计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20140695
53(6):1352-1364,2016

# 增量的动态社会网络匿名化技术

郭彩华　王　斌　朱怀杰　杨晓春
(东北大学信息科学与工程学院　沈阳　110004)
(zhuhjneu@gmail.com)

# Incremental Dynamic Social Network Anonymity Technology

Guo Caihua, Wang Bin, Zhu Huaijie, and Yang Xiaochun
(*College of Information Science and Engineering, Northeastern University, Shenyang* 110004)

**Abstract**　With the rapid development and popularity of social networks, how to protect privacy in social networks has become a hot topic in the realm of data privacy. However, in most of existing anonymity technologies in recent years, the social network is abstracted into a simple graph. In fact, there are a lot of incremental changes of social networks in real life and a simple graph can not reflect incremental society network well, so abstracting the social network into the incremental sequences becomes more realistic. Meanwhile, in real life most of the network contains weight information, that is, a lot of social networks are weighted graphs. Compared with the simple graph, weighted graphs carry more information of the social network and will leak more privacy. In this paper, incremental dynamic social network is abstracted into a weighted graph incremental sequence. In order to anonymize weighted graph incremental sequence, we propose a weighted graph incremental sequence $k$-anonymous privacy protection model, and design a baseline anonymity algorithm (called WLKA) based on weight list and HVKA algorithm based on hypergraph, which prevents the attacks from node point labels and weight packages. Finally, through a lot of tests on real data sets, it proves that WLKA can guarantee the validity of the privacy protection on weighted graph incremental sequence. Compared with WLKA, HVKA is better to retain the original structure properties and improves the availability of weight information, but it also reduces the time cost of anonymous process.

**Key words**　dynamic social network; incremental sequence; data privacy; weight list; hypergraph; loss of information

**摘　要**　随着社会网络的快速发展和普及,如何保护社会网络中的敏感信息已成为当前数据隐私保护研究领域的热点问题.对此,近年来出现了多种社会网络匿名化技术.现有的匿名技术大多把社会网络抽象成简单图,然而实际生活中存在大量增量变化的社会网络,例如email通信网络,简单图并不能很好

**收稿日期**:2014-07-25;**修回日期**:2015-08-11
**基金项目**:国家"九七三"重点基础研究发展计划基金项目(2012CB316201);国家自然科学基金项目(61173031,61272178);国家自然科学基金海外及港澳学者合作基金项目(61129002);高等学校博士学科点专项科研基金项目(20110042110028);中央高校基本科研业务费专项资金项目(N120504001)
This work was supported by the National Basic Research Program of China (973 Program) (2012CB316201), the National Natural Science Foundation of China (61173031, 61272178), the National Natural Science Foundation of China Grant on International Cooperation (61129002), the Research Fund for the Doctoral Program of Higher Education of China (20110042110028), and the Specialized Fund for the Basic Research Operating Expenses Program of Central College (N120504001).
**通信作者**:王斌(binwang@mail.neu.edu.cn)

地刻画这种增量变化,因此,将社会网络抽象成增量序列具有现实意义.同时,在实际生活中大部分网络是带有权重信息的,即很多社会网络以加权图的形式出现,加权图与简单图相比携带了更多社会网络中的信息,也会带来更多的隐私泄露.将增量的动态社会网络抽象成一个加权图的增量序列.为了匿名加权图增量序列,提出了加权图增量序列 $k$-匿名隐私保护模型,并设计了基于权重链表的 baseline 匿名算法 WLKA 和基于超图的匿名算法 HVKA 来防止基于结点标签和权重链表的攻击.最后,通过在真实数据集上的大量测试,证明了 WLKA 算法能够保证加权图增量序列隐私保护的有效性,HVKA 算法则在 WLKA 的基础上更好地保留了原图的结构性质并提高了权重信息的可用性,同时还降低了匿名过程的时间代价.

**关键词** 动态社会网络;增量序列;数据隐私;权重链表;超图;信息损失

**中图法分类号** TP311

随着互联网的发展,出现了越来越多的网络平台,用户在网络上共享信息并参加各种各样的活动.社会网络的出现给人们的生活带来了极大的便利,但随之而来的是越来越多的社会网络数据被发布到网络上,导致隐私泄露.近年来,如何在发布网络数据的同时保护敏感信息成为了研究热点.目前大部分社会网络隐私保护技术主要针对简单图,然而实际生活中存在大量增量变化的社会网络,例如 email 通信网络,简单图并不能很好地刻画这种增量变化,因此,将社会网络抽象成增量序列具有现实意义.

文献[1]考虑了社会网络的动态性,将社会网络抽象成一系列图组成的增量序列,即下一时刻图中结点和边都是增加的,并针对简单图增量序列的隐私问题进行了研究.该文把简单图增量序列定义为 $g=\langle G_0,G_1,\cdots,G_T\rangle$,时刻 $t$ 的社会网络图表示为 $G_t=(V_t,E_t,L_t)$,其中 $V_t$ 代表时刻 $t$ 结点的集合,$E_t$ 代表时刻 $t$ 边的集合,$L_t$ 代表时刻 $t$ 结点标签的集合.由于它是一个增量序列,因此,$V_{t-1}\subseteq V_t$,$E_{t-1}\subseteq E_t$,$L_{t-1}\subseteq L_t$.

文献[1]中的匿名方法保证 $G_t(t=0,1,\cdots,T)$ 符合 $k$ 匿名,即攻击者以结点标签为背景知识识别任意结点的概率不大于 $1/k$.它不仅考虑了攻击者将结点标签作为背景知识,同时也将社会网络的动态变化视为背景知识.图 1(a)(b)分别是简单图增量序列 $g=\langle G_0,G_1\rangle$ 在时刻 $t$ 为 0 或 1 的原始图,时刻 $t=1$ 新增加的结点和边用虚线标出.在文献[1]的匿名算法中通过将标签相近的结点分为一组,使得同组结点基于标签不可区分,$k=2$ 时,时刻 $t$ 为 0 或 1 的分组结果分别如图 2(a)(b)所示,匿名图分别如图 1(c)(d)所示.每个标签均对应 2 个结点,所以攻击者以结点标签为背景知识准确识别任意 1 个结点的概率均不大于 50%.例如,攻击者知道 Alice 的标签是{25,M,TW},首先单独针对图 2(a)(或图 2(b))所示的分组结果中均有 2 个结点 $v_5$,$v_8$ 与标签{25,M,TW}对应,所以攻击者识别 Alice 的概率为 50%.另外,再联合不同时刻 $t$ 为 0 或 1 的分组结果同样不能正确识别结点身份,所以攻击者仍然无法正确识别 Alice.

然而,在现实中进行社会网络分析时,发现结点之间的边是带有权重信息的,即存在许多加权社会网络图.例如在金融交易网络中,边权重表示企业之间交易金额的大小.在文献合作关系网中,边权重表示作者之间合作的次数.在 email 通信网络中,边权重表示实体之间发送邮件的次数.在朋友关系网中,边权重表示朋友之间关系的密切程度.跟简单图相比,加权图携带了更多社会网络中的信息,更能准确地表示现实世界中的社会网络,具有更强的描述能力,同时也会带来更多的隐私泄露,因此对于加权社会网络的隐私保护更为重要.文献[1]中并没有考虑权重信息对结点隐私的影响.图 1(f)是 $G_1$ 的加权图,假设攻击者知道结点 Alice 的标签{25,M,TW}及权重链表⟨3,1⟩(权重链表是指结点与邻居连接边上的权值按降序排列得到的权重序列).攻击者根据标签得到 2 个候选结点 $v_5$,$v_8$,再根据权重链表则可得到唯一的候选结点 $v_8$,正确识别结点的概率为 100%,这就导致了用户身份信息泄露.

现有的动态社会网络隐私保护模型只考虑了简单的无权图,并没有考虑加权图.对此,本文将社会网络抽象成加权图增量序列,并对加权图增量序列上的隐私问题进行研究.本文是第 1 篇解决加权图增量序列隐私保护问题的文章.本文的方法包括 3 部分:

1) 对加权图增量序列最后一时刻的原始图中所有结点分组,分组方法采用文献[1]中对结点分组

的方法,使得同一分组中的结点基于标签不可区分;

2) 对每个分组中所有结点的权重链表匿名,使得同一分组中所有结点基于权重链表不可区分;

3) 根据时刻 $t(t=0,1,\cdots,T)$ 的匿名图推出时刻 $t-1$ 的匿名图,重复这个过程直到所有时刻的图都完成匿名.

