软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(4):768−780 [doi: 10.13328/j.cnki.jos.004565]

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

基于节点分割的社交网络属性隐私保护*

付艳艳[1], 张 敏[1], 冯登国[1], 陈开渠[2]

[1](中国科学院 软件研究所 可信计算与信息保证实验室,北京 100190)

[2](国家超级计算深圳中心(深圳云计算中心),广东 深圳 518055)

通讯作者: 付艳艳, E-mail: fuyy@tca.iscas.ac.cn, http://tca.iscas.ac.cn

摘 要: 现有研究表明,社交网络中用户的社交结构信息和非敏感属性信息均会增加用户隐私属性泄露的风险.针对当前社交网络隐私属性匿名算法中存在的缺乏合理模型、属性分布特征扰动大、忽视社交结构和非敏感属性对敏感属性分布的影响等弱点,提出一种基于节点分割的隐私属性匿名算法.该算法通过分割节点的属性连接和社交连接,提高了节点的匿名性,降低了用户隐私属性泄露的风险.此外,量化了社交结构信息对属性分布的影响,根据属性相关程度进行节点的属性分割,能够很好地保持属性分布特征,保证数据可用性.实验结果表明,该算法能够在保证数据可用性的同时,有效抵抗隐私属性泄露.

关键词: 社交网络;属性隐私;匿名;节点分割

中图法分类号: TP309 **文献标识码**: A

中文引用格式: 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

英文引用格式: 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). http://www.jos.org.cn/1000-9825/4565.htm

Attribute Privacy Preservation in Social Networks Based on Node Anatomy

FU Yan-Yan[1], ZHANG Min[1], FENG Deng-Guo[1], CHEN Kai-Qu[2]

[1](Laboratory of TCA, Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

[2](National Supercomputing Center in Shenzhen (Shenzhen Cloud Computing Center), Shenzhen 518055, China)

Corresponding author: FU Yan-Yan, E-mail: fuyy@tca.iscas.ac.cn, http://tca.iscas.ac.cn

Abstract: Recent research shows that social structures or non-sensitive attributes of users can increase risks of user sensitive attribute disclosure in social networks. Most of the existing private attribute anonymization schemes have many defects, such as lack of proper model, too much distortion on attributes distribution, neglect social structure and non-sensitive attributes' influence on sensitive attributes. In this paper, an attribute privacy preservation scheme based on node anatomy is proposed. It allocates original node's attribute links and social links to new nodes to improve original node's anonymity, thus protects user from sensitive attribute disclosure. Meanwhile, it measures social structure influence on attribute distribution, and splits attributes according to attributes' correlations. Experimental results show that the proposed scheme can maintain high data utility and resist private attribute disclosure.

Key words: social network; attribute privacy; anonymity; node anatomy

随着社交网络(social network,简称 SNS)的日益发展,越来越多的个人信息被网络记录储存下来.用户主动或者被动提交的好友互动记录、兴趣爱好标签、签到信息、消费记录等包含了大量社交结构信息和属性信息,为定向广告、推荐系统等应用提供了丰富的数据来源.用户的需求、喜好、属性、行为以及可能具有的关系等,能够被尽可能详细地加以刻画[1−3].但是,随着用户网络形象的进一步丰富,能够用于确定用户真实身份的信息

* 基金项目: 国家自然科学基金(61232005, 61100237); 深圳市战略新兴产业发展专项资金(CXZZ20120831113048965)
收稿时间: 2013-09-10; 定稿时间: 2013-12-18

也越来越多,用户隐私泄露的隐忧也日益严重[4,5].其中,尤其值得关注的是属性-社交网络场景中用户敏感属性信息泄露的问题.2011 年,ACM Recommender Systems 大会发布的 Last.FM 数据集即为典型的属性-社交网络数据,其中不但包含用户和朋友间形成的社交结构,还包括用户的音乐播放列表、听过的音乐家、音乐标签等个人属性信息.这些社交结构信息和属性信息不但可能导致用户身份泄露[5],还可能进一步导致用户敏感属性信息的泄露.例如,人们可根据用户频繁收听音乐的时间段和音乐类型,推测用户的作息时间、宗教信仰(如是否经常听佛经音乐)等,还可以根据属性关联特征或聚类分析等对用户可能具有的其他属性进行预测[6,7].由于社交关系的影响,未标记隐私属性的用户也可能被推测出可能具有某隐私属性[3,8−10],例如,虽然某用户没有公开其宗教信仰,但其所有朋友都喜欢听佛教音乐,则可以较大概率地推测其喜欢此类音乐.因此,属性-社交网络的数据在发布时,必须同时考虑到社交结构信息对属性分布的影响以及属性分布自身具有的特征,才能更好地实现属性隐私保护的目标.

