基于聚类的社交网络隐私保护方法

周艺华 张 冰 杨宇光 侍伟敏

（北京工业大学信息学部 北京 100124）（可信计算北京市重点实验室 北京 100124）

摘 要 随着社交网络的迅速发展，社交网络积累了大量的数据，它们在一定程度上反映了社会规律。针对如何在保证隐私安全的前提下挖掘出有效知识的问题，提出了基于聚类的社交网络隐私保护方法，该方法具有隐私保护力度自适应、匿名模型安全性和有效性高的特点。该方法基于用户信息和社交关系进行聚类，将社交网络中的所有节点根据节点间的距离聚类为至少包含 k 个节点的超点，并对超点进行匿名化处理。匿名后的超点能够有效地防范以节点属性隐私、子图结构等为背景知识的各类隐私攻击，使攻击者无法以大于 $1/k$ 的概率来识别用户。根据聚类算法和社交网络的特点优化聚类过程中初始节点的选取算法和节点间距的计算方法；同时通过结合自适应思想，优化隐私保护力度的选取方法，有效地减少了信息损失，提高了数据有效性。在 Matlab 上使用不同的数据集进行实验验证，结果表明所提算法在信息损失和运行时间上均优于其他相关方法，进一步证明了它的有效性和安全性。

关键词 社交网络，隐私保护，k-匿名，聚类，信息安全

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/jsjkx.180901749

Cluster-based Social Network Privacy Protection Method

ZHOU Yi-hua ZHANG Bing YANG Yu-guang SHI Wei-min

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

(Beijing Key Laboratory of Trusted Computing, Beijing 100124, China)

Abstract With the rapid development of social networks, social networks have accumulated a large amount of data, which reflect the social laws to some extent. Aiming at mining effective knowledge under the premise of ensuring privacy, this paper proposed a clustering-based social network privacy protection method. The method has the characteristics of adaptive privacy protection strength, high security and effectiveness of anonymity model. Clustering is conducted based on user information and social relationships. It clusters all nodes in the social network into a super point containing at least k nodes according to the distance between nodes, and then the super points are anonymized. Anonymous super points can effectively prevent various types of privacy attacks taking node attribute privacy and sub-graph structure as background knowledge, so that attackers cannot identify users with a probability greater than $1/k$. According to the characteristics of clustering algorithm and social network, the initial node selection algorithm and node spacing calculation method in clustering process are optimized, and by combining the adaptive thinking, the selection method of privacy protection strength is also optimized, which effectively reduces information loss and improves data validity. Experiments were carried out on Matlab platform with different data sets. The results show that the proposed method is superior to other related methods in terms of information loss and running time, which further proves its effectiveness and security.

Keywords Social network, Privacy protection, K-anonymity, Clustering, Information security

1 引言

互联网技术的迅猛发展，促进了各种社交网络平台的兴起。社交网络将人们的生活和人们之间的联系搬到了互联网上，因此积累了大量的信息，这些信息在一定程度上反映了社会规律，形成了一类具有重要研究意义和应用价值的数据。目前，针对社交网络隐私保护问题已涌现出了很多保护技术，在技术实现上最简单的方法就是只对用户身份信息做隐藏处理，而不对其他信息进行处理。虽然这种技术在一定范围内对用户的个人隐私进行了保护，但是恶意者仍然可以通过目标者的社交网络关系的背景知识识别出个体的身份[1-3]，导致用户隐私泄露。如图 1 所示，为保证用户的隐私安全，数据挖掘人员不直接对数据进行分析，而是对经过数据隐私保护处理之后的数据进行数据挖掘和分析。

到稿日期：2018-09-16 返修日期：2019-04-21 本文受北京市自然科学基金资助项目(4182006)，国家自然科学基金项目(61572053)资助。

周艺华(1969－)，男，博士，副教授，主要研究方向为网络与信息安全、密码学、可信计算，E-mail：zhouyh@bjut.edu.cn；**张 冰**(1994－)，女，硕士生，主要研究方向为信息安全；**杨宇光**(1976－)，女，博士，教授，主要研究方向为信息安全与其他学科的交叉；**侍伟敏**(1978－)，女，博士，主要研究方向为网络与信息安全、密码学、信息安全与其他学科的交叉。

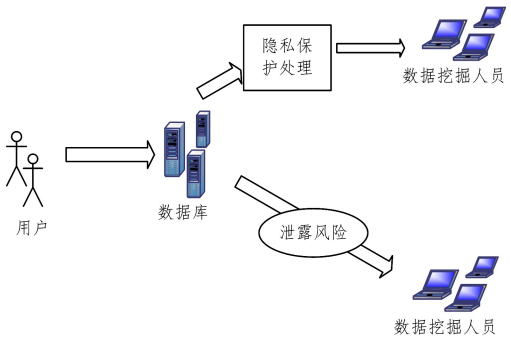

图 1　社交网络数据挖掘模型

Fig. 1　Social network data mining model

在社交网络分析领域中，由于社交网络数据在一定程度上反映了社会规律，因此有效地挖掘社交网络数据的本质具有很重要的研究意义，而有效并准确地进行挖掘的前提是要保证社交网络信息的准确性。当前社交网络分析人员从各个方面和角度研究这些网络的本质特性和规律，因此需要保证数据在宏观上是准确的，局部信息某种程度的不准确性并不会影响到宏观性质的研究。

k-匿名模型[4-5]是一种有效的隐私保护模型。聚类技术就是结合 k-匿名的思想，把一些特定点与边归于相应的集合中，这些集合称为超级节点。在每个超级节点中至少有 k 个“相同”的个体，这些个体的细节将被隐藏[6-7]。聚类已被很好地应用于数据库隐私保护中。社交网络经过聚类进行数据隐私保护后，社交网络分析者仍可利用聚类后的图形特征来考查原始图的宏观特性。

