.cnki.jos.005178]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

图数据发布隐私保护的聚类匿名方法*

姜火文[1,2,3], 占清华[4], 刘文娟[1,3], 马海英[5]

[1](同济大学 计算机科学与技术系,上海 200092)

[2](江西科技师范大学 数学与计算机科学学院,江西 南昌 330038)

[3](嵌入式系统与服务计算教育部重点实验室(同济大学),上海 200092)

[4](江西科技学院 信息工程学院,江西 南昌 330098)

[5](南通大学 计算机科学与技术学院,江苏 南通 226019)

通讯作者: 姜火文, E-mail: hwjiang@tongji.edu.cn

摘 要: 社交网络中积累的海量信息构成一类图大数据,为防范隐私泄露,一般在发布此类数据时需要做匿名化处理.针对现有匿名方案难以防范同时以结构和属性信息为背景知识的攻击的不足,研究一种基于节点连接结构和属性值的属性图聚类匿名化方法,利用属性图表示社交网络数据,综合根据节点间的结构和属性相似度,将图中所有节点聚类成一些包含节点个数不小于 k 的超点,特别针对各超点进行匿名化处理.该方法中,超点的子图隐匿和属性概化可以分别防范一切基于结构和属性背景知识的识别攻击.另外,聚类过程平衡了节点间的连接紧密性和属性值相近性,有利于减小结构和属性的总体信息损失值,较好地维持数据的可用性.实验结果表明了该方法在实现算法功能和减少信息损失方面的有效性.

关键词: 社交网络;隐私保护;聚类匿名;属性图;数据发布

中图法分类号: TP309

中文引用格式: 姜火文,占清华,刘文娟,马海英.图数据发布隐私保护的聚类匿名方法.软件学报,2017,28(9):2323−2333. http://www.jos.org.cn/1000-9825/5178.htm

英文引用格式: Jiang HW, Zhan QH, Liu WJ, Ma HY. Clustering-anonymity approach for privacy preservation of graph data-publishing. Ruan Jian Xue Bao/Journal of Software, 2017,28(9):2323−2333 (in Chinese). http://www.jos.org.cn/1000-9825/5178.htm

Clustering-Anonymity Approach for Privacy Preservation of Graph Data-Publishing

JIANG Huo-Wen[1,2,3], ZHAN Qing-Hua[4], LIU Wen-Juan[1,3], MA Hai-Ying[5]

[1](Department of Computer Science and Technology, Tongji University, Shanghai 200092, China)

[2](School of Mathematics and Computer Science, Jiangxi Science and Technology Normal University, Nanchang 330038, China)

[3](Embedded System and Service Computing Key Laboratory of Ministry of Education (Tongji University), Shanghai 200092, China)

[4](College of Information Engineering, Jiangxi University of Technology, Nanchang 330098, China)

[5](College of Computer Science and Technology, Nantong University, Nantong 226019, China)

Abstract: A huge amount of information in social network has accumulated into a kind of big graph data. Generally, to prevent privacy leakage, the data to be published need to be anonymized. Most of the existing anonymization scheme cannot prevent such attacks by background knowledge of both structure and attribute information among nodes. To address the issue, this investigation proposes a clustering-anonymization method for attribute-graph based on link edges and attributes value among nodes. Firstly, the data in the social

* 基金项目: 国家自然科学基金(61762044, 71561013, 61402244); 江西科技师范大学重点科研项目(2016XJZD002)

Foundation item: National Natural Science Foundation of China (61762044, 71561013, 61402244); Key Research Project of Jiangxi Science & Technology Normal University (2016XJZD002)

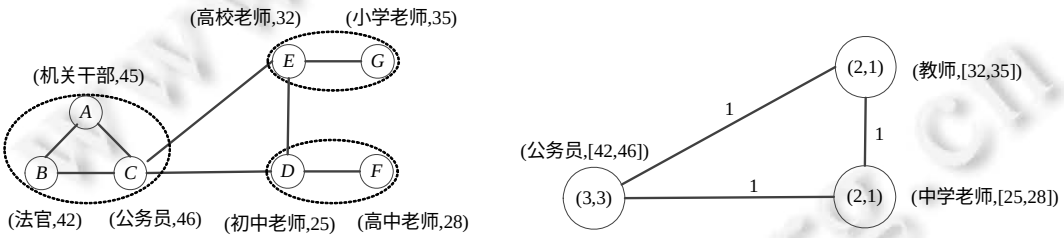
收稿时间: 2016-06-28; 修改时间: 2016-09-04, 2016-11-10; 采用时间: 2017-01-06; jos 在线出版时间: 2017-02-20

CNKI 网络优先出版: 2017-02-20 14:05:20, http://www.cnki.net/kcms/detail/11.2560.TP.20170220.1405.015.html

network is represented by attribute graph. Then all the nodes of this attribute graph are clustered into certain super-nodes according to structural and attribute similarity between two nodes, each of which contains no less than k nodes. Finally, all the super-nodes are anonymized. In this method, the structure masking and attribute generalization for every super-nodes can respectively prevent all the recognition attacks by background knowledge of goals' linkages and attribute information. In addition, it balances the closeness of links among nodes and proximity of attributes value during clustering, therefore can reduce the total loss of information triggered by masking and generalization to maintain the availability of these graph data. Experiment results also demonstrate the approach achieves great algorithm performance and reduces information loss remarkably.

Key words: social network; privacy preservation; clustering-anonymity; attribute graph; data-publishing

随着各类网络社交平台的兴起和应用普及,社交网络中积淀下来海量的个人及社会关系信息,形成了一类具有重要商业价值的图大数据.然而,直接发布这些数据很容易导致个体隐私泄露,从而给人们生活带来困扰,甚至造成重大经济利益损失.因此,针对社交网络数据发布的隐私保护问题日益受到人们关注,其相关研究也已成为一个热点.为了保护隐私,一种简单的方法是直接将用户身份信息隐匿,而保留其他所有信息,这可以一定程度地防止个体隐私暴露.因为即使其中有敏感信息被公开,但由于对应的所有者身份不明确,不会给用户带来直接的困扰.然而面对恶意的隐私盗取,这种简单的匿名方法就无济于事.因为攻击者可以基于与目标相关的背景知识,以较高的概率甚至准确地推测出用户身份,导致隐私泄露[1].例如在图 1(a)中,假定边代表朋友关系,若攻击者获悉"Evi 有最多的朋友数"或"Evi 有不少于 4 个朋友"这个背景知识,则 *C* 节点是 Evi 的身份将被识别.另外在社交网络中,根据用户的某些属性信息,结合其他渠道获得的背景知识,也可能推测出用户身份和其他敏感属性[2,3].