在实现隐私保护的同时,还需要注意避免过度匿名的问题,以免对数据可用性造成不必要的损失,丢失属性数据分布特征.尤其是在属性-社交网络中,考虑到用户属性丰富而稀疏的特点,希望数据匿名时尽量减少属性扰动,从而有利于挖掘属性分布规律,有效预测用户属性,并进行推荐.在针对 Google+数据的研究中发现[9],接近 70%的用户没有显式标注属性,而在标注了属性的用户中,拥有>4 个属性的用户不超过 5%.这表明,社交网络中的属性分布是相当稀疏的.为了保持数据的可用性,匿名结果应尽量保证属性分布稳定.

本文主要关注对属性-社交网络中属性匿名保护的研究,以及对隐藏属性(隐私属性)进行恢复和预测的相关研究.现有的匿名方案包括基于 K 匿名的方案和基于等价类泛化的方案:

- 基于关系数据发布的 K 匿名模型或者更进一步优化的 L 多样化模型的方案通过将属性分布均匀化,实现具有相同度数的用户具有不同属性,从而实现属性的匿名.其中,Yuan 等人[11]提出了一种基于边增删和添加噪声节点的方案,通过调整节点度数,实现度数的 K 匿名,并通过赋予噪声节点不同的敏感属性值,调整敏感属性的出现次数,从而实现属性匿名;基于相同思想,Sun[12]和 Zhou[13]分别提出了不同的属性调整方案;
- 基于等价类的匿名方法将局部区域的用户属性归入同一等价类,针对不同的等价类,分别发布其泛化属性[14].

对属性预测的研究则充分挖掘了属性-社交网络中社交结构对属性的影响.Zheleva[3]和 Mislove[2]分别提出通过用户朋友具有的属性、用户所加入的群组等属性推测用户可能具有的属性的方案.Yin 等人[8]首先提出属性-社交网络的概念,对社交连接和属性间的关系进行了分析,并利用属性特征对用户间能否形成社交关系进行预测.Gong[9]和 Yang[10]等人在此工作的基础上,对属性-社交网络中的用户属性推测作了进一步分析.

通过比较以上属性-社交网络中的属性匿名和推测方案可以看出:现有的属性匿名方案明显缺乏对属性分布与社交结构之间关联的扰动,因此难以抵抗攻击者基于属性联合分布特征或基于社交结构对用户的隐私属性进行推测攻击.而同时,现有匿名方案还存在过度匿名的问题.在现实社交网络中,同一属性的取值是丰富多样的,也可能仅有部分取值是敏感的.而现有方案模型则认为,敏感属性的任何取值均具有相同的敏感性[11,13].对同一属性的所有取值都进行匿名,既不合理,也不必要.

针对社交网络用户属性隐私保护中存在的上述问题,本文提出了一种依据属性分布特征进行节点分割和社交关系分割,进而实现具有隐私属性的用户的身份匿名、属性匿名的隐私保护算法.本文的主要贡献在于:

1) 针对社交结构对属性分布的影响,提出了属性分布矩阵的概念,将社交结构对属性分布造成的影响进行量化,并据此对属性相关性进行精确度量;
2) 基于属性分布的相关性,提出了基于属性分割的节点分割算法,尽可能地保持属性分布的相关性特征,并提高了具有隐私属性的用户节点的匿名性;
3) 将属性匿名的对象由属性精确到某些敏感的属性取值,减少对非隐私属性的扰动和对不具有隐私属性取值的社交群体的扰动.

1 数据模型与定义

属性-社交网络模型是本文工作的基础.在常见的社交网络结构模型 $G(V,E)$中,仅仅考虑了用户节点 V 和节点间的连接关系 E.但在大部分社交网络中,除了结构化数据,每个用户还具有丰富的属性数据,简单的网络结构模型并不能满足社交网络属性匿名分析的需求.为此,我们引入了 Yin 等人提出的属性-社交网络模型[10],并对其进行补充定义和说明.

定义 1(属性-社交网络). 我们将包含用户属性和用户间相互关系的社交网络以属性-社交网络模型描述.形式化的定义如下所示:

$$G=\{VU,VA,EU,EA\},$$

其中,

- VU 为用户节点集合,其中的每一个节点对应于社交网络中的一个真实用户;
- VA 为属性节点所有可能取值的集合,其中的每一个节点对应于社交网络中的一种用户属性取值.例如,对于公司属性,Google 和 Microsoft 为两个不同的属性取值,形成两个不同的属性节点;
- EU 为社交连接关系集合,其中的每一条边对应于用户间的一条社交连接;
- EA 为属性对应关系集合,其中的每一条边对应于用户和属性间的一种对应.

图 1 给出了属性-社交网络的示意图.