现有的大部分基于聚类的隐私保护算法都是基于子图 k-匿名模型。子图 k-匿名模型是假设攻击者以特定子图的结构信息为背景知识进行攻击，构造 k 个相同结构的子图作为候选目标，使目标被识别的概率低于 $1/k$。但是在现实生活中，恶意攻击者获取的背景知识多种多样，主要可以分为社交关系结构知识和用户的属性信息。基于子图 k-匿名模型的隐私保护算法能够有效防范以子图结构特性为背景知识的攻击，却无法有效抵御以用户属性和子图结构为背景知识的攻击。同时，现有的匿名算法对社交网络中的所有节点进行隐私保护时的隐私保护力度是相同的，然而在社交网络中，不同节点对隐私保护力度的需求不同，如果对它们使用相同的隐私保护力度，则会降低需要更高隐私保护力度节点的安全性，并使需要更低隐私保护力度的节点损失更多的信息。

为了提高数据的安全性和有效性，本文提出了一种基于聚类算法的社交网络隐私保护方法。本文的主要贡献如下：

(1)根据社交网络的实际情况，建立实际问题的社交网络模型，并在社交网络图上进行聚类，完成数据匿名化处理，形成匿名社交网络；

(2)根据社交网络的特点，分别量化网络节点之间的用户社交关系和用户信息的“距离”，并计算网络节点之间与聚类形成的超点间的距离来进行聚类；

(3)针对社交网络和聚类算法的特点，以及不同用户对隐私保护力度的需求不同的问题，结合聚类系数和节点密度优化聚类算法，对数据进行不同程度的隐私保护。实验证明，文中所提算法在提高数据安全性的同时降低了信息损失量，提高了数据有效性。

2　相关工作

现有的社交网络隐私保护算法主要分为两类：图结构修改法和图结构泛化法。

图结构修改法是通过一定方法对图的节点和边进行删除、添加等操作来修改图的结构以进行匿名，从而达到保护图的节点和边的目的，防止其被识别。Zou 等[8]提出了一种基于贪心的修改算法进行边的插入，直到目标节点至少与其他 $k-1$ 个节点的邻居结构不可区分为止，达到了 k-匿名的目的。该方法防御了具有与目标节点直接邻接的子图结构的背景知识的攻击。

图结构泛化法最早是由 Zheleva 等[9]提出。Campan 等[10]首次提出了量化信息损失的方法，提出了基于贪心的聚类算法 SaNGreeA，达到了保护节点隐私安全的目的。通常，图匿名方法都是针对无权图，Skarkala[11]等针对带权图提出了基于超点和超边的匿名机制。该机制将各边的权重隐匿在超边权重中，只发布超边的权重，通过超点结构使攻击者无法区分目标节点与其余 $k-1$ 个节点，但其缺少对属性信息的考虑。

文中提出的基于聚类的社交网络隐私保护算法具有隐私保护力度自适应、匿名模型安全性和有效性高的特点，其综合考虑了社交网络的社交关系和用户信息，同时针对社交网络的特点优化了算法，减少了信息损失，提高了聚类效果。

3　社交网络模型

3.1　问题模型

社交网络中主要包含两类数据：用户的个人信息数据和社交关系数据。用户的个人信息数据包含可以直接暴露用户真实身份的身份信息，它反映了用户部分特征个性化的属性信息。用户的真实身份信息数据通常被直接隐匿[12]；其他的属性信息数据，比如年龄、性别等，反映了用户的特征和行为，具有一定的研究和挖掘意义，因此这些数据不会被直接公开，需要经过一定的匿名化处理后发布。用户的社交关系信息是指用户与其他用户之间的特定关联信息，反映了用户之间的社会关联关系，也属于用户隐私。直接对社交关系信息进行发布会暴露用户的隐私，因此也需要对其进行匿名处理。在社交网络中，用户的隐私包括用户个人的属性信息和社交关系信息，用户之间的社会联系可以用图结构进行表示，对社交网络进行隐私保护时，需要同时对这两种信息进行隐私保护。

3.2　相关定义

定义 1(社交网络)　社交网络同时包含用户的个人信息和社交关系信息，可以用一个带标签的无向无权图进行描述，表示为 $G=(V,E,A)$。$V=\{v_1,v_2,\cdots,v_n\}$ 为社交网络中的点集，其中 $v_i(i=1,2,\cdots,n)$ 表示社交网络中的任一用户；$E=\{(v_i,v_j)\mid i\neq j,1\leqslant i,j\leqslant n\}$ 为社交网络中的边集，其中 (v_i,v_j) 表示用户 v_i,v_j 之间的社交关系；$A=\{A_1,A_2,\cdots,A_n\}$ 是社交

网络中的属性集合，也是用户个人信息的集合，其中 $A_i=(a_{i1},a_{i2},\cdots,a_{im})$ 是节点 $v_i(i=1,2,\cdots,n)$ 的 m 维属性序列。