(a) 匿名前的属性图 (b) 匿名后的 2-匿名图

Fig.1 Social network graph and the corresponding anonymity graph

图 1 社交网络模型图及其匿名图

子图 *k*-匿名是一种有效的隐私保护方案,其思想是:当攻击者将目标所在的特定子图结构作为背景知识进行隐私攻击时,构造至少 *k* 个相同结构的子图作为候选,使目标子图导致隐私泄露的概率小于 1/*k*,而与目标相关的图结构信息,如节点的度、邻域图等均可作为攻击者的背景知识[1].可见,子图 *k*-匿名是基于结构变换的匿名方案,也是当前最为典型的匿名方案[4].实际中,攻击者具备的社交网络背景知识可能多种多样,但主要可分为结构信息和属性信息.子图 *k*-匿名能够防范以图结构特性为背景知识的隐私攻击,却不能抵御以节点属性信息为背景知识的隐私攻击.为创新现有匿名方法,使其能够同时防范基于结构和属性信息的隐私攻击,即,能够抵御基于几乎一切背景信息的隐私攻击,本文研究一种基于节点结构和属性相似度的属性图聚类匿名方法.本文的主要贡献为:

(1) 建立问题的属性图模型,并在属性图聚类的基础上完成数据匿名处理;

(2) 分别量化节点间的结构和属性相似性,并藉以计算节点间及节点与超点间的综合相似度,聚类过程根据此综合相似度进行节点聚类;

(3) 对所有属性,不区分准标识属性和敏感属性,一并进行概化处理;但是概化过程区分属性数据为数值型和类别型两种,分别采用不同的概化方法.

1 相关工作

目前,社交网络匿名方法基本分为两大类,即,基于结构变换的匿名方案和基于超级节点的匿名方案[4].基于结构变换的匿名方案也就是子图 k-匿名,为最典型的社交网隐私保护方案,其中又以 k-度匿名和 k-同构子图匿名较为常见.例如:为抵御以节点度数为背景知识的隐私攻击,Liu 等人[5]设计了基于分治法和最大邻差的两种度序列划分算法,并借以构建了 k-度匿名图;为抵御基于子图结构的攻击,Zou 等人[6]提出一种匿名化算法,在网络图中构造了至少 k–1 个与目标子图同构的子图.从度匿名到同构匿名,一般只针对基于度或子图信息的身份识别攻击.为了同时防止根据结构信息推测用户身份和根据多个节点共享同一敏感属性推测用户隐私,Yuan 等人[7]借鉴关系数据保护的(k,l)匿名模型,通过在原图中适当添加噪音节点的方法,构造一种"k-度-l-多样性"匿名模型.针对匿名方法可能对图结构特性的破坏,Masoumzadeh 等人[8]提出了两种启发式方法减少匿名对边的扰乱,以维持图的结构特性,但算法效率和效果还有待完善和理论证实.基于超级节点的匿名方案最早由 Zheleva 等人[9]提出,其后,Campan 等人[10]和 Tassa 等人[11]都对此类匿名方案进行了完善工作[4].Campan 等人[10]提出了一种贪心算法 SaNGreeA,实现节点聚类匿名,同时,首次针对匿名过程产生的结构信息损失给出了一种定量计算方法;Tassa 等人[11]提出了基于有序聚类的匿名方法,实现集中式网络环境下的数据隐私保护,同时,首次针对节点间不完全可见的分布式图数据的隐私保护设计了有序聚类匿名算法.一般匿名方法都是针对无权图,Skarkala 等人[12]针对带权图数据的隐私保护,采用基于超点和超边的匿名机制,通过只发布超边权重隐藏了原来各连接边的具体权重,通过超点结构将用户身份泄露概率降至 $1/k$ 以内,但方法缺少对属性信息的考虑.针对社交网络属性隐私匿名算法中存在的缺乏合理模型、较少考虑社交结构和非敏感属性对敏感属性分布的影响等不足,Fu 等人[13]提出了一种基于节点分割的方法,通过分割节点的属性连接和社交连接提高节点的匿名性,降低了用户属性隐私泄露的风险,但算法的匿名有效性比较受局部区域内属性相关性的影响.近年来,差分隐私保护方法也越来越多地被研究应用于社会网络隐私保护中[14,15].不久前,作者所在的研究小组提出了一种利用遗传算法实现的图聚类匿名隐私保护方法[16],能够实现本文达成的隐私保护目标,但算法运行时间更长.为提高算法中图节点聚类的时间效率,本文进一步研究提出取代遗传算法,而直接根据图节点的结构和属性相似性进行属性图聚类,并在此基础上完成超点匿名.

从形式上说,本文方法也属于基于超级节点的匿名方案,但具有自身特色.例如,与其中两种典型的匿名方法 SaNGreeA[10]和 MASN[17]相比,

(1) SaNGreeA 利用贪心算法,在每一步选取节点聚类时均依据属性信息损失最小原则进行;MASN 在节点聚类过程中以总体属性信息损失最小为目标,但并不保证当前每一步聚类操作使信息损失最小;本文方法也利用到贪心思想,但每一步聚类是根据当前综合相似度最大原则进行,不过,在减小信息损失方面,与 SaNGreeA 有异曲同工之妙;

(2) SaNGreeA 和 MASN 在属性概化时,没有区分属性值为数值型和类别型的不同,采用了统一的概化方法;本文则区分属性值类属分类进行属性概化;

(3) SaNGreeA 和 MASN 均将属性区分为准标识属性(quasi-identifier attribute(s))和敏感属性(sensitive attribute(s)),概化对象为准标识属性;本文则针对所有属性进行概化;

(4) SaNGreeA 的聚类结果不能保证每个聚类中敏感属性的分布符合 l-diversity 特性,可能影响其隐私保护性;MASN 能够满足(k,l)匿名性,匿名保护程度有所增加;而本文方法将敏感属性一并概化,更加增进了其隐私保护强度.