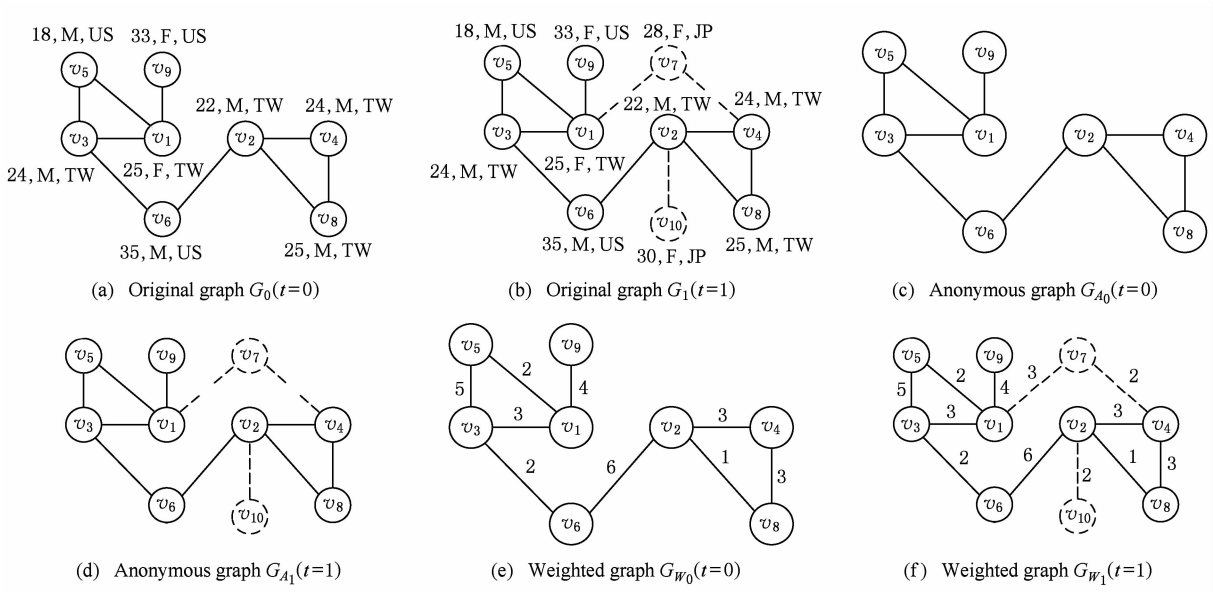



Fig. 1 An example of anonymity incremental sequence $g=\langle G_0,G_1\rangle(k=2)$.

图 1 增量序列 $g=\langle G_0,G_1\rangle$ 的匿名化实例($k=2$)

| | |
|---|---|
| $v_1,v_2$ | 25,F,TW |
| | 22,M,TW |
| $v_3,v_4$ | 24,M,TW |
| | 24,M,TW |
| $v_5,v_8$ | 18,M,US |
| | 25,M,TW |
| $v_6,v_9$ | 35,M,US |
| | 33,F,US |
| $v_7,v_{10}$ | 28,F,JP |
| | 30,F,JP |

(a) Grouping results($t=0$)

| | |
|---|---|
| $v_1,v_2$ | 25,F,TW |
| | 22,M,TW |
| $v_3,v_4$ | 24,M,TW |
| | 24,M,TW |
| $v_5,v_8$ | 18,M,US |
| | 25,M,TW |
| $v_6,v_9$ | 35,M,US |
| | 33,F,US |

(b) Grouping results($t=1$)

Fig. 2 Grouping results of $g=\langle G_0,G_1\rangle(k=2)$.

图 2 增量序列 $g=\langle G_0,G_1\rangle$ 的分组结果($k=2$)

本文的主要贡献如下:

1) 引入了增量序列分类安全条件(ISCSC)来指导匿名过程,并在此基础上提出了加权图增量序列 $k$-匿名隐私保护模型,从而有效地防止攻击者以标签及权重链表为背景知识的隐私攻击;

2) 设计了一个基于权重链表的 baseline 匿名算法 WLKA,WLKA 算法能够保证加权图增量序列隐私保护的有效性;

3) 设计了基于超图的 HVKA 隐私保护算法,在生成匿名图的同时,能够提高图数据的可用性并减少匿名的时间代价;

4) 通过在真实数据集上的大量实验测试和分析,验证了加权图增量序列 $k$-匿名隐私保护模型的有效性以及发布图数据的高可用性.

## 1 相关工作

随着互联网技术的迅速发展,社会网络应用不断涌现,管理和分析社会网络数据成了研究热点.然而在社会网络数据中隐藏着大量的敏感信息,发布和共享社会网络数据会导致隐私泄露,由此,近年来越来越多的研究学者开始把注意力集中到社会网络数据的隐私保护问题上,并提出了多种匿名方法.

文献[2]提出了 $k$-度匿名,用来防止攻击者基于结点度的隐私攻击,该方法使得匿名图中对任意 1 个结点,至少存在 $k-1$ 个其他结点与其度相同.文献[3]设计了 $k$-邻居匿名方法,用来防止基于 1-邻居图的隐私攻击,该方法使得匿名图中对任意一个结点的 1-邻居图,至少存在 $k-1$ 个结点的 1-邻居图与其同构.文献[4]提出了 $k$-自同构,是指图自身存在着 $k$ 个同构映射,使得攻击者进行结点身份识别时,至少存在 $k$ 个候选符合.文献[5]提出了 $k$-同构模型,把社会网络分为 $k$ 个子图,使子图之间互相同构,从而防止隐私泄露.

文献[6-10]把结点与邻居连接边上的属性值序列作为背景知识.文献[6]中把结点与邻居连接边上的属性值序列定义为权重包,并提出 $k$-直方匿名,使得对任意一个结点,至少存在 $k-1$ 个其他结点与其权重包相同;文献[7]提出的加权图匿名化方法是使得对于任意结点的权重包,至少存在 $k-1$ 个其他结点的权重包与其距离小于阈值,而不是相等;文献[8]提出了 $k$-可能隐私保护模型,并通过边权重概括将加权图匿名化为 $k$-可能图;文献[9]讨论了社会网络上的隐私问题,并允许用户设定个性化隐私;文献[10]针对加权社会网络中最短路径识别提出了加权图 $k$-可能路径匿名隐私保护模型来防止基于最短路径隐私攻击;文献[11]针对障碍空间中保持位置隐私的最近邻查询问题提出了一种第三方可靠服务器的方法,该方法能够保证用户在享受基于位置服务所提供的实际准确答案的同时保证位置信息不被泄露.

文献[12]针对攻击者基于结点属性值进行边再识别隐私攻击,提出了基于聚类的模型.文献[1]中用带标签的无向图表示社会网络,并把社会网络的演变抽象成一个增量序列,解决了增量序列上的结点身份再识别问题.但在现实世界中存在大量的加权图,文献[1]忽略了权重对隐私的影响.因此,本文主要研究动态变化的加权社会网络上的结点身份再识别问题.

## 2 背景知识及问题定义

本文主要考虑加权社会网络增量序列的隐私保护问题.下面对文中用到的基本概念进行详细地介绍,为后续工作奠定基础.

**定义 1.** 加权图增量序列.随时间增量变化的加权社会网络表示为 $g_W=\langle G_{W_0},G_{W_1},\cdots,G_{W_T}\rangle$,时刻 $t$ 的社会网络 $G_{W_t}=(V_t,E_t,L_t,W_t)$,其中,$V_t$ 代表时刻 $t$ 的结点集合,$E_t$ 代表时刻 $t$ 结点间边的集合,$L_t$ 代表时刻 $t$ 结点标签的集合,$W_t$ 代表时刻 $t$ 权重链表的集合.假设在 $t(t=0,1,\cdots,T)$ 的下一时刻社会网络中顶点和边的数量非递减,即 $V_{t-1}\subseteq V_t$,$E_{t-1}\subseteq E_t$,$L_{t-1}\subseteq L_t$,那么称 $g_W$ 为加权图增量序列.

图 3(a)(b)分别为加权图增量序列 $g_W=\langle G_{W_0},G_{W_1}\rangle$ 在时刻 $t=0$ 和 $t=1$ 的快照,时刻 $t=1$ 增加的结点和边用虚线给出.

**定义 2.** 结点权重链表.$\forall v\in V$,把 $v$ 邻边上的权重值降序排列,就得到 $v$ 的权重链表,记为 $w=\langle w_1,w_2,\cdots,w_d\rangle$,$(w_1\geqslant w_2\geqslant\cdots\geqslant w_d)$,其中 $d$ 是结点 $v$ 的度.