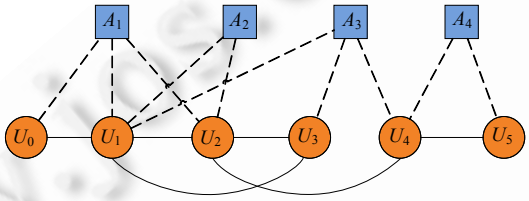


Fig.1 Social-Attribute network

图 1 属性-社交网络示意图

注:●类节点为用户节点,■类节点为属性节点,实线代表社交连接,虚线代表属性连接.

定义 2(局部属性分布矩阵). 为了分析区域内用户属性间的相互关系,我们将子图范围内属性间相互关联的次数以矩阵的形式表示出来.令 Su 表示属性分布矩阵所依赖的属性-社交网络局部子图,其中,包含所有与用户 u 距离小于等于 1 hop 的用户节点、这些用户节点所关联的属性节点和所有对应连接关系,则其局部属性分布矩阵的形式化定义如下:

$$M_{Su}=[W\langle i,j\rangle]_{n\times n},$$

其中,$W\langle i,j\rangle$为 Su 中两属性 i,j 的连接次数.如果某用户同时具有属性 i 和属性 j,那么,计属性 i,j 连接次数 1;如果属性 i,j 通过社交连接$\langle u,v\rangle$相连,那么计属性 i,j 连接次数为 $\frac{1}{\max(D(u),D(v))}$,其中,$D(u),D(v)$分别为节点 u,v 的社交连接度数,同一社交连接对相同属性连接的贡献仅统计 1 次.

在计算属性分布矩阵时,在整个系统范围内计数属性间的相互连接次数,即得到系统属性分布矩阵.

图 2(a)给出了图 1 所示用户节点 U_1 所对应的子图 SU_1,图 2(b)是 SU_1 的局部属性分布矩阵.

定义 3(属性相关系数). 在属性分布矩阵中,我们将两个属性对应的列的相关系数定义为两个属性节点的相关系数,其定义如下:

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}.$$

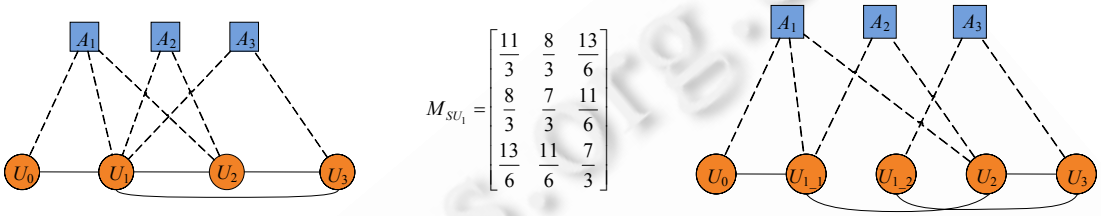
定义 4(节点分割). 为了对用户的敏感属性进行匿名,并避免攻击者通过用户局部区域内的社交连接或者属性连接推测某些敏感属性的存在,我们提出了节点分割的概念,其形式化的描述如下所示:

$$Anatomy(u,Su)\rightarrow\{u_1,Su_1\}\cup\{u_2,Su_2\},$$

其中,

- u 是需要进行分割的用户节点;
- Su 为该节点依赖的属性-社交网络局部子图;
- u_1 和 u_2 为节点分割后形成的两个新节点,用以取代原有用户节点;
- Su_1 和 Su_2 为新节点分别形成的新的属性和连接组成的新的局部结构,其中,原节点的属性按照属性相关度强弱分别分割到两个节点的属性集合中.u_1 的社交邻居集合为所有与其具有共同属性较多的原节点的部分社交邻居集合,u_2 的社交邻居集合为原节点的其他社交邻居集合.

假设图 2(a)中的属性 $A1$,$A2$ 为敏感属性,那么可以根据相关参数对节点 U_1 进行初步分割,得到新的局部结构 SU_1',如图 2(c)所示.需要说明的是,图 2(c)中的节点分割结果并不是最终结果,如果继续对其他带有敏感属性的节点进行分割,得到的结构图仍可能发生进一步变化.



(a) 用户 U_1 对应的 SU_1 (b) 用户 U_1 对应的局部属性分布矩阵 (c) 节点 U_1 的初步节点分割结果

Fig.2 Related structures and matrix for user U_1

图 2 用户节点 U_1 所对应的子图结构及相关信息

2 基于节点分割的属性隐私保护方案

2.1 方案原理

本方案旨在针对社交结构对属性信息的影响以及属性分布特征,进行深入的隐私属性匿名;同时,考虑到社交网络中用户属性丰富而稀疏的特点,尽可能地保留属性数据的特征,以保证数据的可用性.