总结本文方法的特色是:(1) 聚类过程中,综合依据连接关系和属性值的相近性进行节点聚类;(2) 属性概化时,区分数值型和类别型两类属性分类进行概化操作;(3) 对节点的属性,不区分准标识属性和敏感属性,一并经属性概化后发布.本文特色工作的意义主要体现在:(1) 综合结构和属性相近性聚类节点,有利于减小总体信息损失,并改善聚类质量;(2) 属性分类概化有利于降低数值型属性的概化程度,从而减少其信息损失;(3) 通过全部属性的概化,既防止了以属性为背景知识的攻击,又增强了对属性隐私本身的保护.

2 问题模型与定义

社交网络中的数据主要有 3 类:身份数据、社交关系数据和属性数据.身份数据是用户的真实身份信息,通常直接被网络服务商隐匿;社交关系数据是反映用户间某种社会关联关系的信息,很多情况下用户间这种社交关系不愿被公开,即,被视为隐私,需要加以保护;属性数据是反映用户某些特征的个性化信息,可能直接或间接暴露用户隐私,故也有必要适当加以匿名化处理.社交网络中,用户间的复杂联系表现为一种图关系结构,考虑同等关注属性数据的隐私保护,我们定义属性图如下.

定义 **1**(属性图). 对包含了用户属性和用户间相互关系的社交网络,以一个带属性标签的无向无权图 G 描述,形式化表示为一个 3 元组,即:$G=(V,E,A)$,称 G 为属性图.

在属性图 $G=(V,E,A)$中,$V=\{v_1,v_2,\ldots,v_n\}$为图中节点集,每个节点表示社交网络中的一个个体;$E=\{(v_i,v_j)|1\leq i,j\leq n,i\neq j\}$为边集,每条边表示对应两个个体间存在的社交关系;$A=\{A_1,A_2,\ldots,A_n\}$为所有节点的属性值集合,其中,$A_i=(a_{i1},a_{i2},\ldots,a_{is})$为节点 $i(i=1,2,\ldots,n)$的 s 维属性向量值.例如,图 1(a)即表示一个属性图,其节点属性为一个形如(职业,年龄)的二维属性.

定义 **2**(邻接点). 在属性图 $G=(V,E,A)$中,对任意顶点 v_i,若$(v_i,v_j)\in E(G)$,称 v_j 为 v_i 的邻接点.

在属性图中,一条边连接的两个节点互为邻接点.节点 v_i 的所有邻接点的集合记为 $\tau(v_i)$.

定义 **3**(属性图聚类). 根据节点间的相似程度,利用聚类方法将属性图中的所有节点划分成一些聚类(簇),使每个聚类中节点个数大于等于 k 而小于 $2k$,这个过程成为属性图聚类.

属性图聚类可以形式化地表示为属性图 G 上的一种映射关系$\varphi:(v_1,v_2,\ldots,v_n)\rightarrow(V_1,V_2,\ldots,V_m)$,$\varphi$满足以下条件:

(1) $\bigcup_{i=1}^{m}V_i=V(G)$ 且 $k\leq|V_i|<2k$,m 为聚类的数目,$|V_i|$表示 V_i 中的节点数目;

(2) $\forall i,j\in\{1,2,\ldots,m\},i\neq j$,则 $V_i\cap V_j=\varnothing$.

定义 **4**(***k***-匿名图). 对属性图 $G=(V,E,A)$进行属性图聚类操作后,将由每一簇构成的子图用一个$(|V_i|,|E_i|)$的超点表示,超点内各节点的属性向量值概化成一个属性向量值.2 个超点间,若原属性图中没有任何边相连,则无超边相连;若原有 t 条边相连,则仅用一条带权值 t 的超边取代.由此得到的“浓缩”了的图,称为 k-匿名图.其中,$|V_i|$和$|E_i|$分别表示超点 V_i 内的节点数和边数.

本文的研究工作即可归纳为将属性图转化成对应的 k-匿名图对外发布.例如,图 1(b)即是对图 1(a)聚类匿名化处理后的一个 2-匿名图.在 k-匿名图中,由于各超点内部的节点及其连接关系被隐匿,即使该超点被定位,其内部任一节点被识别的概率仍然不超过 $1/k$,故具有很好的隐私保护效果.同时,对于大型社会网络而言,在 k 值不大时(一般取值 5 左右即可保证隐私强度),其 k-匿名图只是相当于在全局范围内做了一些微观浓缩,不会改变原属性图的宏观结构,也基本能够保留其中蕴含的主要信息.故,k-匿名图数据仍然能够维持较高的可用性.

属性图经过聚类匿名后,每个超点只保留一个属性向量值,为此,需要将各超点内部的所有节点属性向量值统一概化成一个属性向量值.

定义 **5**(属性概化). 任取聚类 $V_i(i=1,2,\ldots,m)$,对其中的各节点 v_j 在任意属性上的值 $a_{jt}(t=1,2,\ldots,s)$,选取包含各 a_{jt} 的某一个范围更广的域值作为超点 V_i 在属性 t 上的取值,这个过程称为属性概化.

社交网络中的属性数据一般有数值型和类别型两种,数值型属性指年龄、薪酬等这类属性值为数值的属性;类别型属性指职业、爱好等这类属性值离散的属性.为相对精确地进行属性概化,以减小信息损失,我们对这两种类型的属性分别使用不同的概化方法.对某个超点 V_i,记其中各节点在属性 $t(t=1,2,\ldots,s)$上的取值集合为 $a_t(V_i)$,设属性 t 为数值型,则其属性概化值取为$[\min a_t(V_i),\max a_t(V_i)]$.例如,在某一聚类中,各节点在年龄属性取值上,最大值和最小值分别为 35 和 25,则该超点的属性 t 取值即为[25,35].若属性 t 为类别型,则属性概化需要按预定的概化层次树进行.图 2 所示为一棵概化层次树,其叶子节点对应该类别型属性上各个用户节点的具体属性值,不同层次上的分支节点对应为其所辖各叶子节点的属性概化值,根节点代表最高层次的属性概化值.V_i 内各节点在某类别型属性 t 上的概化属性值为相应概化层次树的一个分支节点,在以该分支节点为根节点的子树

中,各叶子节点值构成的集合即为 $a_t(V_i)$,即,匿名后超点 V_i 在类别属性 t 上的取值对应为概化层次树上以各 a_{jt} 值($a_{jt}\in a_t(V_i)$)为叶子节点的最小上界节点.如图 2 中,“初中老师”和“高校老师”的概化属性值为“教师”,“法官”和“机关干部”的概化属性值为“公务员”.