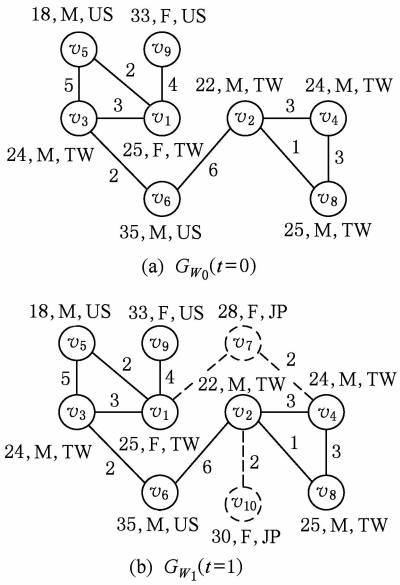



Fig. 3 Weighted graph increment sequence $g_W=\langle G_{W_0},G_{W_1}\rangle$.

图 3 加权图增量序列 $g_W=\langle G_{W_0},G_{W_1}\rangle$

**定义 3.** 权重链表统一化.对图 $G(V,E,L,W)$ 中所有结点进行分组得到分组集合 $C$,每个分组中至少包含 $k$ 个结点,$\forall c\in C$,保证 $c$ 中所有结点的权重链表相同,则称这个过程为权重链表统一化.

对图 3(b)所示的加权图进行分组得到的结果如图 2(b)所示,结点 $v_3$,$v_4$ 被分为一组,权重链表分别为 $w_3=\langle 5,3,2\rangle$,$w_4=\langle 3,3,2\rangle$,进行权重链表统一化后 $w_3=w_4=\langle 5,3,2\rangle$.

**定义 4.** 结点身份泄露.给定加权图增量序列 $g_W=\langle G_{W_0},G_{W_1},\cdots,G_{W_T}\rangle$,如果攻击者以结点标签及权重链表为背景知识能够准确识别出结点 $v$,则结点 $v$ 存在结点身份泄露.

如图 1(f),攻击者根据结点 Alice 的标签{25,M,TW}和权重链表⟨3,1⟩能以 100%的概率识别出结点 $v_8$ 是 Alice,这说明结点 $v_8$ 存在结点身份泄露问题.

**问题定义.** 加权图增量序列匿名化问题.

已知加权图增量序列 $g_W=\langle G_{W_0},G_{W_1},\cdots,G_{W_T}\rangle$、正整数 $k$,发布一个匿名序列 $g'_W=\langle G'_{W_0},G'_{W_1},\cdots,G'_{W_T}\rangle$,使任意时刻图 $G'_{W_t}(t=0,1,\cdots,T)$ 满足以下隐私目标:

1) 对匿名图中任意一条边 $e$,攻击者正确识别其端结点的概率不大于 $1/k$.

2) 对匿名图中任意 2 个结点,攻击者正确判断

该结点间存在连接的概率不大于 $1/k$.

3) 在匿名图中,攻击者以结点标签及权重链表为背景知识准确识别某结点的概率不大于 $1/k$.

## 3 加权图增量序列 $k$-匿名隐私保护模型

为了防止加权图增量序列中结点身份泄露问题,本文提出了基于增量序列分类安全条件的 $k$-匿名隐私保护模型. 为了得到满足隐私目标的 $k$-匿名隐私保护模型,先给出一些例子和性质.

**例 1.** 如图 3(b)中假设 $v_1$ 和 $v_3$ 被分为 1 组,攻击者可以确定出 $v_1$ 和 $v_3$ 之间存在连接,违反了隐私目标. 另外,对如图 4 所示的图分组,得到的分组结果为 $C=\{\{v_1,v_2\},\{v_3,v_4\}\}$,2 个分组间存在 4 条边,攻击者可以确定 $v_1$ 和 $v_2$ 具有相同的朋友 $v_3$ 和 $v_4$,从而造成共同朋友关系泄露.

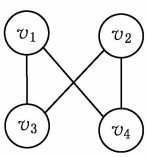



Fig. 4 An example of dense links ($k=2$).

图 4 分组间存在大量边的例子($k=2$)

根据例 1 的分析,我们可以得到性质 1.

**性质 1.** 对于 $g_W$ 中任意时刻原始图的匿名,同一分组及 2 个分组间均不能存在大量的边,否则会造成隐私泄露.

**例 2.** 如图 3 为加权图增量序列 $g_W=\langle G_{W_0},G_{W_1}\rangle$,首先对 $G_{W_1}$ 进行匿名,假设得到的分组结果为 $C=\{\{v_1,v_2\},\{v_3,v_4\},\{v_5,v_8\},\{v_6,v_7\},\{v_9,v_{10}\}\}$. 接下来,通过移除时刻 $t=1$ 添加的结点和边得到时刻 $t=0$ 的匿名图,分组结果变为 $C=\{\{v_1,v_2\},\{v_3,v_4\},\{v_5,v_8\},\{v_6\},\{v_9\}\}$,分组 $\{v_6\}$ 和 $\{v_9\}$ 不符合 $k=2$ 的匿名约束条件,造成隐私泄露. 这样可以分析得到,时刻 $t-1$ 的匿名图是通过移除时刻 $t(t=1,2,\cdots,T)$ 添加的结点和边得到的,由于在移除结点后,包含移除结点的分组的大小有可能小于 $k$,所以结点的移除将会造成隐私泄露.

根据例 2 的分析,我们可以得到性质 2.

**性质 2.** 对于 $g_W$ 中任意时刻原始图的匿名,同一分组中的结点必须来自同一时间戳,否则会造成隐私泄露.

为了防止上述性质 1 和性质 2 中情况的发生,本文引入文献[1]中指导结点分组过程的时间序列分类安全条件,定义加权图增量序列匿名的增量序列分类安全条件(incremental sequence class safety condition, ISCSC),对结点分组的过程中,必须满足 ISCSC.

**定义 5.** 增量序列分类安全条件(ISCSC). 对结点集 $V_t$ 的分组满足 ISCSC,当且仅当 $\forall v\in V_t$ 及 $\forall C\subset V_t$ 满足如下条件:

1) $\forall v\in V_{t'},\forall w\in V_{t'}:v\in C\wedge w\in C\Rightarrow t=t'$.

2) $\forall (v,w)\in E_t:v\in C\wedge w\in C\Rightarrow v=w$.

3) $\forall C_a,C_b\subset V_t,n_e$ 是 $C_a$ 和 $C_b$ 之间边的个数$\Rightarrow n_e\leqslant k$.

条件 1 表明了同一分组中的结点来自同一时间戳;条件 2 限制了对任一时间戳同一分组中不含边;条件 3 限制了对任一时间戳 2 个分组之间边的数量.

**定理 1.** 对加权图增量序列 $g_W=\langle G_{W_0},G_{W_1},\cdots,G_{W_T}\rangle$ 的匿名过程中,满足增量序列分类安全条件(ISCSC)是实现隐私目标的充分条件.

证明. ISCSC 的条件 1($\forall v\in V_{t'},\forall w\in V_{t'}:v\in C\wedge w\in C\Rightarrow t=t'$)保证了在 $\forall$ 时刻 $t(t=0,1,\cdots,T)$ 的匿名图中同一类中的结点属于同一时间戳,也就是说,在时刻 $t-1(t=1,2,\cdots,T)$ 删除的结点都属于同一个类,因此删除后剩余的每个类中仍包含 $k$ 个结点. 攻击者不能根据多重发布识别用户的身份信息.

ISCSC 的条件 2($\forall (v,w)\in E_t:v\in C\wedge w\in C\Rightarrow v=w$)限制了对任一时间戳同一类中不含边. 如果在时刻 $t(t=0,1,\cdots,T)$ 的匿名图中 2 个结点间存在一个边,那么这 2 个结点的真实标签必然属于不同的类. 此外,匿名后每个类有 $k$ 个标签并且各个类的标签不重叠. 因此,对于任意一条边,它的 2 个端点都存在 $k$ 个候选标签,保证了隐私目标 1.

ISCSC 的条件 3($\forall C_a,C_b\subset V_t,n_e$ 是 $C_a$ 和 $C_b$ 之间边的个数$\Rightarrow n_e\leqslant k$)限制了对任一时间戳 2 个类之间边的数量小于或等于 $k$. 假设时刻 $t$ 存在 2 个类 $class_1$ 和 $class_2$,并且用户 $user_1$ 属于 $class_1$,$user_2$ 属于 $class_2$,那么用户 $user_1$ 和 $user_2$ 之间存在边的概率可表示为

$$P(edge(user_1,user_2))=\frac{(|class_1|-1)!\times(|class_2|-1)!\times n_e}{|class_1|!\times|class_2|!},$$

因为 $class_1$ 和 $class_2$ 的大小是 $k$,并且概率 $P(edge(user_1,user_2))$ 应该小于或等于 $1/k$,因此可推导出 $n_e\leqslant k$. 所以 2 个类之间边的数量小于或等于 $k$ 保证了隐私目标 2.