定义 2（匿名化社交网络） 给定一个社交网络 $G=(V,E,A)$，根据计算节点之间的相似度进行聚类，使每个聚类中节点的个数大于或等于 k，将社交网络中的所有节点通过聚类划分成簇集合，也就是匿名化社交网络。匿名社交网络形式化表示为对点集 V 进行聚类，生成了簇集合 $S_{clt}=\{clt_1,clt_2,\cdots,clt_s\}$，$\bigcup_{i=1}^{s} clt_i=V$，$clt_i\cap clt_j=\varnothing$，$i,j\in 1,2,\cdots,n$，$i\neq j$。匿名化社交网络 $G^{ano}=(V^{ano},E^{ano},A^{ano})$，其中 $V^{ano}=\{v_{clt_1},v_{clt_2},\cdots,v_{clt_s}\}$，$v_{clt_i}$ 为匿名化网络的一个匿名节点；$E^{ano}=V^{ano}\times V^{ano}$，$\forall v_{clt_i},v_{clt_j}\in V^{ano}$ 且 $\exists j v_1\in clt_i,v_2\in clt_j,(v_1,v_2)\in E$，那么 $(v_{clt_i},v_{clt_j})\in E^{ano}$。

定义 3（属性泛化） 对于匿名化社交网络中的任意一个簇 clt_i，簇内的各个节点 v_j 的所有属性的值 $a_{ji}(i=1,\cdots,m)$ 都被一个取值范围更广泛的变量所取代，这个泛化过程被称为属性泛化。

社交网络中的任意点的属性信息分为数值类型和非数值类型。在泛化过程中，为了减少信息损失，使用不同的方法对这两种类型的数据进行泛化。数值类型的数据有年龄、收入等，对其进行泛化的方法是用数值范围代替具体数值。令 $clt=\{v_1,v_2,\cdots,v_k\}$ 为匿名化社交网络中的任一超点，即簇，其中所有节点在属性 a_p 上的值被泛化为 $[\max(a_p(clt)),\min(a_p(clt))]$。非数值类型有喜好、职业等，对其进行泛化的方法是根据需要，按特定的泛化层次树进行泛化。图 2 给出一棵职业泛化层次树，树的根节点表示泛化范围最大的属性值，不同层次上的分支节点表示其不同泛化范围下的各叶子节点的属性泛化值。

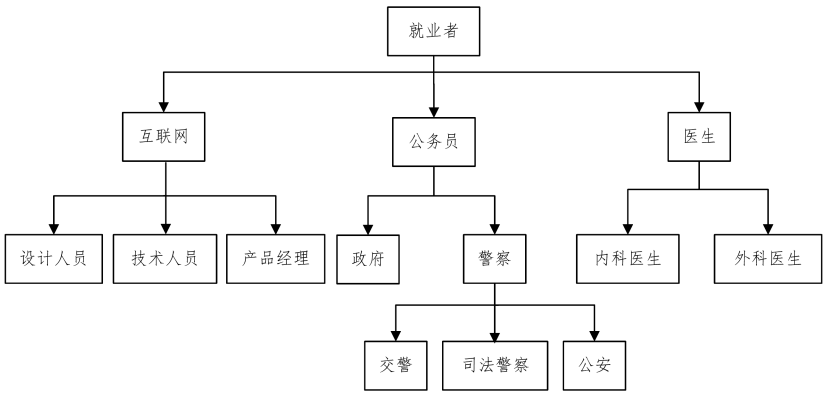

图 2 职业泛化层次树

Fig. 2 Occupational generalization hierarchy tree

定义 4（聚类系数（Clustering Coefficient）） 在图论中，聚类系数表示一个图形中节点的聚集程度。社交网络中的节点具有一定的团体性，有相近的背景信息（如地理位置、社交关系、兴趣爱好等）的节点更容易聚集在一起。建立紧密的社交小团体，小团体内的节点有着相对紧密的连接关系。

在现实世界中，社交网络中节点的局部聚类系数表示其相邻节点形成一个团的紧密程度，节点的聚类系数越大，表示该节点周围形成团的紧密度越大。对于节点 v_i，其直接邻居节点集合为 V_i，V_i 包含的节点数为 N_i，包含的边数为 E_i，则节点的聚类系数为：

$$cfs=\frac{2E_i}{V_i\times(V_i-1)}$$

如图 3 所示，点 A 的邻居节点为 $\{B,C,D\}$，它们之间连接的边数为 2，可能构成的边数为 3，则聚类系数为 2/3；点 D 的邻居节点为 $\{A,B,E\}$，它们之间连接的边数为 1，可能连接的边数为 3，则聚类系数为 1/3。

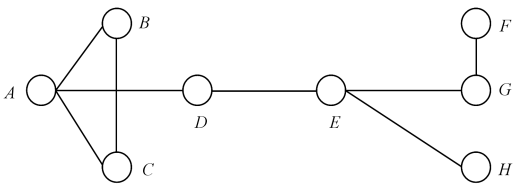

图 3 简单无向图

Fig. 3 Simple undirected graph

定义 5（节点密度） 节点密度表示节点与周围节点的紧密程度，密度值越大表示节点与周围节点越紧密，则其更应该被选取为聚类中心。对于节点集 $V=\{v_1,v_2,\cdots,v_n\}$，其中任一节点 v_i 的密度如下：

$$dens(v_i)=\sum_{j=1}^{n}\frac{1}{\min\limits_{p_{ij}}\sum\limits_{k=1}^{l-1}q^{d(v_k,v_{k+1})}-1}$$