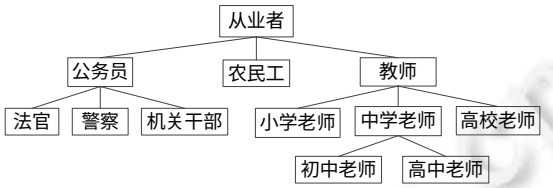


Fig.2 Generalization hierarchy tree for occupation attribute

图 2 职业属性概化层次树

显然,超点的属性概化会造成一定的信息损失.为此,我们依据属性值分属数值型或类别型的不同,分别讨论其信息损失的评估方法.$\forall t\in\{1,2,\ldots,s\}$,若属性值 a_{it} 为类别型,记 a_{it}^* 为其概化值,$h(a_{it}, a_{it}^*)$为对应概化层次树上 a_{it} 到 a_{it}^* 的路径长度,即,叶子节点 a_{it} 到分支节点(甚至是根节点) a_{it}^* 间的层次数.定义超点 V_i 在属性 t 上的概化信息损失值 $infl(V_i,t)$为

$$infl(V_i,t)=\begin{cases}|V_i|\times\dfrac{\max a_t(V_i)-\min a_t(V_i)}{\max a_t(V)-\min a_t(V)}, & \text{if 属性}t\text{为数值型}\\ \displaystyle\sum_{\forall v_j\in V_i}\frac{h(a_{jt},a_{jt}^*)}{h(a_{jt},root)}, & \text{if 属性}t\text{为类别型}\end{cases}.$$

由此可得出属性图 G 的属性信息损失平均值 NAIL(normalized attribute information loss)为

$$NAIL=\sum_{i=1}^{m}\sum_{t=1}^{s}infl(V_i,t)\Big/n,\ NAIL\in[0,1].$$

由属性图到其 k-匿名图,超点“浓缩”也会造成结构信息损失.针对结构信息损失的评估,我们依照文献[10]中 NSIL(normalized structural information loss)的计算方法,$NSIL\in[0,1]$.可见,本文方法造成的信息损失包括结构信息损失和属性信息损失两部分.为此,将两者平均,得出总体信息损失评估值 MTIL(metric for total information loss)计算为 $MTIL=(NAIL+NSIL)/2$.分析 NAIL 的计算不难看出,其实际表示的是平均属性信息损失.文献[10]中 NSIL 的计算方法也表明,其代表的是平均结构信息损失.因此,MTIL 实际表示的是总体上的平均信息损失.

3 节点间综合相似度计算

本文要综合依据节点的社交关系信息和属性信息进行聚类匿名,为此,需要形式化表示节点间连接关系和属性值关联性,我们定义节点间结构相似度和属性相似度来分别衡量彼此连接关系和属性值相关性.

定义 **6**(结构相似度). 表示属性图中节点间的结构相似性程度的实数量,称为结构相似度.$\forall v_i,v_j\in V(G)$,记 $SimS(v_i,v_j)$为 v_i 和 v_j 的结构相似度,$SimS(v_i,v_j)\in[0,1]$.v_i 和 v_j 的结构相似性程度越高,则 $SimS(v_i,v_j)$的值越大.

结构相似度反映社交关系的相近程度.一般而言,两个节点共同拥有的邻接点数越多,则它们间的社交关系越相近,故定义结构相似度的计算为

$$SimS(v_i,v_j)=\frac{|\tau(v_i)\cap\tau(v_j)|}{|\tau(v_i)\cup\tau(v_j)|}.$$

定义 **7**(属性相似度). 表示属性图中节点间的属性相似性程度的实数量,称为属性相似度.$\forall v_i,v_j\in V(G)$,记 $SimA(v_i,v_j)$为 v_i 和 v_j 的属性相似度,$SimA(v_i,v_j)\in[0,1]$.v_i 和 v_j 的属性相似性程度越高,则 $SimA(v_i,v_j)$的值越大.

属性相似度衡量属性向量值的近似程度,由两个属性向量值中相应各属性值对的相似度求均值或加权平均求得.因此,我们先按属性类别为数值型或分类型的不同给出属性值对的相似度计算.$\forall t\in\{1,2,\ldots,s\}$,记 $SA(a_{it},a_{jt})$表示一组属性值对 a_{it} 和 a_{jt} 的相似度,定义 $SA(a_{it},a_{jt})$的计算如下:

$$SA(a_{it},a_{jt})=\begin{cases}1-\dfrac{abs(a_{it}-v_{jt})}{\max a_t(V)-\min a_t(V)}, & \text{if 属性}t\text{为数值型}\\ \dfrac{1}{path(a_{it},a_{jt})}, & \text{if 属性}t\text{为类别型}\end{cases}.$$

这里,$path(a_{it},a_{jt})$表示概化层次树上 a_{it} 和 a_{jt} 间的最短路径长度.规定:$a_{it}=a_{jt}$ 时,$path(a_{it},a_{jt})=1$.

对数值型属性而言,两个属性值相差越大,其相似度越小,上述公式精确地度量了两者的近似程度;而对类别型属性而言,两个属性值越相近,则在概化层次树上以两者的最小上界节点为根节点的子树的高度也越小,两者的最短路径长度就越小,故,公式也基本合理地量化了两者的相似度.一般情况下,不区分各个属性的权重,属性相似度可以由各属性值对的相似度计算平均值得到,故给出属性相似度计算为

$$SimA(v_i,v_j)=\sum_{t=1}^{s}SA(a_{it},a_{jt})\Big/s.$$

定义 **8**(综合相似度). 综合相似度由结构相似度和属性相似度各占一定比例构成,在结构和属性两个方面,综合表示节点间相似程度.$\forall v_i,v_j\in V(G)$,v_i 和 v_j 的综合相似度记为 $Sim(v_i,v_j)$,$Sim(v_i,v_j)\in[0,1]$.$Sim(v_i,v_j)$值越大,则 v_i 和 v_j 在结构和属性上的综合相似程度越高.