综合以上 3 点可知,要想实现隐私目标就必须在分组的过程中满足 ISCSC,所以满足 ISCSC 是实现隐私目标的充分条件. 证毕.

**定义 6.** 加权图增量序列 $k$-匿名隐私保护模型. 给定加权图增量序列 $g_W=\langle G_{W_0}, G_{W_1},\cdots, G_{W_T}\rangle$,参数 $k$.

1) 将图 $G_{W_T}$ 中所有结点分成大小至少为 $k$ 的若干分组,分组过程满足 ISCSC;

2) 对分组中结点进行权重链表统一化,使得同一分组中的结点基于标签及权重链表不可区分,得到匿名图 $G'_{W_T}$;

3) 然后通过移除时刻 $t=T$ 添加的结点和边得到 $G'_{W_{T-1}}$,移除过程不破坏隐私,$G'_{W_{T-1}}$ 仍满足隐私目标,重复这个过程直到所有时刻的图都完成匿名.

通过这 3 个条件,最后得到匿名序列 $g'_W=\langle G'_{W_0}, G'_{W_1},\cdots, G'_{W_T}\rangle$,符合 $k$-匿名隐私保护模型.

为了提高匿名图的可用性,将属性相近的结点分到同一组或者相邻组中,对此我们定义属性优先列表,在分组前根据属性优先列表将结点进行排序. 排序的顺序和 ISCSC 共同决定了结点的分组.

**定义 7.** 属性优先列表. 给定一个属性列表 $a_0, a_1,\cdots,a_n$,按属性对查询的重要程度降序排列,得到属性优先列表,记为 $list_{ap}=\{a_0,a_1,\cdots,a_n\}$,对于任意 $i<j$,说明 $a_i$ 比 $a_j$ 对查询更重要. 本文中假设属性优先列表已给出,记为

$$list_{ap}=\langle location, gender, age\rangle.$$

## 4 加权图增量序列的隐私保护算法

为了获得符合加权图增量序列 $k$-匿名隐私保护模型的匿名序列,本节首先给出一种基于权重链表的 baseline 算法 WLKA. 随后,为了提高匿名图的可用性,给出一种基于超图的高效匿名算法 HVKA.

### 4.1 基于权重链表的 baseline 算法

在研究动机中分析到如果直接采用文献[1]的方法,也可能会造成隐私泄露问题. 本文首先借鉴文献[1]中对结点分组的方法给出了满足 ISCSC 的分组算法 Grouping,然后在分组算法的基础上给出基于权重链表的 baseline 算法 WLKA.

**算法 1.** 分组算法(Grouping).

输入:原始图 $G_{W_T}$、参数 $k$、属性优先列表 $list_{ap}$;

输出:分组集合 $C$.

① 初始化:有序表 $V_{list}=\varnothing$,分组集合 $C=\varnothing$;

② 根据 $list_{ap}$ 排序 $G_{W_T}$ 中结点,返回有序表 $V_{list}$;

③ 创建只包含 $V_{list}$ 中第 1 个结点的分组 $c$,$C=C\cup\{c\}$;

④ FOR $\forall v\in V_{list}$

⑤ FOR $\forall c\in C$

⑥ IF $Size(c)<k$ AND $c=c\cup\{v\}$不违反 ISCSC TEHN

⑦ $c=c\cup\{v\}$, break;

⑧ END IF

⑨ END FOR

⑩ IF 没找到小于 $k$ 的分组或者不满足 ISCSC THEN

⑪ 创建一个新分组 $c$,使 $c=c\cup\{v\}$,$C=C\cup\{c\}$;

⑫ END IF

⑬ END FOR

⑭ FOR $\forall c\in C$ AND $Size(c)<k$

⑮ $Size(c')<k$, 并从 $C$ 中移除 $c'$; /* 找小于 $k$ 的 $c'$ */

⑯ 将$c'$中不违反 ISCSC 的结点 $v$ 加入 $c$ 中,$c=c\cup\{v\}$;

⑰ IF $Size(c')!=0$ /* 分组 $c'$中仍含有结点 */ THEN

⑱ $C=C\cup\{c'\}$; /* 将分组 $c'$加入分组集合 $C$ 中 */

⑲ END IF

⑳ END FOR /* 合并小于 $k$ 的分组 */

㉑ FOR $\forall c\in C$ AND $Size(c)<k$

㉒ 给类 $c$ 中添加 $k$-$Size(c)$个虚拟结点;

㉓ END FOR /* 为结点数小于 k 的分组增加虚拟结点 */

㉔ RETURN $C$.

分组算法 Grouping 的伪代码见算法 1. Grouping 首先根据属性优先列表对结点排序,然后由 ISCSC 指导分组全过程. 算法 1 行①完成初始化工作. 行②根据 $list_{ap}=\langle location, gender, age\rangle$排序 $G_{W_T}$ 中结点生成有序表 $V_{list}$. 行③建立只包含 $V_{list}$ 中第 1 个结点的分组. 行④~⑬对 $v\in V_{list}$,顺序地把 $v$ 插入到结点数小于 $k$ 并且满足 ISCSC 的分组中. 对于某结点 $v$,如果没有找到小于 $k$ 的分组或者不满足 ISCSC,那么创建一个新的分组. 行⑭~⑳对每个小于 $k$ 的分组,检查是否可以和其他分组合并,如果可以,就合并这 2 个小于 $k$ 的分组,以减少小于 $k$ 的分组数量. 行㉑~㉓为结点数仍小于 $k$ 的分组增加虚拟结点,虚拟结点的属性标签随机地从该组中结点的属性

标签中挑选. 例如,假设分组 $c=\{v_1,v_2\}$,$k=3$,结点 $v_1$ 的属性标签为{24,M,TW},结点 $v_2$ 的属性标签为{22,M,TW},则虚拟结点的标签既可以是{24,M,TW},也可以是{22,M,TW}. 行㉕返回分组集合 $C$.

根据文献[1]以及加权图增量序列 $k$-匿名隐私保护模型,可以得到性质 3.

**性质 3.** 分组算法 Grouping 利用属性优先列表以及 ISCSC 指导分组得到集合 $C$;再对 $C$ 中每个分组进行权重链表统一化. 由此得到匿名图是安全的.

根据定义 5,我们首先得到时刻 $t=T$ 的匿名图 $G'_{W_T}$ 后,然后通过移除时刻 $t=T$ 新添加的结点和边得到时刻 $t=T-1$ 的匿名图仍满足隐私目标,以此类推,可以得到时刻 $t(t=0,1,\cdots,T-2)$的匿名图.

结合性质 3 和定义 5,本文提出了基于权重链表的 $k$-匿名算法(WLKA),如算法 2 所示.

**算法 2.** 基于权重链表的 $k$-匿名算法(WLKA).

输入:加权图增量序列 $g_W=\langle G_{W_0}, G_{W_1},\cdots, G_{W_T}\rangle$;

输出:匿名序列 $g'_W=\langle G'_{W_0}, G'_{W_1},\cdots, G'_{W_T}\rangle$.

① 基于 Grouping 算法生成 $G_{W_T}$ 的分组集合 $C$;

② FOR $\forall c\in C$

③ 对 $c$ 中每个结点进行权重链表统一化;

④ END FOR

⑤ 发布匿名图 $G'_{W_T}$;

⑥ WHILE $t!=0$

⑦ 移除 $v\in V_t/V_{t-1}$ 及与其相连的虚拟结点,$e\in E_t/E_{t-1}$;

⑧ 再次对每个分组 $c$ 中的结点进行权重链表统一化;

⑨ 发布匿名图 $G'_{W_{t-1}}$;

⑩ $t=t-1$;

⑪ END WHILE

算法 2 中行①利用 Grouping 算法对时刻 $t=T$ 原始图 $G_{W_T}$ 分组得到若干个大小至少为 $k$ 的分组;行②~④对每个分组中的结点进行权重链表统一化,如果权重链表维数不等,那么添加虚拟结点,虚拟结点的属性标签随机地从该组中结点的属性标签中挑选,使得同一组内所有结点具有相同的权重链表;行⑤发布匿名图 $G'_{W_T}$;行⑥~⑪根据时刻 $t$ 的匿名图 $G'_{W_t}(t=1,2,\cdots,T)$依次生成时刻 $t-1$ 的匿名图 $G'_{W_{t-1}}$ 并发布. 主要做法是移除结点 $v\in V_t/V_{t-1}$ 及与其相连的虚拟结点,边 $e\in E_t/E_{t-1}$,并再次对每组中结点进行权重链表统一化,得到 $G'_{W_{t-1}}$.