其中，q 为密度的伸缩因子，$q>1$；p_{ij} 是节点 v_i,v_j 之间的所有路径；l 是节点 v_i,v_j 之间的节点数目。

定义 6（结构信息损失） 簇内结构信息损失：对簇 clt 来说，超级节点表示为 $(|clt|,|E_{clt}|)$，其表示超级节点有 $|clt|$ 个节点和 $|E_{clt}|$ 条边。对于簇内节点对来说，其只有 $|E_{clt}|$ 条边；对于超级节点来说，无法确定哪些点相连，从而导致了错误连接，即簇内结构损失。

$$instraCost=(e-|E_{clt}|)\times\frac{|E_{clt}|}{e}+|E_{clt}|\times(1-\frac{|E_{clt}|}{e})$$

其中，$e=\binom{|clt|}{2}$ 表示簇 clt 内可能有的边数。

簇间结构信息损失：产生簇间结构信息损失的原因是两个超级节点即簇之间的若干边经过聚类操作聚合成了一条超级边，因此不能确定超级节点之间的哪些边具体连接了哪些真实节点，从而产生了不确定性，导致簇间结构信息损失。

$$interCost(clt_i,clt_j)=2\times|E_{clt_i,clt_j}|\times(1-\frac{|E_{clt_i,clt_j}|}{|clt_i|\times|clt_j|})$$

全局结构信息损失(Structure Information Loss，SIL)：对超级节点对应簇的簇内信息损失和簇间信息损失的总和进行归一化。

$$SIL=\frac{\sum_{i=1}^{m} instraCost(clt_j)+\sum_{i=1}^{m}\sum_{j=i+1}^{m} interCost(clt_i,clt_j)}{n\times(n-1)/4}$$

定义 7(属性信息损失(Attribute Information Loss，AIL))　属性信息的泛化过程如定义 3 所示，在属性泛化的过程中会产生信息损失，根据属性的连续型和离散型分别进行计算。属性信息损失的计算方法如下：

$$ail(clt_i)=\begin{cases}|clt_i|\times\dfrac{\max(a_{pk})-\min(a_{pk})}{\max(a_p(clt_i))-\min(a_p(clt_i))}, & \text{数值型}\\ \sum\limits_{v_j\in clt_i}\dfrac{deepth(a_{pj},a_{pj}^{*})}{deepth(a_{pj},root)}, & \text{离散型}\end{cases}$$

其中，$clt_i=\{v_1,v_2,\cdots,v_k\}$，$a_{pk}$ 表示属性 a_p 在节点 v_k 上的取值。

$S_{clt}=\{clt_1,clt_2,\cdots,clt_s\}$ 为匿名后的社交网络。$A=\{A_1,A_2,\cdots,A_n\}$ 是社交网络中的属性集合，也是用户的个人信息集合，其中 $A_i=(a_{i1},a_{i2},\cdots,a_{im})$ 是节点 $v_i(i=1,2,\cdots,n)$ 的 m 维属性序列。属性信息损失为：

$$AIL=\frac{\sum_{i=1}^{s}\sum_{t=1}^{m} ail_i(clt_i,a_t)}{n}$$

4　基于聚类的隐私保护算法

为了更好地实现对社交网络的匿名保护，提高数据的安全性和有效性，本文提出了一种基于聚类的社交网络隐私保护算法。该算法的主要思想是：将社交网络的节点根据节点间的综合距离聚类成若干个超点，超点中的具体细节被隐匿，只要这两个超点中的节点有一条边相连，则这两个超点之间就有且仅有一条边相连。

本文提出的算法本质上属于基于超点的匿名方法，但也有其独特之处。

1)为了能够更好地抵御背景知识攻击，该算法结合了社交网络的特点，对社交关系结构和社交属性信息共同进行概化，综合结构信息和属性信息来计算节点间的距离。

2)为了提高数据的有效性，针对社交网络的特点，对聚类算法的初始种子节点的选取进行了优化，提高了聚类质量，减少了信息损失。

3)为了提高数据的安全性，考虑每个节点对隐私保护力度的不同需求，本文提出了基于自适应的隐私保护算法。该算法针对不同需求的节点使用不同的保护力度进行隐私保护，在提高数据安全性的前提下减少了数据的信息损失。

4.1　节点间距离的计算方法

定义 8(结构信息特征距离(Structure Information Feature Distance，SIFD))　在社交网络 $G(V,E,A)$ 中，节点集 $V=\{v_1,v_2,\cdots,v_n\}$，其中任意一个节点 v_i 的邻居关系可以表示为 $Neibor=(nb_i^1,nb_i^2,\cdots,nb_i^n)$，若 v_i,v_j 之间存在社交关系，即存在一条边 $(v_i,v_j)\in E$，$\forall i\neq j$，则有 $nb_i^j=0$，否则 $nb_i^j=1$，那么节点之间的结构特征距离为：

$$dist\_str(v_i,v_j)=\frac{|\{\rho|\rho=1,\cdots,n\wedge\rho\neq i,j;nb_i^{\rho}\neq nb_j^{\rho}\}|}{n-2}$$