设结构相似度占权重$\theta\in(0,1)$,属性相似度则占权重$(1-\theta)$,两节点间综合相似度 $Sim(v_i,v_j)$计算如下:

$$Sim(v_i,v_j)=\theta\times SimS(v_i,v_j)+(1-\theta)\times SimA(v_i,v_j) \tag{1}$$

在上述公式(1)中,θ值决定了节点的结构或属性特性在综合相似度计算中所占比例.一般来说,θ越大,意味着聚类时是更多地依据连接关系,不利于属性值相近的节点聚类在一起,将导致属性信息损失增大.为平衡两者的属性信息损失,使总体信息损失趋于更小,本文研究取θ=0.5.对 n 个节点的社交网络图 G,按上述方法求出任意节点间的综合相似度,即构成一个 n 阶相似度矩阵 $R_{Sim}=(r_{i,j})_{n\times n}$,其中,$r_{i,j}$ 表示矩阵中第 i 行、第 j 列的一个元素,显然,$r_{i,j}=Sim(v_i,v_j)$.在下文属性图聚类匿名算法中,我们还要考虑单个节点与超点的相似度,单个节点与超点间的相似度可以在两节点间相似度的基础上定义出来.设任意节点 v_i、超点 V_x,$Sim(v_i,V_x)$表示 v_i 和 V_x 间的综合相似度,则 $Sim(v_i,V_x)=\sum_{v_j\in V_x}Sim(v_i,v_j)\Big/|V_x|$.

4 社交网络图的聚类匿名

为实现社交网络数据发布隐私保护,属性图聚类匿名化后,每个超点内部结构并不公开,只是给出每个超点的统计信息,包括超点中所包含的节点个数、连接边数和概化后的属性向量值.本节提出一种根据节点间综合相似度进行聚类划分的方法来实现社交网络图的 k-匿名化,简称 CAA-VS 算法(clustering-based anonymization algorithm for social network according to vertical similarity).算法的主要思路是:将图中的 n 个节点,根据综合相似度聚类为一些簇,使每个簇中至少包含 k 个节点,将簇内所有节点和边构成的子图用一个超级节点取代,同时隐匿超点内各节点的属性值,只发布其概化属性值.在任意两个簇之间,只要原图中有一条边相连,则对应的两个超级节点间有一条边相连.构造一个聚类的方法是:首先从剩余节点(即还没有归入到任何聚类中的节点)中任选一个作为起始节点,构成一个新聚类,再每次从剩余节点中,按节点与超点相似度最大原则,选取节点添加到聚类中,直到该聚类中节点数大于或等于 k 为止.下面给出 CAA-VS 算法的伪代码描述.

算法 **1**. 属性图的聚类匿名算法 CAA-VS.

输入:属性图 G,匿名参数 k,结构相似度权重θ;

输出:k-匿名图 G^*.

步骤:

1. $R_{sim}=calculating(G,\theta)$; //根据公式(2)求图 G 的相似度矩阵 R_{sim}
2. $V(G^*)=\varnothing$; $i=1$;
3. while ($|V(G)|\geq k$) //逐个构造图节点聚类

   {$\forall v_t\in V(G)$, $V_{i++}=\{v_t\}$; $V(G)=V(G)-\{v_t\}$;

While ($|V_i|<k$) //构造当前聚类 V_i

{$V_i=V_i\cup\{v_p\}$ where $Sim(v_p,V_i)=\max\limits_{v_j\in V(G)} Sim(v_j,V_i)$; //选取综合相似度最大的节点加入 V_i

$V(G)=V(G)-\{v_p\}$;

}

$V(G^*)=V(G^*)\cup V_i$;

}

4. while ($V(G)\neq\varnothing$) //逐个将剩余节点归入各聚类中

{$\forall v_t\in V(G), V_r=V_r\cup\{v_t\}$ where $Sim(v_t,V_r)=\max\limits_{V_i\in V(G^*)} Sim(v_t,V_i)$; //选取相似度最大的类归入

$V(G)=V(G)-\{v_t\}$;

}

5. for each of $V_i\in V(G^*)$ anonymizing V_i; //对每个超点进行匿名化处理

以上 CAA-VS 算法中,主要工作在步骤 1、步骤 3 和步骤 5,步骤 1 求相似度矩阵是要求出$(n\times n)/2$ 个综合相似度值(因为是对称矩阵),然而,根据公式(2)求一次综合相似度值可在 $O(1)$时间里完成,故步骤 1 的时间复杂度为 $O(n^2)$.步骤 3 是一个二层嵌套循环结构,其中:外层循环次数为所划分的聚类个数 m,不难求出 $m\leq n/k$;内层循环次数为 k,内层循环体共执行约 n 次.而内层循环体每次执行主要完成$|V(G)|$次($|V(G)|$表示当前 $V(G)$中剩余节点个数)相似度比较操作,其时间复杂度为 $O(n)$,故步骤 3 的时间复杂度也是 $O(n^2)$.步骤 5 的匿名化处理是对各个超点进行属性概化和子图隐匿,其主要工作在属性概化,已有文献表明,其时间复杂度为 $O(n^2)$,算法中其他步骤的时间复杂度均低于 $O(n^2)$,因此,本算法总的时间复杂度为 $O(n^2)$.

CAA-VS 算法根据综合相似度将属性图 G 的节点聚类为一些超点,以超点内节点数和边数的统计信息取代超点内的子图结构,反映了节点及结构隐藏的思想,能够抵御各类基于图结构信息的攻击,如基于节点度数的识别攻击和基于节点 d-邻域的匹配攻击.以超点为单位将其内部各节点的属性值概化成一个统一的属性值,并以超点概化属性值取代该超点内各节点的属性值发布,从而隐藏了具体用户的属性信息,这体现为(值域)泛化技术的思想,能够抵御以各种属性信息为背景知识的链接攻击和针对属性隐私本身的盗取.由此可见:本文算法融合了隐私保护技术中的隐匿和泛化等技术,能够有效防范针对图节点属性隐私以及基于图结构或节点属性等一切背景知识的各类隐私攻击,将其攻击成功率至少降至 $1/k$ 以内,因而具有很强的隐私保护效果.

信息损失直接关联数据的可用性,理论上说,数据匿名化处理造成的信息损失量越大,意味着原数据得到了更大程度的隐匿,相应地,数据的可用性将变得越差;相反,匿名操作带来的信息损失越小,则意味着数据可用性得到了更大程度的保持.在上述 CAA-VS 算法中,聚类过程中的每一步聚类操作均依据综合相似度最小原则选取节点进行,符合信息损失量最小原则.因为根据前文第 3 节分析,节点间属性相似度越大,则属性值越相近,在属性概化时,各节点属性值越为相近,则概化后的属性值越为接近各个原属性值,由此带来的属性信息损失量就越小;另外,一般节点间结构相似度越大,则社交关系越为相近,故理论上连接紧凑的节点更容易聚为一类,由文献[10]定义的 NSIL 计算办法可知,据此聚类带来的结构信息损失也越小.可见,本文的属性图聚类匿名方法能够保证总体信息损失趋于最小,故数据可用性较高.