以图 3(a)(b)所示的加权图增量序列 $g_W=\langle G_{W_0},G_{W_1}\rangle$为例描述 WLKA 的匿名过程,匿名结果如图 5 所示. 由于对 $G_{W_1}$ 进行权重链表统一化过程中 $v_6$ 和 $v_9$ 的权重链表维数不同,因此在 $v_9$ 上添加了虚拟结点 $v_{11}$,同理也在 $v_{10}$ 上添加了虚拟结点 $v_{12}$.

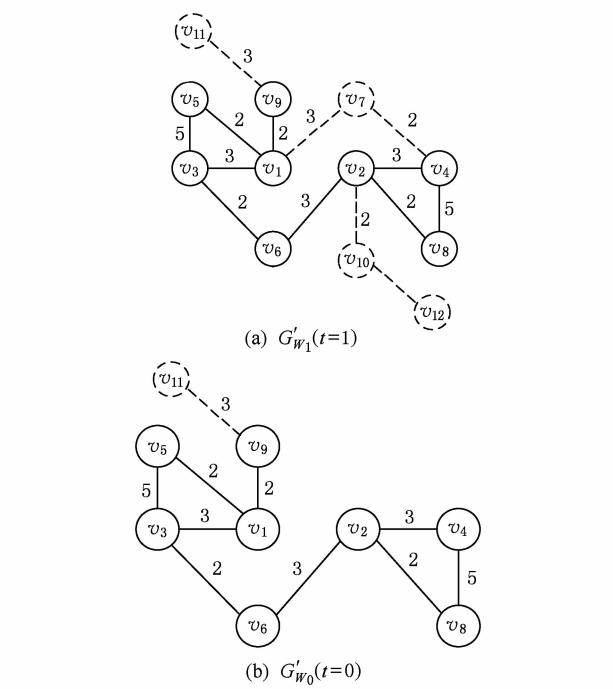



Fig. 5 Anonymous results of WLKA on $g=\langle G_{W_0},G_{W_1}\rangle$.

图 5 $g_W=\langle G_{W_0},G_{W_1}\rangle$的 WLKA 匿名结果

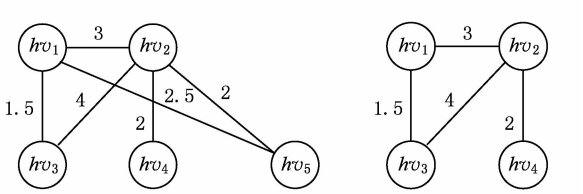
### 4.2 基于超图的 $k$-匿名算法

算法 2 为加权图增量序列提供了一种有效的匿名算法. WLKA 算法为了实现权重链表统一化,会添加虚拟结点,并改变边权重,这将导致可用性下降,而且 WLKA 算法需要对每一时刻单独进行权重链表统一化,带来了较大的时间代价. 为了提高匿名序列的可用性和减少时间代价,本节提出一种基于超图的 $k$-匿名算法 HVKA.

给出基于超图的 $k$-匿名算法 HVKA 之前,先给出超图的基本定义以及超图可用性的性质.

**定义 8.** 超图、超边、超点. 超图是指将原图中所有结点和边聚合成若干超点和超边后构成的图. 其中超点 $hv_i$ 代表图中的一组结点,超边 $e_{hv_i,hv_j}$ 代表超点 $hv_i$ 和 $hv_j$ 间所有可能的边.

例如,对图 3(b)所示的 $G_{W_1}$ 中所有结点聚合得到图 6(a)所示的超图,结点 $v_1,v_2$ 聚类成超点 $hv_1$,结点 $v_3,v_4$ 聚类成超点 $hv_2$,超边 $e_{hv_1,hv_2}$ 代表了 $hv_1$ 和 $hv_2$ 间所有的边 $e_{v_1,v_3},e_{v_2,v_4}$.




(a) $G'_{W_1}(t=1)$ (b) $G'_{W_0}(t=0)$

Fig. 6 Anonymous results of HVKA on $g=\langle G_{W_0}, G_{W_1}\rangle$.

图 6 $g_W=\langle G_{W_0}, G_{W_1}\rangle$的 HVKA 匿名结果

下面给出计算超边的权重定义.

超边的权重定义为原始图边权重的平均值:

$$w'(hv_i, hv_j)=\frac{\sum_{e\in I,J} w_e}{IJ}, \tag{1}$$

其中,$IJ=E\bigcap(hv_i\times hv_j)$.

例如,由图 3(b)所示的 $G_{W_1}$ 经聚合得到图 6(a)所示的超图,结点 $v_1$,$v_2$ 聚类成超点 $hv_1$,结点 $v_3$,$v_4$ 聚类成超点 $hv_2$.根据等式(1)可算得 $hv_1$ 和 $hv_2$ 间的权重为

$$w'(hv_1, hv_2)=[w(v_1, v_3)+w(v_2, v_4)]/2=3.0.$$

同理可得:

$$w'(hv_1, hv_3)=1.5,$$
$$w'(hv_1, hv_5)=2.5,$$
$$w'(hv_2, hv_3)=4.0,$$
$$w'(hv_2, hv_4)=2.0,$$
$$w'(hv_2, hv_5)=2.0.$$

基于超图的 $k$-匿名算法同样利用分组算法 Grouping 获得 $G_{W_T}$ 的分组集合 $C$;然后利用聚合算法 Aggregating 得到时刻 $t=T$ 的匿名图 $G'_{W_T}$;最后根据时刻 $t(t=1,2,\cdots,T)$ 的匿名图得到时刻 $t-1$ 的匿名图;继而生成匿名序列.

算法 3 给出了聚合算法 Aggregating 的伪代码.聚合算法 Aggregating 的作用是将分组集合 $C$ 中同一分组中的结点聚类成超点,不同分组间的边聚类成超边,同时计算超边上的权重,从而得到超图,也就是匿名图.

**算法 3.** 聚合算法(Aggregating).

输入:原始图 $G_{W_T}$ 及分组集合 $C$;

输出:匿名图 $G'_{W_T}$.

① FOR $\forall c_i\in C$

② $hv_i=\{v_i\}$; /* 建立包含 $c_i$ 中一个结点的超点 */

③ FOR $\forall v_j\in c_i$

④ $hv_j=\{v_j\}$; /* 建立包含 $c_i$ 中另一个结点的超点 */

⑤ $hv_{\text{new}}=hv_i\bigcup hv_j$; /* 将 2 个超点合并 */

⑥ $u_i=hv_i$ 的邻结点; /* 记录 2 个超点的邻结点 */

⑦ $u_j=hv_j$ 的邻结点;

⑧ FOR $\forall u\in u_i\bigcup u_j$

⑨ 建立超边 $e_{u,hv_{\text{new}}}$; /* 建超点与邻结点间超边 */

⑩ 计算权重 $w'_{u,hv_{\text{new}}}$; /* 计算超边上权重 */

⑪ 删除边 $e_{u,hv_i}$,$e_{u,hv_j}$; /* 删除原边 */

⑫ 删除结点 $hv_i$,$hv_j$; /* 删除原结点 */

⑬ END FOR

⑭ END FOR /* $c_i$ 中所有结点和边都完成聚合 */

⑮ $hv_i=hv_{\text{new}}$; /* $hv_i$ 为 $c_i$ 中所有结点聚合后得到超点 */

⑯ END FOR /* 每个分组结点和边都聚合成超点、边 */

基于超图的 $k$-匿名算法 HVKA,如算法 4 所示,其形式化地描述了 HVKA 算法的整个匿名过程.算法 4 的行①利用 Grouping 算法对时刻 $t=T$ 原始图 $G_{W_T}$ 分组得到若干个大小至少为 $k$ 的分组;行②利用 Aggregating 算法将分组中结点和边聚集成超点和超边,并为每个超边计算权重得到时刻 $t=T$ 的匿名图 $G'_{W_T}$;行③发布匿名图 $G'_{W_T}$;行④~⑧移除 $G'_{W_t}(t=1,2,\cdots,T)$ 中由结点 $v\in V_t/V_{t-1}$ 构成的超点及与其相连的超边,得到 $G'_{W_{t-1}}$.

**算法 4.** 基于超图的 $k$-匿名算法(HVKA).

输入:加权图增量序列 $g_W=\langle G_{W_0}, G_{W_1},\cdots, G_{W_T}\rangle$;

输出:匿名序列 $g'_W=\langle G'_{W_0}, G'_{W_1},\cdots, G'_{W_T}\rangle$.