其表示在社交网络中两个节点之间在结构即社交关系上的距离，距离越大则表示两个节点在社交关系上的相似度越低，在同一个小团体的概率越低；距离越小则表示两个节点的相似度越大，其越应该被聚类在同一个超级节点中。

节点到簇的距离表示一个节点与一个超级节点之间的距离，距离越小则表示该节点与该超级节点的相似度越高，其越应该被划分在该簇中。计算方法为：

$$dist\_str(v,clt)=\frac{\sum_{v_i\in clt} dist(v,v_i)}{|clt|}$$

定义 9(个人信息特征距离(Personal Information Feature Distance))　在社交网络 $G(V,E,A)$ 中，$A_i=(a_{i1},a_{i2},\cdots,a_{im})$ 是节点 $v_i(i=1,2,\cdots,n)$ 的 m 维属性序列。对于每个节点，在个人属性信息上的距离为所有属性的距离和。本文定义：1)连续数值型属性信息损失的计算方法为两者之差；2)离散型数据的信息损失计算方法为当两属性相等时距离为 0，两者不相等时距离为 1。

由于在社交网络中，每个节点的不同属性对节点之间计算距离的影响不同，两个节点之间的距离越大表示相似度越低，其在一个小团体的可能性就越低。比如，相同地域或相似年龄的人更容易建立一个小团体；而具体的工作日期或者毕业院校对节点之间距离的影响则不如地域属性的影响大。因此，在计算节点之间的个人信息特征距离时，应乘以该属性的权重系数，权重系数越大，表示该属性的影响越大。

定义 10(网络节点特征综合距离(Comprehensive Distance，CD))　通过参数 α 将节点的结构信息特征距离和个人信息特征距离结合为综合距离来测量两个节点之间的距离，在聚类算法中的距离越近，节点越应该聚合在一个超点中。

$$CD=\alpha\times SIFD+(1-\alpha)\times PIFD$$

4.2　初始种子节点的优化算法

k-means 算法是一种被广泛研究和应用的聚类算法，有着理论可靠且算法简单的优点。因此，很多算法将该算法应用于社交网络中进行聚类，以达到隐私保护的目的。需要注意的是，将 k-means 算法应用在其他领域时，应结合相应的特点才能达到良好的聚类效果。

k-means 算法对初始聚类中心敏感，大部分算法在使用 k-means 时，结合社交网络的特点，使用度大的节点作为初始节点，一个节点的度越大，代表该用户节点的好友越多，同时其周围节点较为密集。本文提出基于聚类系数和密度的计算方法来选取初始中心。结合上文提到的局部聚类系数和用户节点密度来表示用户节点的周围节点用户的聚集程度，聚集程度越高，将其选为初始中心进行聚类的效果将越好。聚集密度的公式如下：

$$clt\_dens(v_i)=1+cfs(v_i)\times dens(v_i)$$

根据聚集密度对点集中的所有点进行排序，取聚集密度最大的点作为第一个初始中心。同时，相距较远的种子节点更具有代表性，且可以避免将同一个簇的节点选为种子节点，因此在高聚集密度的区域选取相距较远的点作为种子节点。初始节点的优化算法如算法 1 所示。

算法 1 初始节点的优化算法

function init_seed(G,cluster_num)

1. //获取高聚集密度区域,得到集合 $High_{clt\_dens}$;
2. 选取聚集密度最大的节点作为第一个初始中心节点 $seed_1$;
3. 在集合 $High_{clt\_dens}$ 中选取距离 $seed_1$ 最远的点形成一个集合,在该集合中选取密度最大的点作为 $seed_2$;
4. 以此类推,初始种子节点的计算方法如下:

$$seed_i=\max(\max(\sum_{j=1}^{i-1}dist(v_k,seed_j))),v_k\in High_{clt\_dens}$$

现有的基于聚类的社交网络隐私保护算法的隐私保护力度均为 k,因此初始节点的数目就是$\lfloor n/k \rfloor$。

在本文提出的算法中,节点的隐私保护力度不同,即每个簇中的节点数目不同,因此初始节点的数目是在一个范围内而不是固定的取值:$\lfloor n/high\_k \rfloor \leqslant cluster\_num \leqslant \lfloor n/k \rfloor$。

4.3 隐私保护力度的自适应

众所周知,在社交网络中,不同节点对隐私保护力度的需求不同,因此本文提出了核心点和非核心点的概念来区分使用不同的隐私保护力度。需要高隐私保护力度的点称为核心点,其余的点称为非核心点。这样,一方面提高了需要加强隐私保护力度的节点的安全性;另一方面,减少了不需要高隐私保护力度的节点信息损失,提高了数据的有效性。

定义 11(核心节点(Kernel Node)) 对于节点 v_i,在一定范围(ζ)内,节点数据大于或等于 χ,则该节点为核心节点。核心节点的公式化表示为:

$$v_i,\text{if } |\{v_j,Dist(v_i,v_j)\leqslant\zeta\}|\geqslant\chi,j\neq i$$

核心节点表示该节点密度大,在一定范围内节点密集,容易被恶意攻击者根据背景知识进行识别攻击,因此需要加大隐私保护力度。同时,假设隐私保护力度为 k,由于核心点密集程度高,在首次聚类后,簇中节点数大于 k;而非核心节点在首次聚类后,簇中节点可能会不足 k。一般的聚类算法会对多余节点进行再分配,将多余节点分配至不足 k 的簇里,这增加了数据损失,并且降低了数据的隐私保护力度和安全性。本文提出的算法在保证非核心节点最基本隐私保护力度的基础上,对核心节点进行了高力度的保护,提高了数据的安全性,减少了数据信息损失。