5 实验与结果分析

本节通过实验分析本文算法 CAA-VS 的有效性,考虑典型聚类匿名算法 SaNGreeA[10]和我们前期提出的 GA-CAG[16]算法在图聚类的功能要求、信息损失计算方法、隐私保护的对象和目标等方面与本文方法有更近的相似性和可比性,我们选择这三者进行实验对比和分析.实验数据来自 UC Irvine Machine Learning Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html)[10]的 Adult 数据集,分别从中任意取 600 个节点和 1 000 个节点构成的社交网络,记为 *DB*1 和 *DB*2,各组实验均在这两个真实数据集上运行.实验环境为:Intel(R) Core(TM) i5-4210U CPU@1.70GHz(2394MHz);4.00GB(1600MHz)内存;希捷 ST ST500LM000-SSHD-8GB(500GB)主硬盘;

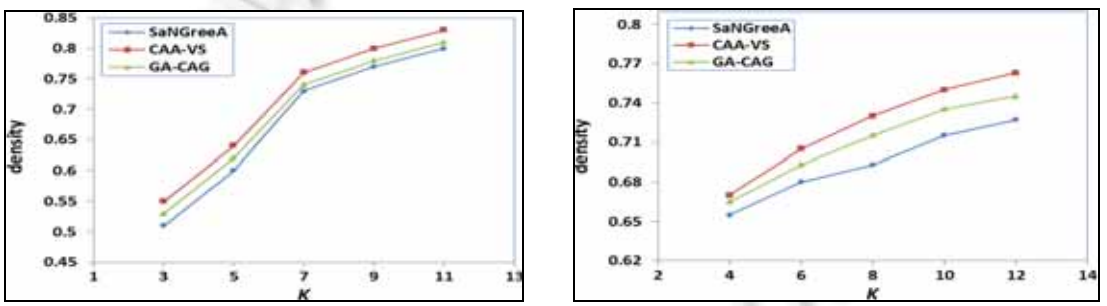
Microsoft Windows 8.1 中文版(64 位)操作系统;算法均采用 Microsoft Visual C++ 7.0 实现.考虑算法在构造新聚类时随机选取初始节点会导致聚类结果有细微差别,每组实验重复进行 5 次,结果取其平均值.

5.1 聚类质量分析

本文方法的主要工作基础是依据节点综合相似度进行属性图聚类,故本小节对图聚类质量进行实验分析.引入聚类密度和聚类熵两个参数来衡量聚类质量,它们可以分别反映连接结构和节点属性上的聚类效果[18],两者分别定义如下.

- 聚类密度: $density(\{V_i\}_{i=1}^m)=\sum_{i=1}^{m}\frac{|\{(v_p,v_q)\,|\,v_p,v_q\in V_i,(v_p,v_q)\in E\}|}{|E|}$;
- 聚类熵: $entropy(\{V_i\}_{i=1}^m)=\sum_{i=1}^{m}\frac{|V_i|}{|V|}\left(-\sum_{t=1}^{s}\sum_{\alpha_{jt}\in a_t(V_i)}(p_{ij}^{\alpha_{jt}}\log p_{ij}^{\alpha_{jt}})\right)$, $p_{ij}^{\alpha_{jt}}$ 表示聚类 i 中的节点在属性 t 上取值为 $\alpha_{jt}\in a_t(V_i)$的概率.

本节进行两组实验,第 1 组实验是考察算法的聚类密度.图 3(a)和图 3(b)分别显示了在 *DB*1 和 *DB*2 上各算法所得 *density* 值随 *k* 值变化的规律.从中可见:在两组数据集上,*density* 值均随着 *k* 值的增大而明显增大.这是因为 *k* 值增大则各聚类中节点数目增多,聚类数量减少,导致聚类外部连接各聚类的边数减少,属于聚类内部边的数目相应增多,故 *density* 值增大.在 *DB*1 上,3 种算法的 *density* 值一直很接近;而在 *DB*2 上,随着 *k* 值的增大,*density* 值差距有略微增大.这反映在聚类密度衡量的聚类效果上,3 种算法在数据量不大或聚类很小时非常相近;而随着数据和所得聚类的规模变大,差别也更为明显;但总体比较接近,本文算法略优.

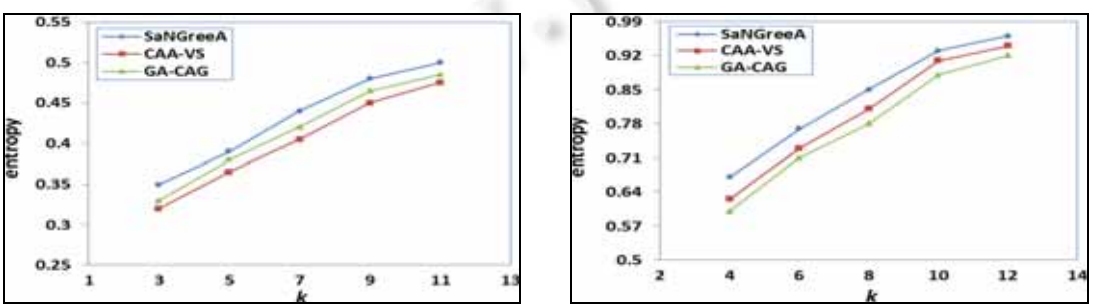


(a) *DB*1 (b) *DB*2

Fig.3 The impacts of changing *k* on the cluster densities in the datasets of *DB*1 and *DB*2

图 3 在 *DB*1 和 *DB*2 数据集上的聚类密度变化情况

第 2 组实验是分别在两个数据集上比较算法的聚类熵,图 4 显示了该组实验结果.



