doi:10.11959/j.issn.1000-436x.2016116

# 基于信息损失量估计的匿名图构造方法

苏洁，刘帅，罗智勇，孙广路
（哈尔滨理工大学计算机科学与技术学院，黑龙江 哈尔滨 150080）

**摘 要**：首先分析了在进化的社会网络序列中，攻击者利用节点度信息，通过识别目标节点的方法对局部社会网络进行攻击过程，分析了利用 *k* 匿名方法对该类攻击进行隐私保护时存在的信息损失问题，针对该问题，提出了一种基于信息损失量估计的 *k* 匿名图流构造方法，通过子图节点属性泛化、子图内部结构的泛化控制图重构的信息损失，通过禁止子图内部扰动阻止网络攻击。定义匿名过程中由于图重构造成的节点和结构信息损失的估算方法，建立了基于贪婪聚类算法的网络节点的 *k* 匿名聚类算法，根据信息损失估计实现匿名分组，在进化的社会网络中以最小信息损失量构造匿名社会网络，在医疗诊断数据集上的实验表明所提方法能够较理想地控制信息损失量。
**关键词**：社会网络；隐私保护；*k* 匿名；信息损失估计
**中图分类号**：TP309.2 **文献标识码**：A

# Method of constructing an anonymous graph based on information loss estimation

SU Jie, LIU Shuai, LUO Zhi-yong, SUN Guang-lu
（School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China）

**Abstract:** A potential attack based on degree information by re-identifying target vertexes from a sequence of published graphs was analyzed. To deal with this kind of attack, a *k*-anonymous graph stream constructing method based on information loss estimation was provided. Information loss caused by re-constructing graph was controlled by using the method of attributes generalization of nodes and the structure generalization of sub-graph. The disturbance in sub-graph was forbidden to prevent the attack. The method of measuring the information loss of nodes and structures during the anonymous process due to re-construction of graph was defined. A *k*-anonymity cluster algorithm based on greedy clustering algorithm was build, which realized anonymous partition according to the information loss. Finally, a method of constructing anonymous social network for the evolving social network with the least information loss was provided. The experiments on medical diagnostic data set show that the algorithm of constructing anonymous graph based on the information loss estimation can be used to control the loss of information.
**Key words:** social network, privacy protection, *k*-anonymity, information loss estimation

## 1 引言

随着社会网络分析方法在各个社会研究领域中的广泛应用，越来越多的研究人员开始关注社会网络相关问题[1]，其中，社会网络的隐私保护成为该研究领域的关键问题之一。发布社会网络时，需要保护私人的敏感信息和社会关系，而社会网络的攻击方试图通过数据挖掘等技术发现社会网络中的敏感信息。社会网络通常以图的形式发布，网络中的节点表示个体，边表示个体间的关系。在社会

**收稿日期**：2015-11-12；**修回日期**：2016-04-26
**基金项目**：黑龙江省自然科学基金资助项目（No.A201301）；黑龙江省教育科学规划课题基金资助项目(No.GBC1211062)；黑龙江省普通高等学校新世纪优秀人才培养计划基金资助项目(No.1155-ncet-008)；黑龙江省博士后基金资助项目（No.LBH-Z12082）
**Foundation Items:** The Natural Science Foundation of Heilongjiang Province(No.A201301), Scientific Planning Issues of Education in Heilongjiang Province(No.GBC1211062), Research Fund for the Program of New Century Excellent Talents in Heilongjiang Provincial University (No.1155-ncet-008), Post Doctoral Fund of Heilongjiang Province(No.LBH-Z12082)

网络图中，每个节点由实体—属性集合描述，有唯一的标识符，由于对社会网络的研究可以利用图工具实现，越来越多的研究人员通过研究图匿名方法来解决隐私保护问题。文献[2]将匿名方法进行了分类，分析了图的匿名方法，指出由于动态社会网络需要定期发布网络数据来支持动态分析，因此会造成信息的泄露。文献[3, 4]提出了基于群和分类的匿名图，文献[5]提出了一种社会网络中数据和结构化匿名的聚类方法，文献[6]介绍了在图中怎样保护敏感关系，文献[7]提出一种基于差分隐私模型的随机扰动方法，实现边及边权重的强保护，文献[8]验证了匿名图中节点的再识别问题，文献[9]提出了通过发布和分析合成图的方法来保护社会网络中的个人社会关系，文献[10]提出一种在共享有意义的图形数据集的同时保护个人隐私的解决方案，文献[11, 12]研究了进化社会网络中的匿名图问题。然而，隐私保护技术仍然处于研究的初级阶段，网络的攻击方仍然能够在发布的匿名图中根据背景知识找到社会网络中感兴趣的个体和相关信息，文献[13, 14]证明现有的图匿名方法并未取得 $k$ 匿名方法的理想结果。

现有的图的匿名方法分为 3 类：1）基于 $k$ 匿名的方法，通过调整图的结构保护敏感信息[15, 16]，采用 $k$ 匿名方法，网络节点无法识别子图内的 $k$–1 个节点；2）基于概率的方法，通过随机添加/删除边或切换边的方法保护敏感信息[17]；3）基于泛化的方法，通过隐藏个人细节信息的隐私保护方法[4,5]。在社会网络发布过程中，通过更换节点的识别信息或者通过增加/删减边来改变结构信息，实现社会网络隐私保护。由于存在大量可获取的历史发布数据和节点度信息，社会网络攻击者会在某一时刻插入一个目标节点，在发布网络序列中利用背景信息识别该目标节点，实现网络攻击。针对该类攻击的匿名方法包括：1）采用节点度泛化的匿名方法，针对攻击者利用指定个体社会关系的先验知识对网络进行的攻击，通过插入或删除边的方法实现基于度匿名图的重构，使每个节点至少与 $k$–1 个节点有相同的度；2）采用 $k$ 邻域匿名方法，利用贪婪图调整算法生成节点标签，插入边，使每个邻接节点能够区分 $k$–1 个节点，该方法避免了攻击者根据已知目标节点的邻接子图进行的网络攻击；3）利用 $k$ 子图同构的匿名方法，重构图至少包含 $k$ 个子图的同构子图，避免攻击者通过识别指定个体的任意子图进行的网络攻击。利用此类方法保护社会网络需要重构社会网络图，在此过程中产生的信息损失既包括节点属性信息损失，又包括结构信息损失。

在分析 $k$ 匿名方法的基础上，针对社会网络发布过程中潜在的安全问题及匿名过程中的信息损失问题，本文利用匿名图工具，提出了在进化的社会网络中通过信息损失估计的方法，利用边的泛化构造 $k$ 匿名图。本文创新之处如下。