4.4 基于聚类的社交网络隐私保护算法

本文提出的算法如算法 2 所示。

算法 2 基于聚类的社交网络隐私保护算法

输入:社交网络模型 G=(V,E,A)

输出:k-匿名图 $G^*=(V^*,E^*,A^*)$,$V^*=\{cluster_1,cluster_2,\cdots,cluster_m\}$

1. Seed=init_seed(G,cluster_num)//根据上文中的 init_seed 算法分配初始种子节点
2. Dist=calculate(G,α)//根据定义 9 计算距离矩阵
3. //计算核心点集

   kernel=$\{v_p||Dist(v_p,v_i)\leqslant r|\geqslant min\_node,v_i\in V,i\neq j\}$
4. //进行首次聚类

   for(i=0;i<cluster_num;i++)

   $cluster_i=cluster_i\cup\{Seed(i)\}$

   $V=V-Seed(i)$

   $V^*=V^*\cup\{cluster_i\}$

   end//将种子节点分配至每个簇中

   while(|V|≥0)

   {//将未分配节点逐一分配至距离最近的簇中

   $cluster_i=cluster_i\cup\{v_p\}$ where $\{Dist(v_p,cluster_i)=\min_{v_j\in V}Dist(v_j,cluster_i)\}$

   }
5. for (i=0;i<cluster_num;i++)

   {

   $v_j\in kernel?\omega=1:\omega=0,v_j\in cluster_i$

   $k=k+k\times\omega$

   $level_i=k$

   }//计算每个簇的隐私保护力度
6. //根据隐私保护力度做再分配

   for(i=0;i<cluster_num;i++)

{

   $V=V\cup\{v_j\},\forall v_j\in cluster_i\&|cluster_i|\geqslant level_i\&\max(Dist(v_j,cluster_i))$

   $cluster_i=cluster_i-v_p$

   }//构造待分配节点集合

   while($|V|\geqslant0$ & $|cluster_i|\leqslant level_i,\forall cluster_i\in V^*$)

   {

   $cluster_i=cluster_i\cup\{v_p\}$ where $\{Dist(v_p,cluster_i)=\min_{v_j\in V}Dist(v_j,cluster_i)\}$

   }//将待分配节点分配至未满足对应隐私保护力度的簇中
7. for each $cluster_i$ of V^* anonymizingc $luster_i$

4.5 算法分析

本节主要从 3 方面对算法进行分析:时间复杂度、安全性和正确性。

本文算法的主要工作集中在步骤 1、步骤 2、步骤 5、步骤 6 和步骤 7 中。步骤 1 是获取种子节点集,计算所有节点的聚集密度,并在高聚集密度的集合里生成种子节点集,其外层循环为簇的个数,内层循环为高聚集密度集合,因此步骤 1 的时间复杂度为 $O(n^2)$。步骤 2 为计算距离矩阵,根据定义 10 求解一次综合距离可以在 O(1)时间里完成,因此步骤 2 的时间复杂度为 $O(n^2)$。步骤 5 是将未分配节点逐一分配,外层循环为未分配节点数目,内层循环为簇的个数 $cluster\_num$,可以得出 $cluster\_num\leqslant n/k$,因此步骤 5 的时间复杂度为 $O(n^2)$。同理,步骤 6 是将待分配节点分配至未满足对应隐私保护力度的簇中,待分配节点数目小于或等于 n,未满足簇数目小于或等于 $cluster\_num$,分配方法相同,因此步骤 6 的时间复杂度为 $O(n^2)$。步骤 7 是对各个簇中的节点进行匿名化处理,已有文献表明属性概化工作的时间复杂度为 $O(n^2)$。因此,本文提出的算法的复杂度为 $O(n^2)$。

本文提出了基于聚类的隐私保护算法。该算法结合社交网络中节点的特点计算综合距离,将网络中的节点聚合成超级节点,以超级节点内的节点个数和边数的统计信息代替子图结构,将节点和子图结构隐匿在超级节点中,体现了隐匿思想。因此,该算法可以有效地抵御基于结构的攻击,如基于度信息和子图结构的攻击。同时,该算法将超级节点中各个节点的属性信息泛化为一个概化值,体现了泛化的思想,这使得该匿名模型可以有效地抵抗以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。同时,考虑到社交网络各个节点对隐私保护力度的不同需求,提出了基于自适应调节的隐私保护力度。本文算法对社交网络中的所有节点有一个基

础的隐私保护力度，根据节点密度的稠密程度可将节点分为核心节点和非核心节点，对核心节点加大隐私保护力度，即 $k=k+k\times\omega$，非核心节点的隐私保护力度为 k。可以看出，所有节点的隐私保护力度均大于或等于 k。综上所述，本文算法结合了隐私保护技术中的隐匿和泛化等技术，结合了个性化和自适应的思想，能够有效地防范针对社交网络中以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，因此具有很强的隐私保护效果。