(a) *DB*1 (b) *DB*2

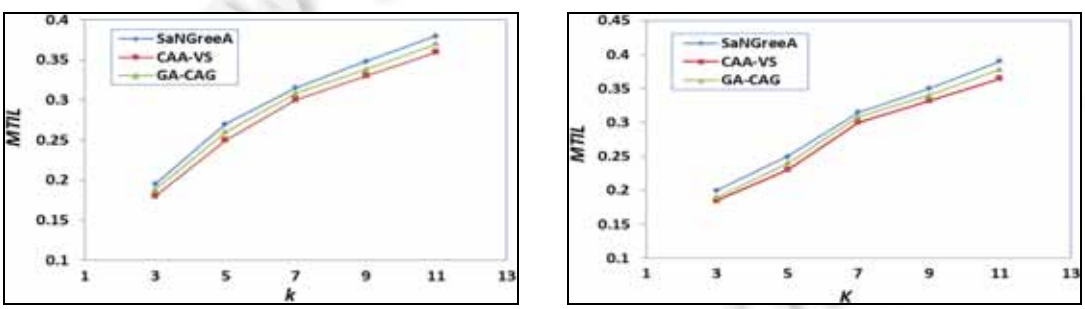
Fig.4 The impacts of changing *k* on the cluster entropies in the datasets of *DB*1 and *DB*2

图 4 在 *DB*1 和 *DB*2 数据集上的聚类熵变化情况

由于 k 值增大带来的聚类变大,即$|V_i|$值变大,使得类中节点属性值更丰富,反映在 $p_{ij}^{\alpha_{jt}}\log p_{ij}^{\alpha_{jt}}$ 值上的各属性值的离散程度也呈变大趋势.故根据聚类熵计算公式,其值随 k 的增大而增大.图 4 也验证了这点.另外,当数据规模变大,不同属性值的个数将更多,则属性值分布的离散程度会变大,也会带来聚类熵增大.与图 4(a)相比,图 4(b)上聚类熵值偏大正证实了这点.图 4 还表明:3 种算法聚类熵值相差较小,在 *DB*1 上,CAA-VS 和 SaNGreeA 分别有最低和最高的熵值;而在 *DB*2 上,GA-CAG 具有更小的熵值.反映当数据量变大时,GA-CAG 强大的搜索优化能力显现一定优势.

5.2 算法性能分析

本节从数据信息损失和算法运行时间两个方面,各通过一组实验讨论算法性能.本文算法构成的信息损失来自超点内子图隐匿带来的结构信息损失,以及超点属性概化带来的属性信息损失,因此,我们以本文第 2 节定义的 MTIL 计算值来代表数据信息损失量.图 5 反映了本文算法分别在两组数据上运行时,MTIL 值随 k 值变化的情况.图中显示,MTIL 值随 k 值正向变化.这是由于 k 值变大,则超点内节点个数增多,在任意属性上的节点属性值将更加丰富,对其概化时,就可能需要取一个更为宽泛的概化属性值.即:各个属性上的概化程度都可能变大,导致平均属性信息损失 NAIL 变大;同时,各超点子图规模变大,子图隐匿时将造成更多的边被隐藏,导致平均结构信息损失 NSIL 值也增大.比较图 5(a)和图 5(b)发现:两者的 MTIL 值非常接近,说明数据规模对 MTIL 值没有多大影响.这是因为 MTIL 表示的是平均信息损失,主要与匿名方法有关;而总体看,图 5(b)的 MTIL 略微偏大.主要是因为 n 变大可能使平均每个聚类中有更大比例的不同属性值被隐匿,以致 NAIL 有细微变大趋势.图 5 也表明,CAA-VS 有更低的信息损失值,反映出本文算法在减小信息损失方面的有效性.

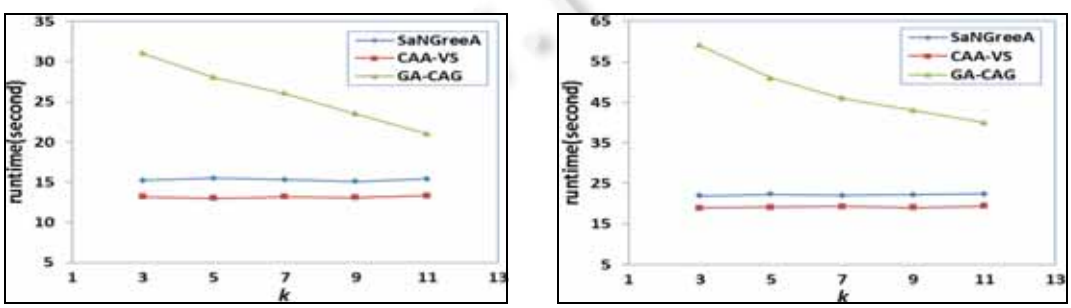


(a) *DB*1 (b) *DB*2

Fig.5 The impacts of changing k on the information loss (MTIL) in the datasets of *DB*1 and *DB*2

图 5 在 *DB*1 和 *DB*2 数据集上的数据信息损失变化情况

为考察本文算法的时间效率,最后一组实验就 3 种算法分别在 *DB*1 和 *DB*2 上的运行时间做出测试分析.图 6 显示了本组实验结果.



(a) *DB*1 (b) *DB*2

Fig.6 The impacts of changing k on the runtime in the datasets of *DB*1 and *DB*2

图 6 在 *DB*1 和 *DB*2 数据集上的运行时间变化情况

图 6 显示:随着 k 值的变大,GA-CAG 的运行时间有明显减小趋势,而 CAA-VS 和 SaNGreeA 的基本没变.这说明对 CAA-VS 和 SaNGreeA 算法来说,k 值的变化基本不会影响运行时间.分析两者的算法过程也可以发现,k 值的变化不会明显影响算法运算量.而对 GA-CAG 算法来说,由于 k 值的增大会减少交叉算子的运算量,故其运行时间有比较明显的减少.图 6 也显示:本文算法用时最少,而 GA-CAG 有最大用时.说明 CAA-VS 一次完成图节点的聚类划分,有着更好的时间效率;而 GA-CAG 利用遗传算法来迭代搜索最优解,需要更多的耗时.比较图 6(a)和图 6(b)可见,各算法在 *DB*2 上的运算时间更高.这是因为 *DB*2 的数据规模更大,算法运行过程需要处理更多的数据,故耗时更多.

6 结束语

在社交网络图中,个体节点或边蕴含的隐私信息可能因遭受到恶意盗取而泄露.考虑同时保护两者信息,防止一切以连接边和属性值为背景知识的攻击,从而保障社交网络图数据发布的隐私安全,本文提出一种基于结构和属性综合相似度的属性图聚类匿名方法.该方法分别定义了节点间的结构相似度和属性相似度,依据两者的综合相似度利用贪心法实现属性图聚类划分,最后对各个聚类进行属性概化和子图隐匿的匿名处理.实验也验证了该方法在实现聚类质量和算法性能方面的有效性.后续工作中,我们将针对目前社交网络中个性化隐私保护不足的问题展开深入研究.