① 基于 Grouping 算法生成 $G_{W_T}$ 的分组集合 $C$;

② 基于 Aggregating 算法生成超图 $G'_{W_T}$;

③ 发布匿名图 $G'_{W_T}$;

④ WHILE $t!=0$

⑤ 移除 $G'_{W_t}$ 中结点 $v\in V_t/V_{t-1}$ 构成的超点及相连超边;

⑥ 发布匿名图 $G'_{W_{t-1}}$;

⑦ $t=t-1$;

⑧ END WHILE

HVKA 算法为加权图增量序列提供了高效的匿名方法.利用 HVKA 算法匿名如图 3 所示的加权图增量序列 $g_W=\langle G_{W_0}, G_{W_1}\rangle$,得到的匿名序列如图 6 所示.匿名图 $G'_{W_1}$ 和 $G'_{W_0}$ 中每个超点都包含 2 个结点,因此攻击者正确识别结点身份的概率不大于 1/2,

同时,超边上的权重不同于原始图中的权重,攻击者以权重链表为背景知识无法正确识别身份,避免了结点身份泄露.由于匿名过程中不需要添加结点,也不需要单独对每一时刻的权重链表统一化,因此HVKA算法具有高可用性和较低的时间代价.

**定理 2.** 在匿名过程中,边权重改变量越小,匿名图就能更好地保留原图的性质,可用性就越高.

HVKA算法借助超图的思想在匿名过程中对权重的改变量较小,并且不需要添加大量虚拟结点.根据定理2可知,经HVKA匿名得到的匿名图具有高可用性.

图的匿名过程必然会造成图数据可用性的下降.将加权图 $G_{W_T}$ 进行匿名后得到匿名图 $G'_{W_T}$,由权重变化所导致的信息损失率定义为

$$IL(G_W, G'_{W_T}) = \sum_{(i,j)\in E_T} \frac{|w(i,j) - w'(i,j)|}{W(G_{W_T})}, \quad (2)$$

其中,$w'(i,j)$为匿名后的权重,$W(G_{W_T})$为原始图 $G_{W_T}$ 中所有边权重取值之和.

以图3(a)(b)为例说明WLKA算法和HVKA算法在匿名过程中造成的信息损失率.对图3(b)所示的 $G_{W_1}$ 进行WLKA匿名后得到如图5(a)所示的 $G'_{W_1}$.权重变化导致的信息损失率:$IL(G_{W_1}, G'_{W_1}) = 26/36 = 0.72$.对图3(b)所示的 $G_{W_1}$ 进行HVKA匿名后得到如图6(a)所示的 $G'_{W_1}$.权重变化导致的信息损失率:$IL(G_{W_1}, G'_{W_1}) = 4/36 = 0.11$.可以看出HVKA匿名带来的信息损失比WLKA匿名带来的信息损失小,根据定理2可知,经HVKA得到的匿名图更好地保留了原图的性质,可用性高.

## 5 性能分析与评价

本节对WLKA和HVKA隐私保护算法进行性能分析和评价,采用真实社会网络数据集Cond-Mat-2005①和Astro-Ph②进行实验分析和测试,数据集的描述如表1和表2所示.

数据集Cond-Mat-2005为科研人员在凝聚态物理研究领域合作发表论文的加权网络,分别获取该网络在1995-12-03, 1998-03-03, 2001-04-03, 2005-05-18的图数据,作为加权图增量序列的时刻 $t=0,1,2,3$ 的社会网络图.表1展示了每一时刻增加的结点和边的数目,总共包含了40 421个结点和175 692条边.每条边的权重表示科研人员之间合作发表论文的数目[13],Cond-Mat-2005的平均边权重值为0.28.

Astro-Ph描述了天体物理学研究领域中科研人员之间的合作发表论文的加权网络,表2所示的时刻 $t=0,1$ 的社会网络图是该网络在1998-11-06和1999-02-01的图数据,总共包含了16 706个结点和121 251条边.Astro-Ph的平均边权重值为0.51.

**Table 1 Cond-Mat-2005 Datasets Description**

**表 1 数据集 Cond-Mat-2005 描述**

| $t$ | Timestamp | Nodes | Edges | Nodes Added | Edges Added |
|---|---|---|---|---|---|
| 0 | 1995-12-03 | 21 224 | 106 758 | 21 224 | 106 758 |
| 1 | 1998-03-03 | 26 753 | 128 684 | 5 529 | 21 926 |
| 2 | 2001-04-03 | 36 306 | 164 143 | 9 553 | 35 459 |
| 3 | 2005-05-18 | 40 421 | 175 692 | 4 115 | 11 549 |
| | Total | 124 704 | 575 277 | 40 421 | 175 692 |

**Table 2 Astro-Ph Datasets Description**

**表 2 数据集 Astro-Ph 描述**

| $t$ | Timestamp | Nodes | Edges | Nodes Added | Edges Added |
|---|---|---|---|---|---|
| 0 | 1998-11-06 | 11 434 | 99 804 | 11 434 | 99 804 |
| 1 | 1999-02-01 | 16 706 | 121 251 | 5 272 | 21 447 |
| | Total | 28 140 | 221 055 | 16 706 | 121 251 |

由于数据集Cond-Mat-2005和Astro-Ph都不含标签,所以本文中人工生成了标签和权重,所有值满足统一分布.其中标签含有3个属性:年龄(10～60)、性别(男或女)以及位置(50个国家).实验测试的软硬件环境如下:1)硬件环境. Intel Core 2 Quad 2.66 GHz CPU,4 GB DRAM内存. 2)操作系统平台. Microsoft Windows 7. 3)编程环境. Java,Eclipse.

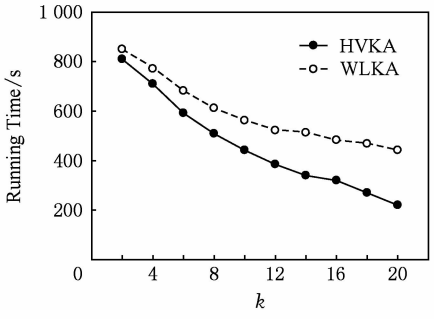
### 5.1 执行时间

图7显示了在不同数据集上,WLKA和HVKA算法的执行时间随 $k$ 值大小的变化情况. 从实验结果可以看出,HVKA算法的执行时间小于WLKA算法,这是因为WLKA算法要单独为每一时刻进行权重链表统一化,而HVKA算法则只需在时刻 $t=T$ 对结点和边进行处理,这就大大减少了匿名边权重造成的时间代价.随着 $k$ 值的增大,WLKA和HVKA算法的执行时间均逐渐减少,这是由于 $k$ 越

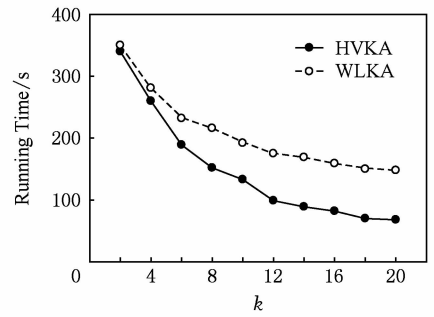
① http://www.cise.ufl.edu/research/sparse/matrices/Newman/cond-mat-2005

② https://snap.stanford.edu/data/ca-AstroPh.html

大分组数量越少,验证是否符合 ISCSC 的次数也越少. 但 HVKA 算法的执行时间始终小于 WLKA 算法的执行时间.




(a) Cond-Mat-2005




(b) Astro-Ph

Fig. 7 Running time for different $k$ values.

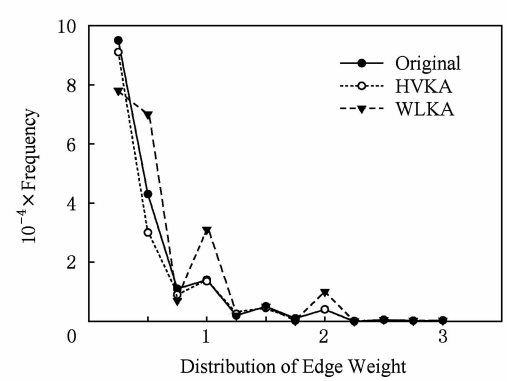
图 7 执行时间随 $k$ 值的变化

## 5.2 数据可用性

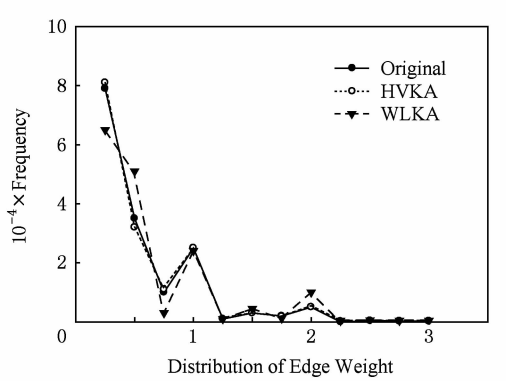
首先考虑匿名权重对图数据可用性的影响,即发布的匿名加权图增量序列中边权重信息的保持程度.