1) 在社会网络发布过程中，利用信息损失估计方法构建 $k$ 匿名子图，在节点信息损失、结构信息损失和社会网络安全级别方面取得折衷的最优值。

2) 利用节点信息、子图结构信息泛化构建的匿名子图，避免了扰动攻击对网络安全的威胁。

3) 在发布的社会网络中，通过判断网络结构图的变化选择构建子图方法，提出以极小的信息损失代价平衡网络子图构建的时间复杂性的方法，最大程度保证动态网络的稳定性。

## 2　基于节点度的社会网络攻击

社会网络图被定义为 $G(V,E)$，$V$ 是节点集合，表示社会网络中的个体，$E$ 是边的集合，表示个体之间的关系。$G$ 用邻接矩阵 $\boldsymbol{M}=(m_{ij})_{n\times n}$ 表示, $n$ 为节点数。若节点 $i$ 和节点 $j$ 相邻接，则 $m_{ij}=1$，否则 $m_{ij}=0$。节点 $i$ 的度表示为 $d_i$，$d_i=\sum_j m_{ij}$，$j=1,2,\mathrm{L}\ ,n$。匿名的发布图表示为 $G'(V',E')$，用邻接矩阵 $\boldsymbol{M}'=(m_{ij}')_{n\times n}$ 表示。节点 $i$ 的度在发布的社会网图中表示为 $d_i'$，当社会网络随时间进化时，任意时刻 $t$ 的社会网络图和发布的社会网络图分别表示为 $G_t$ 和 $G_t'$, $t=0,1,2,\mathrm{L}\ ,n$。在社会网络图中，关键信息包括节点属性、表示节点间关系的边属性、节点度、目标个体的邻接关系、子图等，其中标识符不会随着时间改变。图 1 中用来描述节点的属性信息分成 3 类[12]：1）唯一标识信息，例如身份证号、驾驶证号、社会安全保障号 SSN 等，该类属性在社会网络发布前已经被隐藏；2）近似标识信息，例如邮政编码、家庭住址等；3）敏感属性信息，如家庭、收入等。社会网络的攻击方能够访问到 $G_0',G_1',\mathrm{L}\ ,G_n'$ 的部分网络数据，利用节点度信息，通过查询节点相关匹配信息的候选集来识别插入在图 $G_0$ 中的目标节点，该过程如图 1 所示。

对图 $G$ 采用边扰动的方法得到图 $G'$，由文献[18]得到利用邻接背景知识识别图 $G'$ 中的目标节点的概

率小于$\frac{1}{k}$。为了保护个体间的联系，采用连接扰动方法，向图中随机添加和删除相等数量的边，通过向图中随机添加噪声的方法隐藏链接信息。社会网络进化过程中，社会网络的攻击方根据节点的背景知识，通过分析节点的拓扑特征识别目标节点。

图 1 中，$G_0$ 为初始社会网络图，序列 $G_0$,L , $G_i$,L , $G_j$ 表示社会网络随时间的进化过程，序列 $G_0'$,L ,$G_i'$,L ,$G_j'$ 是匿名发布社会网络，$G_0$ 通过连接扰动方法得到发布图 $G_0'$。该社会网络中，目标节点攻击方试图识别目标节点 $C$。假设已知节点 $C$ 的度，扰动律为 10%，攻击方推断 $G_0'$ 中目标节点的度为 [2,4]，查询图 $G_0'$，所有节点的度都在 2~4，候选目标节点序列为 $\{A, B, C, D, G\}$。在图 $G_i$ 中新插入 6 个节点 $E$、$F$、$I$、$J$、$K$ 和 $H$，为了在 $G_0'$ 中识别目标节点，攻击者用节点 $E$、$F$、$I$、$J$ 和 $K$ 连接节点 $C$。$G_i$ 中 $C$ 的度为 6，候选节点的度应为[5,7]，因此推断 $G_i'$ 的候选节点集合是 $\{C,D\}$。经过一段时间后，攻击者在 $G_j$ 中删除子网到节点 $C$ 的连接后，候选节点的度应该为 [2,4]，推断 $G_j'$ 中的目标节点的候选节点集合是 $\{C\}$。因此，攻击者有机会在社会网络的进化中，利用匿名图工具，通过识别目标节点的方法实现网络攻击。

上述分析表明，利用边扰动的方法实现动态社会网络匿名，攻击者可以利用收集到的节点信息实现局部网络攻击。虽然 Facebook、Twitter 等社会网络已经限定网络用户的访问范围，但是基于应用的需要，攻击者仍然能够利用上述方法攻击局部开放网络。

采用 $k$ 邻域匿名方法能够有效控制攻击者利用已知目标节点的邻接子图信息进行的网络攻击，但是构建 $k$ 匿名图的过程中会有大量信息损失，3.1 节中给出了利用贪婪图调整算法生成节点标签，通过构建信息损失量估计算法，预估计构建 $k$ 匿名子图的损失量，实现最小信息损失匿名图构造。

## 3 构建社会网络的匿名子图

### 3.1 基于泛化的 $k$ 匿名

$k$ 匿名是隐私保护的经典方法，每个数据组至少包含 $k$ 个无法区分的节点。传统方法通过插入或删除边的扰动方法保护节点不被识别，该匿名方法构造过程中会造成信息损失，影响数据的可信性。基于属性泛化的方法能够降低对原图结构的破坏，降低信息损失。

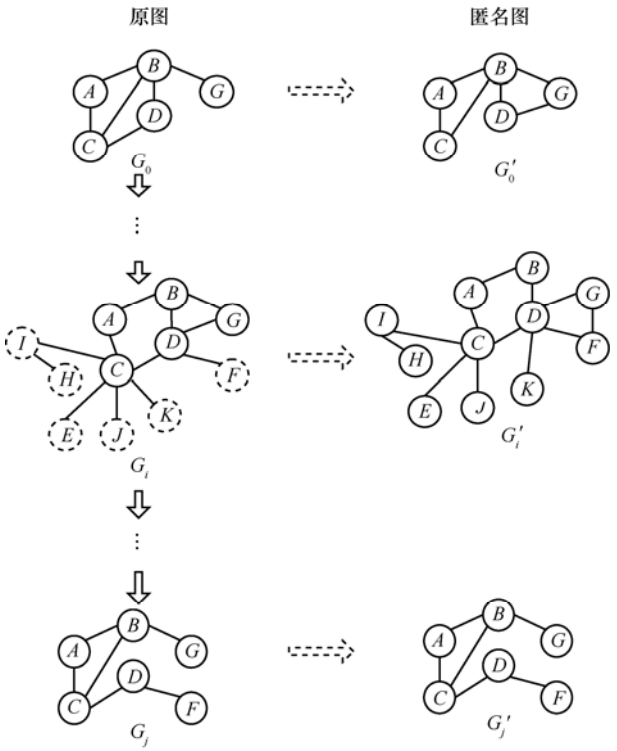



图 1 基于节点度的社会网络攻击

为了构建 $k$ 匿名子图，既要对节点信息泛化，也要对子图内部结构和子图间的联系泛化。表示子图之间的关系的边显示了网络的结构特征，以实现某些应用。匿名网络结构中，子图内部不允许使用扰动方法，有效防止了基于节点度的攻击。利用节点信息、子图内部关系和子图间的关系，通过估计信息损失，构造 $k$ 匿名图。