算法的正确性不仅体现在匿名模型的安全性，还体现在匿名模型的数据有效性。数据的安全性越高，意味着数据的隐匿程度越大。在本文提出的算法中，每个超点中的节点数目越多，则 k 越大，超点中每个节点被识别的概率 $1/k$ 就越小，同时数据被隐匿的信息越多，信息损失越大。首先，本文算法在进行聚类时遵循距离最小原则，节点每次进行聚类时都选择距离自己最小的簇，即相似度最大的簇，簇内节点越相似。根据定义 3，节点越相似，概化时信息损失越小。同时，节点间结构距离越近，则社交关系越相似，连接紧凑的点更容易凑成一类，根据定义 6，聚类带来的结构信息损失也越小。综上，所提算法符合信息损失量最小原则。其次，本文对核心点与非核心点区别了隐私保护力度，根据每个节点的个性化决定其隐私保护力度，对非核心节点选取与核心点不同的隐私保护力度，减少了再分配时造成的信息损失。综上，本文基于聚类的匿名方法可以保证总体信息损失趋于最小，数据有效性较高。

5　实验与结果分析

本文采用了两组数据集进行实验，从斯坦福大规模网络数据集大全（Stanford Large Network Dataset Collection，SNAP）和来自 UC Irvine Machine Learning Repository 的 Adult 数据集中任意取 300 个、400 个节点构成社交网络。实验环境为 Intel(R)Core(TM)i5，4.0 GB(1.6 GHz)内存。

本文从数据有效性和算法运行时间两方面来分析算法性能。对数据有效性的评估重点在于该算法对原始社交网络匿名后造成的信息损失。SaNGreeA 算法是基于贪心的典型聚类匿名算法，CAA-VS 算法[13]是基于节点连接结构和属性值的属性图聚类匿名化方法，它们与本文提出的聚类匿名算法（Adaptive-Dens-Cfs）在聚类功能、信息泛化损失的计算方法、匿名化目标上有一定的相似性和可比性，因此选择它们进行实验对比和分析。

如图 4、图 5 所示，在相同的隐私保护力度下（k 不变），本文提出的算法有更高的数据可用性，明显减少了信息损失，提高了数据有效性。

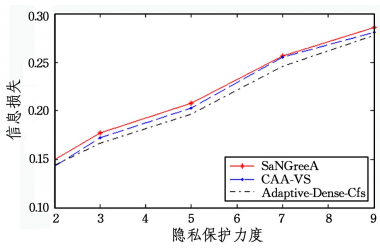

图 4　DB1 下 3 种算法的信息损失对比

Fig. 4　Comparison of information loss for three algorithms in DB1

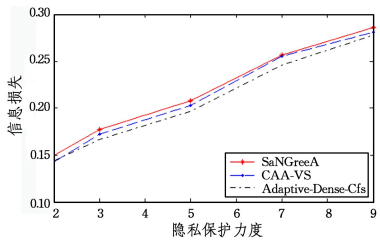

图 5　DB2 下 3 种算法的信息损失对比

Fig. 5　Comparison of information loss for three algorithms in DB2

从图 4、图 5 可以看出，随着 k 的增长，信息损失也在增长，这是因为随着 k 的增大，簇内隐匿泛化的节点数目变多，意味着更多的边被隐匿。同时，由于节点增多，各个属性的值域变大，隐匿时泛化程度变大，从而导致信息损失变大。当 k 很小时，Adaptive-Dens-Cfs 算法并没有明显的优势。这是由于当 k 很小（如 $k=1$）时，所有算法的信息损失为 0，但是由于加入隐私力度自适应的思路，对于核心点所在的簇，其中的节点数多于 1，造成了一定的信息损失。因此 k 很小时，由于 Adaptive-Dens-Cfs 算法对部分点进行了更高的隐私保护，导致其有效性有限。但是随着 k 的不断增长，可以看出，该算法降低了信息损失，提高了数据有效性。

如图 6、图 7 所示，在相同隐私保护力度（k 相同）下，Adaptive-Dens-Cfs 算法所用时间更少，有着更高的时间效率。随着 k 的增大，Adaptive-Dens-Cfs 算法所用时间基本不变，因此 k 值的变化对运行效率影响不大。

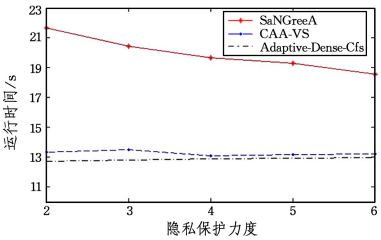

图 6　DB1 下 3 种算法的运行时间对比

Fig. 6　Comparison of running time for three algorithms in DB1

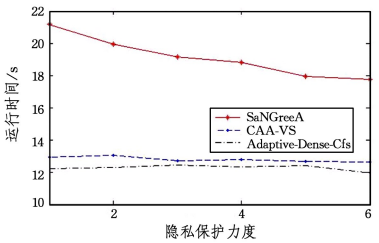

图 7　DB2 下 3 种算法的运行时间对比

Fig. 7　Comparison of running time for three algorithms in DB2

结束语　本文使用带标签的无向无权图来模拟社交网络，图中的节点和边代表用户的个人信息和社交关系等隐私信息。本文提出的隐私保护算法可应用于分析和挖掘社交网络本质特征的场景中，在对社交网络进行分析和挖掘时需要保证在宏观上的准确性，局部的不准确性不影响整个宏观性质。该隐私保护算法基于 k-means 聚类，结合社交网络的特性进行改进，将自适应与隐私保护力度相结合，同时考虑了用户的个人信息和社交关系，通过参数将其结合为综合距离来