References:

[1] Liu XY, Wang B, Yang XC. Survey on privacy preserving techniques for publishing social network data. Ruan Jian Xue Bao/Journal of Software, 2014,25(3):576−590 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4511.htm [doi: 10.13328/j.cnki.jos.004511]

[2] Sweeney L. *K*-Anonymity: A model for protecting privacy. Int'l Journal on Uncertainty, Fuzziness, and Knowledge-Based Systems, 2002,10(5):557−570. [doi: 10.1142/S0218488502001648]

[3] Gong NZ, Talwalkar A, Mackey L, Huang L, Shin ECR, Stefanov E, Shi E, Song D. Jointly predicting links and inferring attributes using a social-attribute network (SAN). In: Proc. of the SNA-KDD. 2012.

[4] Fu YY, Fu H, Xie X, Sun GZ, Zhang M. Anonymity and privacy preservation for social network. Communications of the CCF, 2014,10(6):51−58 (in Chinese).

[5] Liu P, Li XX. An improved privacy preserving algorithm for publishing social network data. In: Proc. of the IEEE Int'l Conf. on High Performance Computing and Communications & IEEE Int'l Conf. on Embedded and Ubiquitous Computing. 2013. 888−895. [doi: 10.1109/HPCC.and.EUC.2013.127]

[6] Zou L, Chen L, Özsu MT. *K*-Automorphism: A general framework for privacy preserving network publication. Proc. of the VLDB Endowment, 2009,2(1):946−957. [doi: 10.14778/1687627.1687734]

[7] Yuan M, Chen L, Yu PS, Yu T. Protecting sensitive labels in social network data anonymization. IEEE Trans. on Knowledge and Data Engineering, 2013,23(3):633−647. [doi: 10.1109/TKDE.2011.259]

[8] Masoumzadeh A, Joshi J. Preserving structural properties in edge-perturbing anonymization techniques for social networks. IEEE Trans. on Dependable and Secure Computing, 2012,9(6):877−889. [doi: 10.1109/TDSC.2012.65]

[9] Zheleva E, Getoor L. Preserving the privacy of sensitive relationships in graph data. In: Proc. of the Privacy, Security, and Trust in KDD. Berlin, Heidelberg: Springer-Verlag, 2008. 153−171. [doi: 10.1007/978-3-540-78478-4_9]

[10] Campan A, Truta TM. Data and structural *k*-anonymity in social networks. In: Proc. of the Privacy, Security, and Trust in KDD. Berlin, Heidelberg: Springer-Verlag, 2009. 33−54. [doi: 10.1007/978-3-642-01718-6_4]

[11] Tassa T, Cohen DJ. Anonymization of centralized and distributed social networks by sequential clustering. IEEE Trans. on Knowledge and Data Engineering, 2013,25(2):311−324. [doi: 10.1109/TKDE.2011.232]

[12] Skarkala ME, Maragoudakis M, Gritzalis S, Mitrou L, Toivonen H, Moen P. Privacy preservation by *k*-anonymization of weighted social networks. In: Proc. of the IEEE/ACM Int'l Conf. on Advances in Social Networks Analysis and Mining. 2012. 423−428. [doi: 10.1109/ASONAM. 2012.75]

[13] Fu YY, Zhang M, Feng DG, Chen KQ. Attribute privacy preservation in social networks based on node anatomy. Ruan Jian Xue Bao/Journal of Software, 2014,25(4):768−780 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4565.htm [doi: 10.13328/j.cnki.jos.004565]

[14] Hay M, Li C, Miklau G, Jensen D. Accurate estimation of the degree distribution of private networks. In: Proc. of the IEEE Int'l Conf. on Data Mining. 2009. 169−178. [doi: 10.1109/ICDM.2009.11]

[15] Sala A, Zhao X, Wilson C, Zheng H, Zhao BY. Sharing graphs using differentially private graph models. In: Proc. of the ACM SIGCOMM Conf. on Internet Measurement Conf. 2011. 81−98. [doi: 10.1145/2068816.2068825]

[16] Jiang HW, Zeng GS, Hu KK. A graph-clustering anonymity method implemented by genetic algorithm for privacy-preserving. Journal of Computer Research and Development, 2016,53(10):2354−2364 (in Chinese with English abstract).

[17] Wang R, Zhang M, Feng D, Fu Y. A clustering approach for privacy-preserving in social networks. In: Proc. of the Information Security and Cryptology (ICISC 2014). Springer Int'l Publishing, 2014. 193−204. [doi: 10.1007/978-3-319-15943-0_12]

[18] Han QL, Zhao HB, Pan HW, Yin GS, Chang JY. Research on spatio-temporal object graph clustering algorithm based on structure and attribute. Journal of Computer Research and Development, 2013,50(Suppl.):154−162 (in Chinese with English abstract).

附中文参考文献:

[1] 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. http://www.jos.org.cn/1000-9825/4511.htm [doi: 10.13328/j.cnki.jos.004511]

[4] 付艳艳,付浩,谢幸,孙广中,张敏.社交网络匿名与隐私保护.中国计算机学会通讯,2014,10(6):51−58.

[13] 付艳艳,张敏,冯登国,陈开渠.基于节点分割的社交网络属性隐私保护.软件学报,2014,25(4):768−780. http://www.jos.org.cn/1000-9825/4565.htm [doi: 10.13328/j.cnki.jos.004565]

[16] 姜火文,曾国荪,胡克坤.一种遗传算法实现的图聚类匿名隐私保护方法.计算机研究与发展,2016,53(10):2354−2364.

[18] 韩启龙,赵洪斌,潘海为,印桂生,常吉羽.基于结构——属性的时空对象图聚类算法的研究.计算机研究与发展,2013,50(增刊):154−162.

姜火文(1974－),男,江西南昌人,博士,副教授,CCF 学生会员,主要研究领域为隐私安全,软件演化,智能计算.

刘文娟(1989－),女,硕士,主要研究领域为系统扩展,大数据处理.

占清华(1979－),男,硕士,讲师,主要研究领域为信息安全,病毒防治和入侵检测.

马海英(1977－),女,博士,副教授,主要研究领域为隐私保护,公钥密码学,网络安全.