社会网络图表示为 $G(V,E)$，$|V|=N$，$N$ 为节点数，根据网络节点度信息，采用贪婪聚类算法对网络图进行节点 $k$ 匿名分组，使分组满足以下 2 个条件：

1) 每个分组至少包含 $k$ 个节点；

2) 估计聚类的信息损失，降低匿名过程的信息损失量。

因此，需要定义一种信息损失的估算方法。

### 3.2 基于信息损失估计的匿名图重构方法

基于信息损失估计的匿名图重构方法将具有相似属性且具有最小信息损失的 $k$ 个节点聚为一个集合，聚类分组过程中，用于损失估计的信息包括图重构信息和节点与分组的结构信息。

设 $V$ 表示有序节点序列 $\{v_0, v_1, \text{L}, v_N\}$，节点间的邻接关系表示为邻接矩阵 $\boldsymbol{A}=\{a_{i,j}\}$，$i=0,1,2,\text{L}$，$N, j=0,1,2,\text{L}$ ,$N$。$v_i$ 和 $v_j$ 直接连接时，$a_{i,j}=1$，否则 $a_{i,j}=0$。利用该矩阵检索邻接节点，任意节点

间的距离定义如下。

**定义 1**　对于任意的 $i, j = 0,1,2,\mathrm{L}\ ,N$，节点 $v_i$ 和 $v_j$ 间的距离定义为 $v_i$ 和 $v_j$ 间的最短距离

$$D(v_i,v_j)=\frac{\left|d\middle|d=a_{i,k}+a_{k,q}+\mathrm{L}\ +a_{p,j},i\neq k\neq q\neq\mathrm{L}\ \neq p\neq j\right|}{mn} \tag{1}$$

其中，$a_{i,k}=a_{k,q}=\mathrm{L}\ =a_{p,j}=1$，表示节点 $v_i$ 和 $v_j$ 之间的最短路径，$mn$ 是该最短路径上的节点数。任意节点 $v_i(v_i\notin s_k)$ 与聚类集合 $s_k$ 间的距离定义为

$$\forall v_j\in s_k, D(v_i,s_k)=\frac{\left(\sum_{v_j\in s_k}D(v_i,v_j)\right)}{\left|s_k\right|} \tag{2}$$

其中，节点间的距离、节点与聚类集合间的距离取值为[0, 1]。

在图 $G$ 中选择度最大的节点作为聚类集合的中心节点，选择 $k$−1 个与当前聚类集合有最小距离但未分配的节点来构造新的聚类集合。节点间的距离和结构距离分别表示为 $D(v_i,v_j)$ 和 $D(v_i,s_k)$。

根据节点属性计算聚类分组过程的信息损失包括泛化信息损失和结构信息损失[14]。泛化信息损失用于计算节点描述性信息的损失[20]，定义泛化信息损失为

$$GLoss(G,CL)=\frac{\sum_{j=1}^{m}\left(\left|s_j\right|(\mathrm{Attr}(s_j,N)+\mathrm{Cate}(s_j,C))\right)}{n(p+q)} \tag{3}$$

其中，$CL=\{s_1,s_2,\mathrm{L}\ ,s_m\}$ 是采用聚类方法得到的分组，$\left|s_j\right|$ 是 $s_j$ 的基，$N=\{N_1,N_2,\mathrm{L}\ ,N_p\}$ 和 $C=\{C_1,C_2,\mathrm{L}\ ,C_q\}$ 分别表示数字化属性和分类结构属性的集合，$s_j$ 的生成属性表示为 $\mathrm{gen}(s_j)$，生成信息损失因子分别表示为 $\mathrm{Attr}(s_j,N)$ 和 $\mathrm{Cate}(s_j,C)$。$\mathrm{Attr}(s_j,N)$ 定义如式(4)所示，$\mathrm{Cate}(s_j,C)$ 定义如式(5)所示。

$$\mathrm{Attr}\left(s_j,N\right)=\sum_{k=1}^{p}\frac{\mathrm{size}\left(\mathrm{gen}\left(s_j\right)\left[N_k\right]\right)}{\max_{X\in N}\left(X\left[N_k\right]\right)-\min_{X\in N}\left(X\left[N_k\right]\right)} \tag{4}$$

$$\mathrm{Cate}\left(s_j,C\right)=\sum_{k=1}^{q}\frac{\mathrm{height}(M(\mathrm{gen}(s_j)[C_k]))}{\mathrm{height}(H_{C_k})} \tag{5}$$

其中，$\mathrm{gen}(s_j)$ 包含数字和分类属性，定义为

$$\mathrm{size}(\mathrm{gen}(s_j)[N_k])=[\min\{X^1[N_k],\mathrm{L}\ ,X^u[N_k]\},\ \max\{X^1[N_k],\mathrm{L}\ ,X^u[N_k]\}] \tag{6}$$

与分类相关的层次属性定义为 $H_{C_k},\mathrm{gen}(s_j)[C_k]$ 定义为 $H_C$ 的最近祖先，满足式（7）

$$\mathrm{size}(\mathrm{gen}(s_j)[N_k])=\max\{X^1[N_k],L,X^u[N_k]\}-\min\{X^1[N_k],L,X^u[N_k]\} \tag{7}$$

其中，$M(\mathrm{gen}(s_j)[C_k])$ 是 $H_C$ 中以 $\mathrm{gen}(s_j)[C_k]$ 为根的子层，$\mathrm{height}(H_{C_k})$ 定义为 $H_C$ 的子层高度。