计算用户节点特征的综合距离,将社交网络中的每个节点聚合成超点,最后进行了属性泛化,将用户节点隐匿在超点中,完成了对社交网络的隐私保护,量化并计算了社交网络聚类匿名过程中的信息损失。经过实验证明,该算法在降低信息损失和聚类效果方面具有有效性。

未来的研究方向如下:1)由于现实生活中的社交网络无时无刻不在更新,如何对社交网络中的数据进行实时隐私保护,在保证隐私保护力度的前提下减少时间开销是值得进一步研究的问题;2)本文提出的隐私保护算法重点考虑了在进行隐私保护的过程中减少数据的信息损失,下一步将研究如何权衡隐私保护力度和数据有效性。

参 考 文 献

[1] LIU X Y,WANG B,YANG X C. Survey on Privacy Preserving Techniques for Publishing Social Network Data[J]. Journal of Software,2014,25(3):576-590.

[2] LIU K,TERZI E. Towards identity anonymization on graphs [C]// ACM SIGMOD International Conference on Management of Data. ACM,2008:93-106.

[3] LIU K,DAS K,GRANDISON T,et al. Privacy-Preserving Data Analysis on Graphs and Social Networks[OL]. https://www.researchgate.net/publication/290110049 _ Privacy-Preserving _ Data_Analysis_on_Graphs_and_Social_Networks.

[4] CEN T T,HAN J M,WANG J Y,et al. Survey of K-anonymity research on privacy preservation[J]. Computer Engineering & Applications,2008,44(4):130-134. (in Chinese)
岑婷婷,韩建民,王基一,等. 隐私保护中 K-匿名模型的综述[J]. 计算机工程与应用,2008,44(4):130-134.

[5] WANG Z B,SHEN M Y,ZHAO J. Privay protection technology discussion of social network based on node reconstruction[J]. Computer Engineering and Applications,2017,53(11):131-136. (in Chinese)
王正彬,沈明玉,赵皎. 基于节点重构的社交网络的隐私保护技术探讨[J]. 计算机工程与应用,2017,53(11):131-136.

[6] XIE Y,ZHENG M. A Differentiated Anonymity Algorithm for Social Network Privacy Preservation [J]. Algorithms, 2016, 9(4):85.

[7] ZHELEVA E,GETOOR L. Preserving the Privacy of Sensitive Relationships in Graph Data[C]// ACM SIGKDD International Conference on Privacy, Security, and Trust in Kdd. Springer-Verlag,2008:153-171.

[8] ZOU L,CHEN L,ÖZSU M T. k-automorphism:a general framework for privacy preserving network publication [M]. VLDB Endowment,2009.

[9] GU Y H,LIN J C,GUO D . Clusterin-based bynamic privacy preserving method for social networks[J]. Journal on Communications,2015,36(S1),126-130. (in Chinese)
谷勇浩,林九川,郭达. 基于聚类的动态社交网络隐私保护方法[J]. 通信学报,2015,36(S1):126-130.

[10] CAMPAN A,TRUTA T M. Data and Structural k-Anonymity in Social Networks[M]// Privacy, Security, and Trust in KDD. Springer Berlin Heidelberg,2008:33-54.

[11] SKARKALA M E,MARAGOUDAKIS M,GRITZALIS S,et al. Privacy Preservation by k-Anonymization of Weighted Social Networks[C]// IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2012: 423-428.

[12] LAN L H ,JU S G,JIN H. Social Networks Data Publication Based on k-anonymity[J]. Computer Science,2011,38(11):156-160. (in Chinese)
兰丽辉,鞠时光,金华. 社会网络数据的 k-匿名发布[J]. 计算机科学,2011,38(11):156-160.

[13] JIANG H W,ZHAN Q H,LIU W J,et al. Clustering-Anonymity Approach for Privacy Preservation of Graph Data-Publishing [J]. Journal of Software,2017,28(9):2323-2333. (in Chinese)
姜火文,占清华,刘文娟,等. 图数据发布隐私保护的聚类匿名方法[J]. 软件学报,2017,28(9):2323-2333.

[14] RONG H,MA T H,TANG M L,et al. A novel subgraph K^+-isomorphism method in social network based on graph similarity detection [J]. Soft Computing,2018,22(8):2583-2601.

[15] SAMANTHULA B K,LEI C,WEI J,et al. Privacy-Preserving and Efficient Friend Recommendation in Online Social Networks [J]. Transactions on Data Privacy,2015,8(2):141-171.

[16] LIU P,BAI Y,WANG L,et al. Partial k-Anonymity for Privacy-Preserving Social Network Data Publishing [J]. International Journal of Software Engineering & Knowledge Engineering, 2016,27(1):71-90.

[17] YU F H,CHEN M J,YU B L,et al. Privacy preservation based on clustering perturbation algorithm for social network[J]. Multimedia Tools & Applications,2018,77(9):1-18.

[18] QIAN J W,LI X Y,ZHANG C H,et al. Social Network De-Anonymization and Privacy Inference with Knowledge Graph Model[J]. IEEE Transactions on Dependable and Secure Computing,2017,PP(99):1-1.