图 8 分别显示了在 2 个数据集上不同时刻边权重取值的分布情况. 对于 HVKA 算法生成的匿名序列中与结点相连的边及边权重都是不确定的,对此采用抽样一致图方法随机抽取与匿名图一致的图,然后在此抽样图上进行分析. 主要的做法是为每个结点分配候选边及边权重得到抽样图,然后统计抽样图的权重分布,重复这个步骤得到多个抽样图的权重分布,最后取这些权重分布的平均值作为最终结果. 当 $k=20$ 时,在图 8(a)~(f)中,经过 HVKA 算法生成的匿名图和原始图的权重分布十分接近,WLKA 算法生成的匿名图与原始图有较大的偏差. 当 $k=50$ 时,在图 8(g)(h)中,经过 HVKA 算法生成的匿名图和原始图的权重分布仍然非常接近,而 WLKA 算法生成的匿名图与原始图有了更大的偏差.

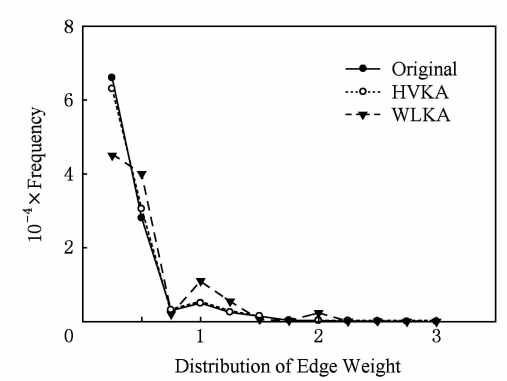
本文利用单跳查询来评估匿名加权图增量序列的可用性. 单跳查询涉及同一周期中一对不同属性的结点,形式化描述为:在一个时间周期具有某一属性的结点和另一属性结点间存在多少连接. 例如,



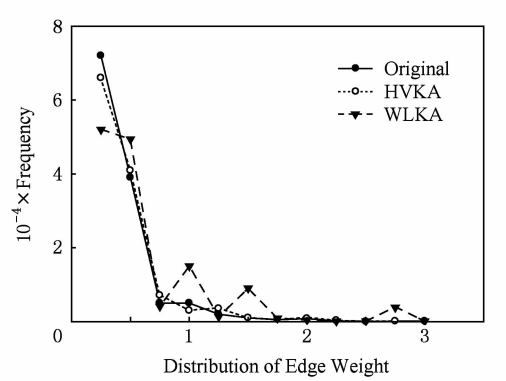

(a) Cond-Mat-2005, $t=3, k=20$



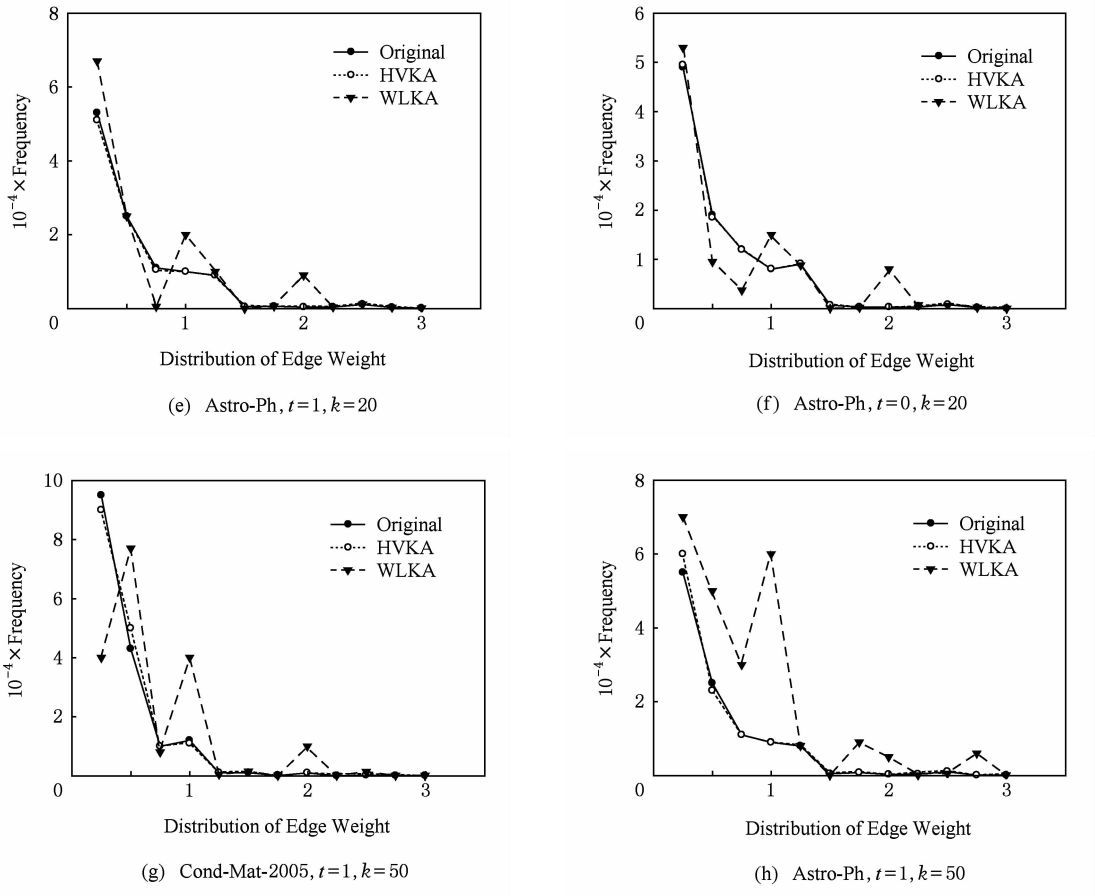

(b) Cond-Mat-2005, $t=2, k=20$




(c) Cond-Mat-2005, $t=1, k=20$




(d) Cond-Mat-2005, $t=0, k=20$




(e) Astro-Ph, $t=1, k=20$ (f) Astro-Ph, $t=0, k=20$

(g) Cond-Mat-2005, $t=1, k=50$ (h) Astro-Ph, $t=1, k=50$

Fig. 8 Distribution of edge weight and frequency.

图 8 边权重与频率分布情况

在一个测量周期,美国的用户和日本的用户间存在多少朋友关系. 本文通过计算查询结果的平均相对误差来度量匿名方法的可用性. 定义相对误差为

$$R=\frac{q-q'}{q},$$

其中,$q$ 是在原始图上的查询结果,$q'$是匿名图上的查询结果. 由于匿名图中结点标签是不确定的,所以无法在匿名图中执行查询操作,对此同样使用抽样一致图方法随机抽取与匿名图一致的图,然后在此抽样图上进行查询.

图 9 展示了在不同的属性优先队列下,对 WLKA 算法在 2 个数据集上得到的匿名序列进行单跳查询所得结果的相对误差随 $k$ 值的变化情况. 从图 9 的 4 幅图中均可看出,$k$ 越大,每一时刻单跳查询的相对误差就越大,这是由于随着 $k$ 的增加,结点的候选标签数随之增加,从而造成了查询结果的相对误差越来越大. 图 9(a)为没有考虑属性优先队列的情况,所有时间戳的平均相对误差都超过3.2%. 图 9(b)则为匿名前按属性优先队列 $list_{ap}=\langle location, gender, age\rangle$(LGA)对结点进行排序的情况,平均相对误差降到 0.4%以下,这是因为排序后同一分组中结点具有相似的属性,从而导致单跳查询的相对误差变小. 属性优先队列对数据集 Astro-Ph 的影响如图 9(c)(d)所示,再次验证了属性优先队列对减小查询结果相对误差的重要性.