参数 $\alpha$ 和 $\beta$ 由用户设置，用来控制重构信息和距离信息的重要性。基于信息损失估计的匿名聚类算法如算法 1 所示。

**算法 1**　基于信息损失估计的匿名聚类算法

**输入**　图 $G$;
　　参数 $k$、参数$\alpha$和参数$\beta$;
　　初始聚类集合$S$=∅

**输出**　聚类分组 $CL=\{s_1,s_2,\mathrm{L}\ ,s_m\}$。

1) $m$=|$CL$|=0;　// 聚类分组数
2) $n$=|$V$|; //初始 $n$ 值
3) while( $n>0$ ) // $n$ 是尚未分配节点数
//遍历节点找出最大度节点作为聚类的种子节点
4)　Seed= $v_i$, $v_i$ 有当前最大度 $d_i$;
5)　$s_j$={ $v_i$ }; // $v_i$ 加入到分组 $j$ 中
6)　$V$=$V$−$v_i$;
7)　while(|$s_j$|<$k$)
8)　$MinLoss(\alpha GLoss(G_j,CL)+\beta D(v,s_j))$;
//查找最少损失节点 $v$，$GLoss(G_j,CL)$ 是 $v$ 加入分组造成的信息损失，$G_j$ 是由 $s_j$ 和 $v$ 生成的子图，$CL$ 是包含 $s_j$ 和 $v$ 的分组，$D(v,s_j)$ 是 $v$ 和分组 $s_j$ 的距离
9)　$s_j=s_j\cup v$;
10)　$n=n-\left|v\right|$;
11)　if $(n==0)$
12)　return;
13)　end if
14)　end while
15)　if(|$s_j$|<$k$)
16)　$\forall v\in s_j$,遍历分组 $s_p, p\neq j$，将 $v$ 加入最小损失分组 $s_p$ 中;
17)　$s_p=s_p\cup v$;
18)　else
19)　$CL=CL\cup\left\{s_j\right\}$
20)　$m$=$m$+1;
21)　end if

22) end while

社会网络在进化过程中，会有新用户加入，或者旧用户退出，在网络图中表现为插入新的节点和边或者删除某些节点和边，由此造成的社会网络结构的变化定期更新发布。更新时间间隔表示为 $\Delta t$，图流序列表示为 $G_0, G_1, \cdots, G_t$，$t$ 时刻的图结构变化定义为如式（8）所示。

$$ADJ(t)=\{\phi(V_t),\omega(E_t)\} \tag{8}$$

图 $G_t$ 中节点及边的结构变化定义如式（9）和式（10）所示。

$$\phi(V_t)=diff(V_{t-1},V_t) \tag{9}$$

$$\omega(E_t)=diff(E_{t-1},E_t) \tag{10}$$

其中，$\omega(E_t)$ 根据 $\{d(V)\}$ 定义，$\{d(V)\}$ 为节点度变化集合。用 $A(V_t)$ 表示 $t$ 时刻由边/节点增加导致的结构变化节点属性集合，用 $D(V_t)$ 表示 $t$ 时刻由边/节点删除导致的结构变化节点属性集合。$t$ 时刻图结构变化率定义为

$$\mathrm{R}(t)=\mu\mathrm{R}(V_t)+\pi\mathrm{R}(E_t) \tag{11}$$

节点的变化率定义为 $\mathrm{R}(V_t)\propto\phi(V_t)$，结构变化率定义为 $\mathrm{R}(E_t)\propto\omega(E_t)$，$\mathrm{R}(E_t)$ 定义为：$\mathrm{R}(E_t)=\dfrac{\omega(E_t)}{\sum_{i=1}^{n}d_i}$，$\mathrm{R}(V_t)$ 表示为 $\mathrm{R}(V_t)=\dfrac{\phi(V_t)}{|V_t|}$。在 3.4 节中定义了基于图的变化率的图流聚类算法。

### 3.3 匿名子图信息损失评价

利用基于信息损失估计的匿名聚类算法将社会网络图 $G$ 划分成分组集合 $CL=\{s_1,s_2,\mathrm{L},s_m\}$，结构信息损失 $SLoss(G,CL)$ 由类内结构损失和类间结构损失 2 部分组成[21]，定义如(12)所示。

$$SLoss(G,CL)=\frac{\sum_{j=1}^{m}\left(\mathrm{intraSL}(s_j)\right)+\sum_{i=1}^{m}\sum_{j=i+1}^{m}\left(\mathrm{interSL}(s_i,s_j)\right)}{\frac{n(n-1)}{4}} \tag{12}$$

其中，类内结构损失为 $\sum_{j=1}^{m}\left(\mathrm{intraSL}(s_j)\right)$，类间结构损失为 $\sum_{i=1}^{m}\sum_{j=i+1}^{m}\left(\mathrm{interSL}(s_i,s_j)\right)$，满足

$$\mathrm{intraSL}(s_j)=2\left|E_{s_j}\right|\left(1-\frac{\left|E_{s_j}\right|}{\binom{|s_j|}{2}}\right) \tag{13}$$

$$\mathrm{interSL}(s_i,s_j)=2\left|E_{s_i,s_j}\right|\left(1-\frac{\left|E_{s_i,s_j}\right|}{|s_i||s_j|}\right) \tag{14}$$

由式（14）可以推出，当 $\left|E_{s_i,s_j}\right|=\dfrac{(|s_i||s_j|)}{2}$ 时，$s_i$ 与 $s_j$ 的类间结构损失取得最大值 $|s_i||s_j|$。匿名图构建过程中的类内结构最大损失和类间结构最大损失定义如式(15)和式(16)所示。

$$\max\left(\sum_{j=1}^{m}\mathrm{intraSL}(s_j)\right)=\sum_{j=1}^{m}\left(\frac{|s_j|\left(|s_j|-1\right)}{4}\right)=\frac{1}{4}\sum_{j=1}^{m}|s_j|^2-\frac{1}{4}\left(\sum_{j=1}^{m}|s_j|\right) \tag{15}$$

$$\max\left(\sum_{i=1}^{m}\sum_{j=i+1}^{m}\left(\mathrm{interSL}(s_j,s_j)\right)\right)=\sum_{i=1}^{m}\sum_{j=i+1}^{m}\left(\frac{|s_i||s_j|}{4}\right) \tag{16}$$

### 3.4 基于图的变化率的图流聚类算法

对于初始的社会网络 $G$，采用聚类算法得到分组 $CL=\{s_1,s_2,\mathrm{L},s_m\}$ 后，聚类分组 $\{s_1,s_2,\mathrm{L},s_m\}$ 对应的节点核记为 $\{cl_1,cl_2,\mathrm{L},cl_m\}$，$cl_i=[\mathrm{gen}(s_i),(|s_i|,|E_{s_i}|)]$，$(|s_i|,|E_{s_i}|)$ 是类内生成对，$cl_i\cap cl_j=\varnothing$，$i,j=1,\mathrm{L},m,i\neq j$。匿名社会网定义为

$$G_m=\left(\{cl_1,cl_2,\mathrm{L},cl_m\},\{cl_1,cl_2,\mathrm{L},cl_m\}\times\{cl_1,cl_2,\mathrm{L},cl_m\}\right) \tag{17}$$

上述定义中，对于任意边 $e(v_k,v_p)$，其中，$v_k\in s_i$，$v_p\in s_j$，满足下条件 $(cl_i,cl_j)\in\{cl_1,cl_2,\mathrm{L},cl_m\}\times\{cl_1,cl_2,\mathrm{L},cl_m\}$。根据节点泛化信息和类内、类间结构信息损失估计构造 $k$ 匿名图。当满足条件 $|cl_1|\geqslant k$ 时，创建 $k$–匿名网络。社会网络发展过程中，基于图的变化率的图流聚类算法如算法 2 所示。

**算法 2** 基于图的变化率的图流聚类算法

**输入** $G_{t-1}, G_t$;

$CL_{t-1}=\{s_1,s_2,\mathrm{L},s_m\}$；// $CL_{t-1}$ 是 $t-1$ 时刻的分组 $\delta_v,\delta_E$

**输出** $cl_t=\{s_1',s_2',\mathrm{L},s_m'\}$

1) 读取 $ADJ(t)=\{\phi(V_t),\omega(E_t)\}$;

2) if $\mathrm{R}(V_t)<\delta_v$

3) if $\mathrm{R}(E_t)<\delta_E$

4) 结构变化的节点集合 $A(V)$;

5) for $\forall v\in A(V)$

6) 遍历分组，找到满足条件 $MinLoss(\alpha GLoss(G_j,cl_j)+\beta D(v,s_j))$ 的 $s_j$；// $GLoss(G_j,cl_j)$

是 $v$ 加入分组造成的信息损失，$G_j$ 是由 $s_j$ 和 $v$ 生成的子图，$cl_j$ 是包含 $s_j$ 和 $v$ 的分组，$D(v,s_j)$ 是 $v$ 和分组 $s_j$ 的距离；

7) end for
8) 结构变化的节点集合 $D(V)$；
9) for $\forall v \in D(V)$)
10) $v$ 所在分组 $s_i$
11) if($|s_i|$<$k$)
12) $\forall v \in s_i$，遍历分组找到最少损失分组 $s_p$，$p \neq j$，将 $v$ 加入 $s_p$
13) $s_p = s_p \cup v$
14) end if
15) end for
16) end if
17) end if
18) else
19) 网络结构变化明显，对整体网络匿名分组
20) end else

在社会网络更新过程中，采用基于图的变化率的图流聚类算法，计算图流的结构变化，通过估算最小信息损失量方法实现匿名聚类。

## 4 仿真实验

表 1 提供了用于急性髓细胞白血病诊断的病历数据。$t_0$ 时刻图 $G_0$ 的节点集合为 $\{v_0, v_1, \text{L}, v_{20}\}$。病患资料的个人信息中，SSN 和驾驶证等已经被隐藏，表中给出了年龄、性别、邮政编码、婚姻状况等近似标识符。为了保护病人的隐私，采用本文匿名方法，属性集定义为 $Al = \{N_1, C_1, C_2, C_3\}$，$P = 1, Q = 3$，$N_1 = \text{Age}$，$\{C_1, C_2, C_3\} = \{\text{Gender, Zip code, Marriage}\}$。

图 2 为根据诊断数据和社会关系构建的社会网络，

**表 1**　　**诊断的病历数据**

| ID | Gender | Age | Zip code | Marriage | Diagnosis | Blast | Category | Mutation | Cytogenetics |
|---|---|---|---|---|---|---|---|---|---|
| 1 | F | 56 | 33613 | Divorce | AA | 1 | AA | normal | DEL(7) |
| 2 | F | 60 | 33647 | Marriage | AA | 1 | AA | normal | Normal |
| 3 | M | 81 | 34660 | Single | AA | 1 | AA | normal | Normal |
| 4 | M | 34 | 32801 | Divorce | AA | 1 | AA | U2AF1 | Normal |
| 5 | M | 56 | 32211 | Marriage | AA | 1 | AA | U2AF1 | TRI(1Q) |
| 6 | M | 34 | 32868 | Marriage | AA | 1 | AA | normal | TRI(6) |
| 7 | F | 73 | 34768 | Single | AA | 1 | AA | normal | Normal |
| 8 | F | 77 | 33102 | Marriage | AA | 1 | NON | DNMT3A | Normal |
| 9 | F | 84 | 32855 | Single | AA | 1 | AA | normal | DEL(2) |
| 10 | F | 68 | 33709 | Marriage | ACML | 1 | MDS/MPN | ASXL1 | Normal |
| 11 | F | 66 | 34302 | Marriage | ACML | 1 | MDS/MPN | DNMT3A NPM1 TET2 | Normal |
| 12 | M | 73 | 34565 | Single | AML | 4 | AML/MDS | DNMT3A | Normal |
| 13 | M | 59 | 32652 | Marriage | AITCL | 1 | LYMPHOMA | ASXL1 | Normal |
| 14 | F | 63 | 33615 | Marriage | AML | 4 | AML | TP53 | DEL(20q) EL(5q31) |
| 15 | F | 19 | 75865 | Single | AML | 4 | AML | normal | Normal |
| 16 | M | 48 | 33650 | Single | AML | 4 | AML/MDS | normal | Normal |
| 17 | M | 76 | 75677 | Single | AML | 4 | AML | normal | DEL(5q) DEL(7q) |
| 18 | F | 77 | 33218 | Marriage | AML | 4 | AML | normal | Normal |
| 19 | M | 65 | 34813 | Marriage | AML | 4 | AML | ASXL1 | DEL(7) |
| 20 | M | 71 | 32556 | Marriage | AML | 4 | AML | normal | Normal |
| 21 | F | 67 | 33451 | Marriage | AML FROM MDS | 4 | AML/MDS | NEGATIVE | POSITIVE |
| 22 | M | 60 | 33648 | Marriage | AML WITH MDS | 4 | AML/MDS | FLT3 RUNX1 SF3B1 | Normal |