图 10 为对 2 个数据集的 HVKA 匿名序列进行单跳查询所得结果的相对误差随 $k$ 值的变化情况. 对比图 9(b)和图 10(a)、图 9(d)和图 10(b)可以看出,HVKA 匿名序列单跳查询的相对误差要比 WLKA 匿名序列单跳查询的相对误差小. 这是由于单跳查询只考虑不同属性结点间的关联,HVKA 算法的匿名过程中添加的虚拟结点不包含边,所以添加虚拟结点并没有改变匿名序列中边的总数,因此对单跳查询结果的影响并不大;而 WLKA 算法的匿名过程中为了实现权重链表统一化会添加虚拟边,导致单跳查询结果的相对误差增大.

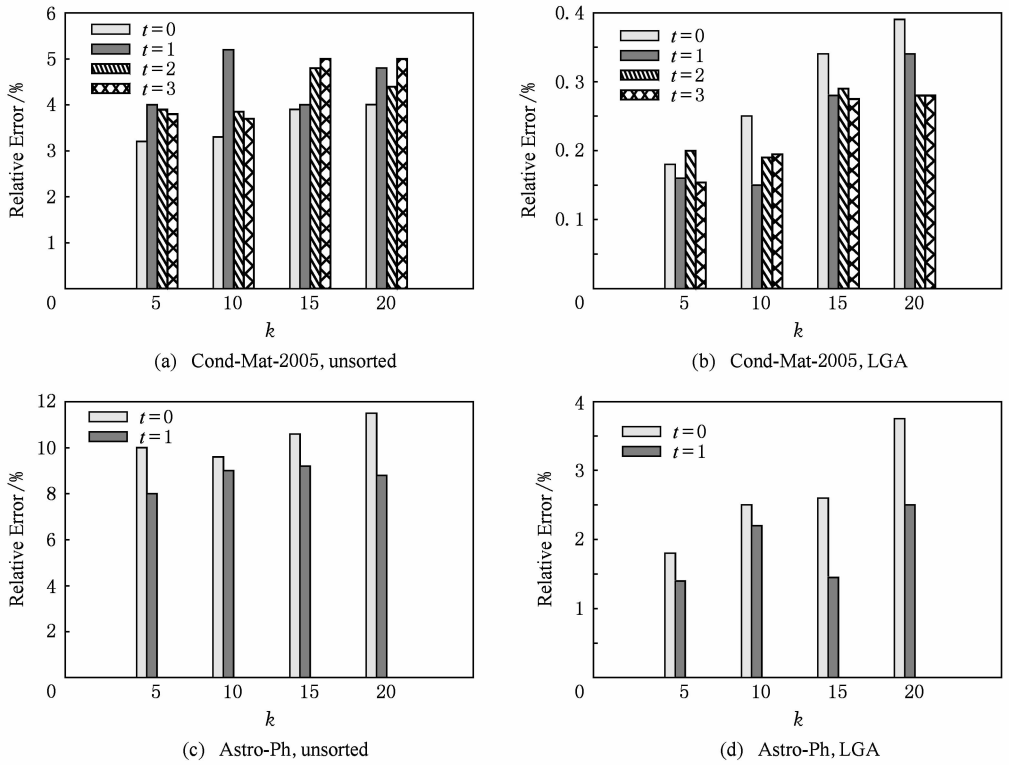



Fig. 9 The relative error of single hop query on WLKA anonymous sequences.

图 9 WLKA 匿名序列单跳查询的相对误差

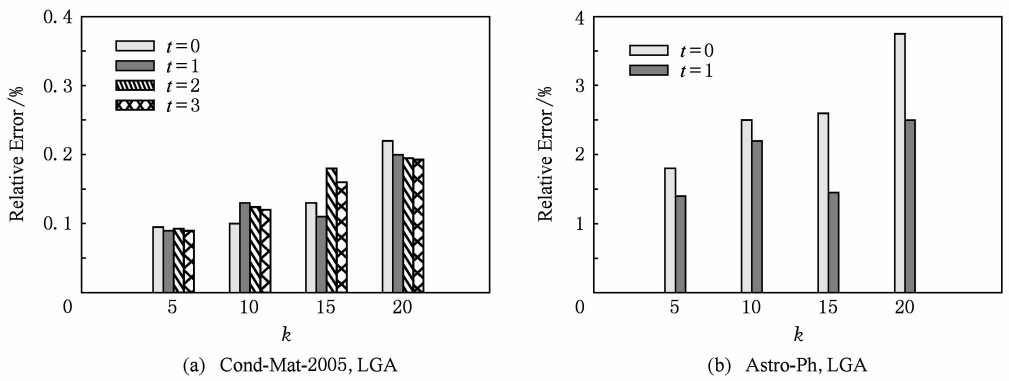



Fig. 10 The relative error of single hop query on HVKA anonymous sequences.

图 10 HVKA 匿名序列单跳查询的相对误差

## 6 结束语

本文研究了加权图增量序列的隐私保护问题.为了匿名加权图增量序列,本文提出了加权图增量序列 $k$-匿名隐私保护模型,并在此基础上设计了基于权重链表的 baseline 算法 WLKA 和基于超图的 $k$-匿名算法 HVKA,从而有效地防止结点身份再识别和边权重泄露等隐私攻击.最后基于大量实验测试和分析,验证了 WLKA 算法能够有效地防止结点身份泄露和边权重泄露;HVKA 算法则在保证了加权图增量序列隐私保护有效性的基础上提高了匿名图的可用性,同时也降低了匿名过程的时间代价.

## 参 考 文 献

[1] Wang C J L, Wang E T, Chen A L P. Anonymization for Multiple Released Social Network Graphs [M] //Advances in Knowledge Discovery and Data Mining. Berlin: Springer, 2013: 99-110

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

[3] Zhou B, Pei J. Preserving privacy in social networks against neighborhood attacks [C] //Proc of the 24th Int Conf on Data Engineering. Piscataway, NJ: IEEE, 2008: 506-515

[4] Zou L, Chen L, Özsu M T. K-automorphism: A general framework for privacy preserving network publication [J]. The VLDB Journal, 2009, 2(1): 946-957

[5] Cheng J, Fu A W, Liu J. K-isomorphism: Privacy preserving network publication against structural attacks [C] //Proc of the 2010 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2010: 459-470

[6] Yuan M, Chen L. Node protection in weighted social networks [C] //Proc of Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2011: 123-137

[7] Li Y, Shen H. Anonymizing graphs against weight-based attacks [C] //Proc of the Int Conf on Data Mining Workshops (ICDMW). Piscataway, NJ: IEEE, 2010: 491-498

[8] Liu X, Yang X. A generalization based approach for anonymizing weighted social network graphs [C] //Proc of Int Conf on Web-Age Information Management. Berlin: Springer, 2011: 118-130

[9] Yuan M, Chen L, Yu P S. Personalized privacy protection in social networks [J]. The VLDB Journal, 2010, 4(2): 141-150

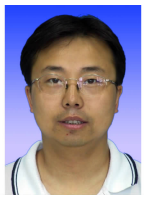
[10] Chen Ke, Liu Xiangyu, Wang Bin, et al. Anonymizing weighted social network graph against path-based attacks [J]. Journal of Frontiers of Computer Science and Technology, 2013, 7(11): 961-972 (in Chinese)
(陈可, 刘向宇, 王斌, 等. 防止路径攻击的加权社会网络匿名化技术[J]. 计算机科学与探索, 2013, 7(11): 961-972)

[11] Zhu Huaijie, Wang Jiaying, Wang Bin, et al. Location privacy preserving obstructed nearest neighbor queries [J]. Journal of Computer Research and Development, 2014, 51(1): 115-125 (in Chinese)
(朱怀杰, 王佳英, 王斌, 等. 障碍空间中保持位置隐私的最近邻查询方法[J]. 计算机研究与发展, 2014, 51(1): 115-125)

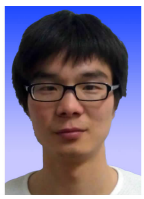
[12] Bhagat S, Cormode G, Krishnamurthy B, et al. Class-based graph anonymization for social network data [J]. The VLDB Journal, 2009, 2(1): 766-777

[13] Newman M E J. The structure of scientific collaboration networks [J]. Journal of the National Academy of Sciences, 2001, 98(2): 404-409

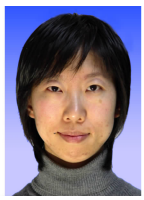
**Guo Caihua**, born in 1990. Master. Her main research interests include social network privacy protect.



**Wang Bin**, born in 1972. Associate professor. His main research interests include query processing and optimization, and distributed data management.



**Zhu Huaijie**, born in 1988. PhD candidate. Student member of China Computer Federation. His main research interests include spatial database and management, location privacy.



**Yang Xiaochun**, born in 1973. Professor and PhD supervisor. Senior member of China Computer Federation. Her main research interests include database technique and theory, and data quality analysis.