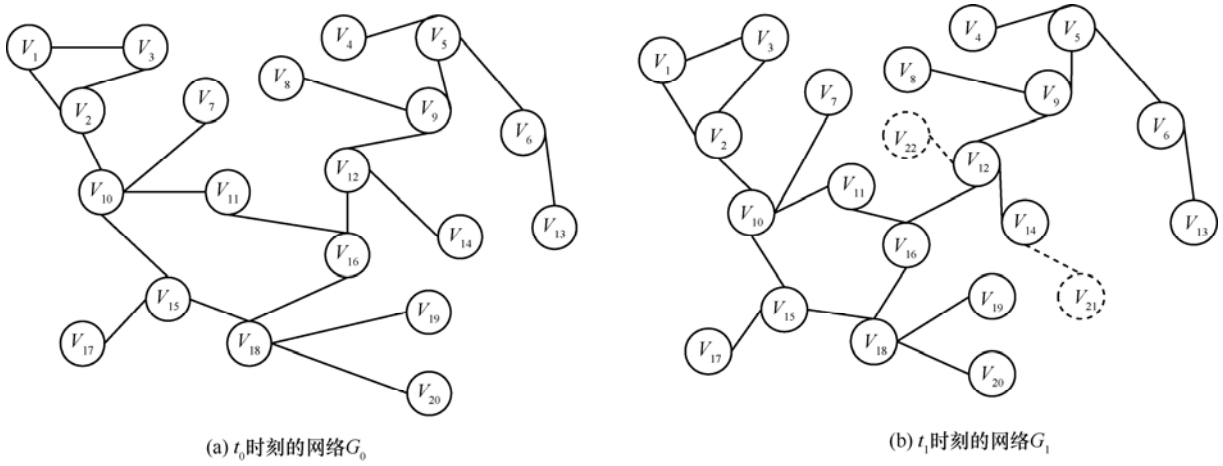



图 2　$t_0$ 和 $t_1$ 时刻的网络

$t_0$时刻的网络表示为图 $G_0$，$t_1$时刻的网络表示为图 $G_1$，层次结构属性 $C_1$、$C_2$、$C_3$，如图 3 所示。表 2 给出了 $k$ 取 3、6，$a$ 分别取 0、0.6、1 时的聚类分组结果。

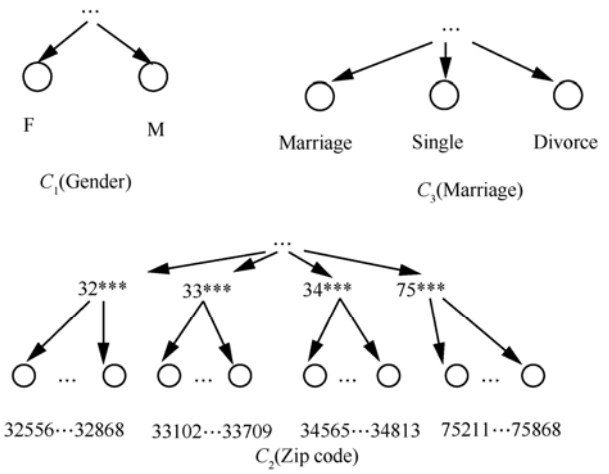



图 3　层次结构属性

图 4 给出了 $k$ 取 2~10，$a$ 取 0~1 时节点的泛化信息损失和结构信息损失情况。图 4 数据表明，最终发布的匿名数据集中包含的匿名组数目越多，$k$ 值越小，信息损失越少，匿名化的数据越接近真实数据，该数据集的可信任度越高。$a$=1 时的节点泛化信息损失高于 $a$=0 时的泛化信息损失；$a$=0 时的结构信息损失低于 $a$=1 时的结构信息损失。

在 $t_1$ 时刻插入新的节点如图 2(b)所示。$\delta_v=\delta_E=0.1$, $G_1$ 结构变化率 $\mathrm{R}(V_t)=0.090\,9<0.1$，$\mathrm{R}(E_t)=0.086\,9<0.1$，取 $k$=4，$a$ 分别为 0 和 1 时，采用算法 2 聚类分组结果为 $\mathrm{A}_1$，对 $G_1$ 采用基于完全信息算是估计的聚类分组结果为 $\mathrm{A}_2$，如表 3 所示，$\mathrm{A}_1$ 和 $\mathrm{A}_2$ 的聚类分组信息损失如图 5 所示。

聚类分组误差定义为

$$\delta(GLoss)=\frac{|\mathrm{A}_1(GLoss)-\mathrm{A}_2(GLoss)|}{\mathrm{A}_2(GLoss)} \tag{18}$$

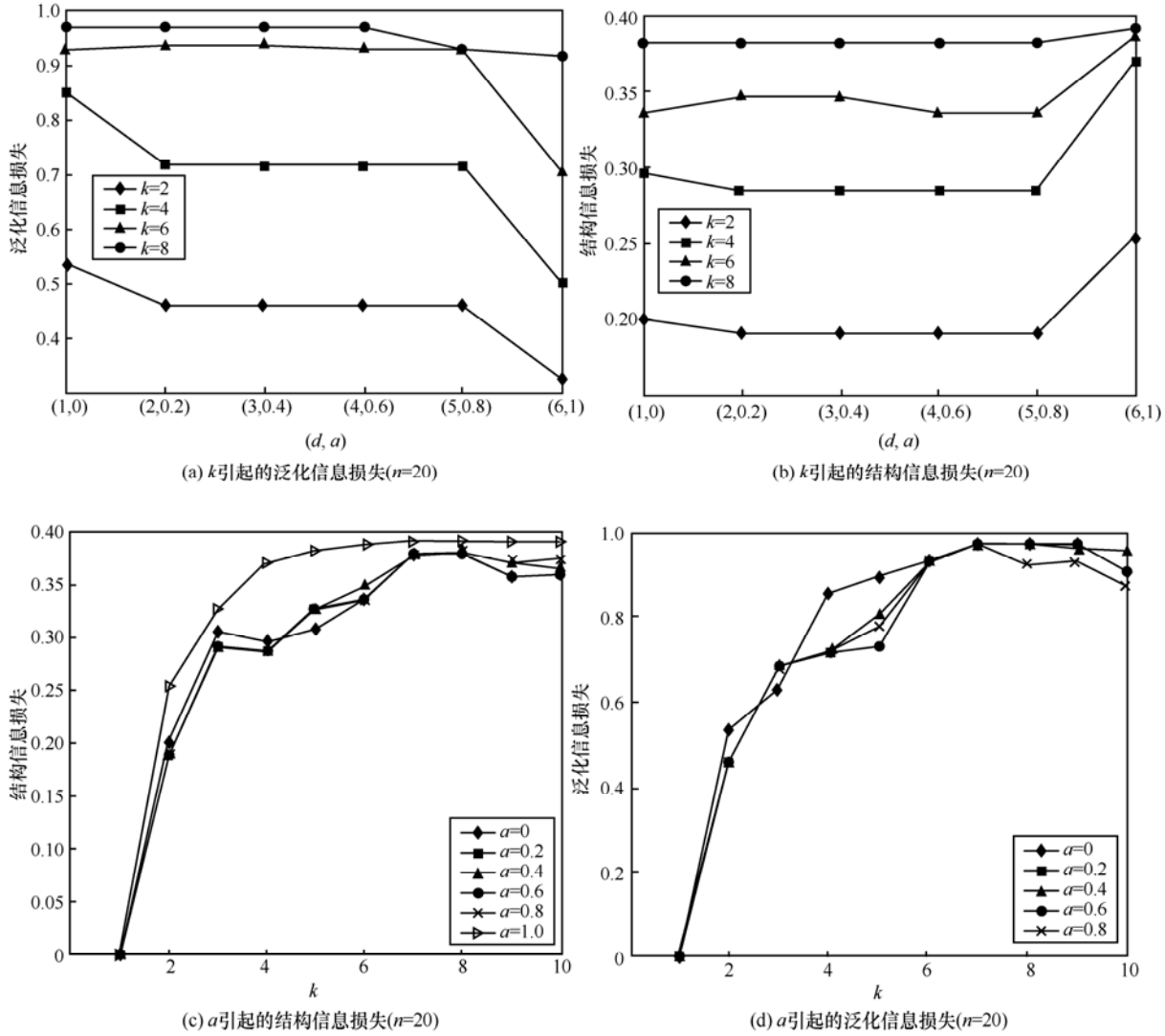
$$\delta(SLoss)=\frac{|\mathrm{A}_1(SLoss)-\mathrm{A}_2(SLoss)|}{\mathrm{A}_2(SLoss)} \tag{19}$$

$k=4$，$a$ 分别取 0、1 时，采用算法 2 的信息损失误差如图 6 所示。

**表 2**　　**基于信息损失估计的聚类分组**

| $k$ 值 | $a=1$ | $a=0.6$ | $a=0$ |
|---|---|---|---|
| $k$=3 | $S^1_{k=3}(G_0)=\left\{\begin{matrix}\{v_{10},v_{14},v_2,v_1,v_{15}\},\{v_{18},v_8,v_{11}\},\\ \{v_5,v_{13},v_6\},\{v_{12},v_3,v_{17}\},\\ \{v_{19},v_{16},v_4\},\{v_{20},v_9,v_7\}\end{matrix}\right\}$ | $S^{0.6}_{k=3}(G_0)=\left\{\begin{matrix}\{v_{10},v_2,v_{11}\},\{v_{18},v_{15},v_{16},v_4,v_{20}\},\\ \{v_5,v_6,v_{13}\},\{v_9,v_8,v_{12}\},\\ \{v_1,v_3,v_7\},\{v_{19},v_{14},v_{17}\}\end{matrix}\right\}$ | $S^0_{k=3}(G_0)=\left\{\begin{matrix}\{v_{10},v_2,v_1,v_{14}\},\{v_{18},v_{15},v_{16},v_{17}\},\\ \{v_5,v_4,v_6\},\{v_9,v_8,v_{12}\},\\ \{v_{19},v_3,v_7\},\{v_{20},v_{11},v_{13}\}\end{matrix}\right\}$ |
| $k$=6 | $S^1_{k=6}(G_0)=\left\{\begin{matrix}\{v_{10},v_{14},v_2,v_8,v_{18},v_{11},v_{19},v_{20}\},\\ \{v_5,v_{13},v_6,v_4,v_{16},v_1\},\\ \{v_9,v_7,v_3,v_{12},v_{17},v_{15},\}\end{matrix}\right\}$ | $S^{0.6}_{k=6}(G_0)=\left\{\begin{matrix}\{v_{10},v_2,v_{11},v_7,v_{15},v_{18},v_{19},v_{20}\},\\ \{v_5,v_6,v_{13},v_4,v_9,v_8\},\\ \{v_{12},v_{16},v_{14},v_1,v_3,v_{17}\}\end{matrix}\right\}$ | $S^0_{k=6}(G_0)=\left\{\begin{matrix}\{v_{10},v_2,v_1,v_3,v_7,v_{11}\},\\ \{v_{18},v_{15},v_{16},v_{12},v_9,v_{14},v_{19},v_{20}\},\\ \{v_5,v_4,v_6,v_{13},v_8,v_{17}\}\end{matrix}\right\}$ |

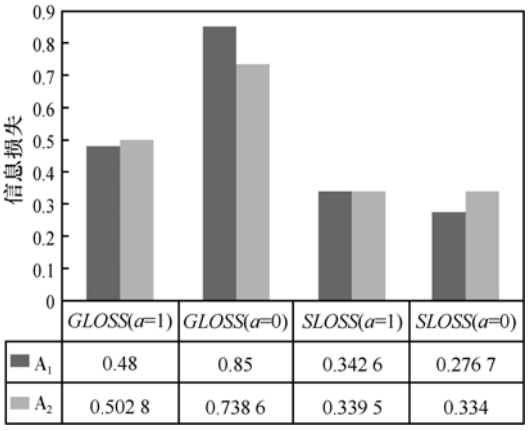
2016116-7




(a) $k$引起的泛化信息损失($n$=20)　(b) $k$引起的结构信息损失($n$=20)

(c) $a$引起的结构信息损失($n$=20)　(d) $a$引起的泛化信息损失($n$=20)

图 4　信息损失评估




| | GLOSS($a$=1) | GLOSS($a$=0) | SLOSS($a$=1) | SLOSS($a$=0) |
|---|---|---|---|---|
| $A_1$ | 0.48 | 0.85 | 0.342 6 | 0.276 7 |
| $A_2$ | 0.502 8 | 0.738 6 | 0.339 5 | 0.334 |

图 5　$A_1$ 和 $A_2$ 的聚类分组信息损失($k$=4)

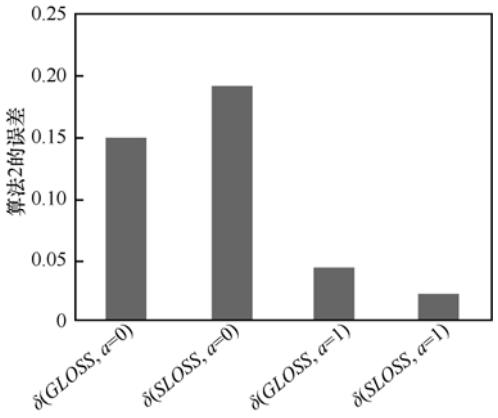



图 6　算法 2 的信息损失误差($k$=4)

该数据表明，最终发布的匿名数据集中包含的匿名组越多，该数据集包含的信息越丰富，且数据集的平均匿名组规模越小，信息损失越小，匿名化的数据越接近原来的真实数据，该数据集的可用性越高。

表 3　　基于聚类分组信息损失估计的聚类分组($k$=4)

| 结果 | $a$ | 聚类分组 |
|---|---|---|
| $A_1$ | $a=1$ | $S^1_{k=4}(G)=\{\{v_{10},v_{14},v_2,v_8,v_{21}\},\{v_{18},v_{11},v_7,v_9\},\{v_5,v_{13},v_{20},v_{19},v_{22}\},\{v_1,v_4,v_6,v_{16}\},\{v_{15},v_{17},v_{12},v_3\}\}$ |
| | $a=0$ | $S^0_{k=4}(G_1)=\{\{v_{10},v_2,v_1,v_3\}\{v_{18},v_{15},v_{16},v_{19}\},\{v_5,v_4,v_6,v_9\},\{v_{12},v_{14},v_7,v_8,v_{22}\},\{v_{11},v_{13},v_{17},v_{20},v_{21}\}\}$ |
| $A_2$ | $a=1$ | $S^1_{k=4}(G_1)=\{\{v_{10},v_{14},v_2,v_8,v_{22}\},\{v_{18},v_{11},v_7,v_9\},\{v_5,v_{13},v_{20},v_{19},v_6\},\{v_{12},v_3,v_{17},v_{16}\},\{v_{21},v_1,v_{15},v_4\}\}$ |
| | $a=0$ | $S^0_{k=4}(G_1)=\{\{v_0,v_1,v_2,v_3,v_{20},v_{22}\}\{v_{12},v_4,v_5,v_6\},\{v_{18},v_7,v_8,v_9\},\{v_{11},v_{13},v_{14},v_{15}\},\{v_{21},v_{16},v_{17},v_{19}\}\}$ |

## 5　结束语

针对社会网络发布过程中的安全问题，本文分析了基于扰动的攻击过程及相应的解决方法，建立了基于信息损失估计的 $k$ 匿名方法，通过子图节点属性信息泛化和子图结构信息泛化构建子图，在降低重构图的信息损失的同时，阻止扰动攻击。在社会网络更新过程中，首先判断网络结构变化，在网络变化率较小的情况下，通过损失部分信息的方法平衡网络计算时间复杂性，同时减少网络结构的破坏。后续研究工作中，将继续研究在动态的社会网络中如何以最少的信息损失取得最优的匿名级别问题。

## 参考文献：


[1] 韩毅, 方滨兴, 贾焰,等. 基于密度估计的社会网络特征簇挖掘方法[J]. 通信学报, 2012, 33(5):38-48.
HAN Y, FANG B X, JIA Y, et al. Mining characteristic clusters: a density estimation approach[J]. Journal on Communications, 2012, 33(5):38-48.

[2] WU X, YING X, LIU K. A survey of privacy-preservation of graphs and social networks[M]. Managing and mining graph data. Springer US, 2010: 421-453.

[3] CASAS-ROMA J, HERRERA-JOANCOMARTÍ J, TORRA V. Anonymizing graphs: measuring quality for clustering[J]. Knowledge & Information Systems, 2015, 44(3):1-22.

[4] BHAGAT S, CORMODE G, KRISHNAMURTHY B. Class based graph anaonymization for social network data[C]//35th International Conference on Very Large Data Base. c2009: 766-777.

[5] WANG R, ZHANG M, FENG D, et al. A clustering approach for privacy-preserving in social networks[C]//Information Security and Cryptology-ICISC 2014. Springer International Publishing, c2014: 193-204.

[6] JING Y, GOSSWEILER III R C. Using visualization techniques for adjustment of privacy settings in social networks[P]. US8832567. 2014.

[7] AGGARWAL C C, LI Y, YU P S. On the anonymizability of graphs[J]. Knowledge & Information Systems, 2015, 45(3):571-588.

[8] 兰丽辉, 鞠时光. 基于差分隐私的权重社会网络隐私保护[J]. 通信学报, 2015, 36(9):145-159.
LAN L H, JU S G. Privacy preserving based on differential privacy for weighted social networks[J]. Journal on Communications, 2015, 36(9):145-159.

[9] KARWA V, SLAVKOVIC A B, KRIVITSKY P N. Differentially private exponential random graphs[C]//Privacy in Statistical Database-UNESCO Chair in Data Privacy, International Conference, PSD 2014. Ibiza, Spain, c2014: 143-155.

[10] SALA A, ZHAO X, WILSON C. Sharing graphs using differentially private graph models[C]//The 2011 ACM SIGCOMM Conference on Internet Measurement, ACM, c2011: 81-98.

[11] MEDFORTH N, WANG K. Privacy risk in graph stream publishing for social network data[C]//The 2011 IEEE 11th International Conference on Data Mining. c2011: 437-446.

[12] ROSSI L, MUSOLESI M, TORSELLO A. On the *k*-anonymization of time-varying and multi-layer social graphs[J]. arXiv preprint arXiv: 1503. 06497, 2015.

[13] ZHOU B, PEI J. The *k*-anonymity and *l*-diversity approaches for privacy preservation in social networks against neighborhood attacks[J]. Knowledge and Information Systems, 2011, 28(1): 47-77.

[14] LIU C G, LIU I H, YAO W S, et al. *K*-anonymity against neighborhood attacks in weighted social networks[J]. Security & Communication Networks, 2015, 18(8): 3864-3882.

[15] LIU K, TERZI E. Towards identity anonymization on graphs[C]//The 2008 ACM SIGMOD International Conference on Management of Data. ACM, c2008: 93-106.

[16] CHENG J, FU A W, LIU J. *K*-isomorphism: privacy preserving network publication against structural attacks[C]//The 2010 ACM SIGMOD International Conference on Management of Data. ACM, c2010: 459-470.

[17] MICHEAL H, GEROME M, DAVID J. Resisting structural re-identification in anonymized social networks[J]. Proceedings of the VLDB Endowment, 2008, 1(1): 102-114.

[18] FUNG B C M, JIN Y, LI J, et al. Anonymizing social network data for maximal frequent-sharing pattern mining[M]//Recommendation and Search in Social Networks. Springer International Publishing, 2015:77-100.

[19] SWEENEY L. *k*-anonymity: a model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05): 557-570.

[20] BYUN J W, KAMRA A, BERTINO E, et al. Efficient *k*-anonymization using clustering techniques[C]//Dasfaa. Springer Berlin Heidelberg, c2007: 188-200.

[21] HAN J, KAMBER M. Data mining: comcepts and techniques[J]. San Francisco, 2006, 29(1): 1 - 25.



**作者简介：**

**苏洁**（1979-），女，山东淄博人，哈尔滨理工大学副教授、硕士生导师，主要研究方向为智能信息处理。

**刘帅**（1988-），男，山东济宁人，哈尔滨理工大学硕士生，主要研究方向为智能信息处理。

**罗智勇**（1978-），男，黑龙江大庆人，哈尔滨理工大学副教授、硕士生导师，主要研究方向为智能信息处理。

**孙广路**（1979-），男，黑龙江哈尔滨人，哈尔滨理工大学教授、硕士生导师，主要研究方向为计算机网络与信息安全、机器学习。