为了实现隐私属性的匿名,必须考虑与隐私属性分布相关的两大因素:(1) 社交结构与隐私属性间的相关性;(2) 隐私属性与非隐私属性间的相关性.前者要求在属性匿名时,应对相关社交结构对属性分布带来的影响进行衡量,避免攻击者根据社交结构对用户可能具有的属性进行推测;后者要求为了实现隐私属性的匿名,必须对用户具有的其他非隐私属性进行清理,避免攻击者依据属性间的相关性对用户的隐私属性进行推测.

此外,从保护数据的可用性角度出发,要求匿名方案能够充分保护用户的社交结构,保护属性间的相关性,以便未来进行数据的挖掘利用.尤其是考虑到属性-社交网络中用户属性稀疏的特点,在可能的情况下,用户具有的属性连接都应该得到保留.

结合属性-社交网络模型分析可以得出,实现隐私属性的匿名有两种基本方式:1) 保持用户身份,扰动用户具有的隐私属性;2) 保持隐私属性,匿名具有此属性的用户身份.为了保护数据可用性,要求尽量保持图中的属性节点 VA、用户社交连接 EU、用户属性连接 EA 的各项分布特征基本不变.如果选择方式 1)进行隐私属性匿名,那么,或者将造成隐私属性缺失(简单删除)、相关属性的分布特征丢失(全局范围内扰动隐私属性);或者无法降低属性被推测概率(在用户邻居范围内扰动隐私属性).因此,本文采用第 2 种方式进行属性匿名.从匿名需求的分析中得出:匿名方案必须实现社交连接 EU 和属性 VA 的分割,才能减弱用户社交结构和非隐私属性两者对隐私属性的影响.因此,结合匿名具有隐私属性的用户身份的需求,本算法通过对具有隐私属性的用户节点 VU 进行分割,形成两个新的独立节点,并根据属性间的相关性对原节点的属性进行对应分割.新节点分别继承原节点的部分属性连接和社交连接,从而实现 EU 和 VA 的分割,进而实现隐私属性 VA 的匿名,并保证 VA,EU,EA 的

分布特征稳定.

同时,为了抵抗攻击者基于背景知识的攻击,本方案的匿名属性集可包含多个部分:敏感属性值、非敏感属性值、识别度高的属性值等等.这样,即使攻击者可以确定攻击目标位于匿名图中,并被分割,攻击者也无法继续确定攻击目标被分割的具体原因.

2.2 方案描述

基于属性-社交网络模型,本方案通过节点分割实现隐私属性相关的社交连接关系和属性连接关系的分割清理,从而实现隐私属性的匿名.本方案的具体步骤描述如下:

步骤 1. 生成隐私属性匿名需求节点序列.

本方案首先对系统中存在的隐私属性进行统计,并对包含隐私属性的节点进行排序,形成优先级队列.节点的优先级计算依据以下原则:1) 节点具有的隐私属性越多,优先级越高;2) 节点的社交邻居节点具有的隐私属性连接越多,该节点优先级越高;3) 对于无法按照原则 1)、原则 2)排序的节点,在其集合所处的优先级范围内随机排序.节点优先级越高,表明其需要匿名的属性越多,或者与隐私属性的关联越多.因此,在匿名过程中尤其需要关注其隐私性和数据可用性.

步骤 2. 序列节点依次分割.

按照优先级队列顺序进行节点分割,并由生成的新节点分别继承原节点对应的属性关系和社交关系.整个分割过程如算法 1 所示.

算法 1. 属性匿名 *AttributeAnonymize*.

输入:属性-社交网络 G,节点优先级队列 L,匿名属性集合 P;

输出:节点分割结果集合.

1) *Anonymizednodes*; //建立已匿名节点集合
2) For each User-node u in L //按照节点优先级队列,依次处理节点
3) $Su \leftarrow Local(G,u)$; //获得节点 u 对应的局部子图 Su
4) $M_{Su} \leftarrow Attmatrix(Su)$; //生成节点 u 的局部属性分布矩阵
5) $C \leftarrow Corr(M_{Su})$; //根据属性相关性分割非敏感属性关系和社交关系
6) $Su' \leftarrow NodeAnatomy(C,Su,P,Anonymizednodes)$;
7) $L \leftarrow L.update(Su)$; //更新在此轮匿名中被扰动的优先级较低节点
8) $Anonymizednodes \leftarrow u.newnode_1,u.newnode_2$; //更新已匿名节点
9) $G \leftarrow update(G,Su',Su)$; //更新图
10) End For;
11) Return G;

如算法 1 所示,节点分割过程包括 3 个部分:1) 节点局部属性分布矩阵生成;2) 节点属性连接分割;3) 节点社交关系分割.下面分别介绍这 3 个部分的处理过程.

步骤 2.1. 节点局部属性分布矩阵生成.

在本步操作中,需要统计当前节点及所有社交邻居节点的所有属性,以及这些属性节点间的连接关系,并计算连接关系的权重.本算法首先统计局部子图中各个节点自身的属性和属性连接,然后统计基于社交连接关系形成的新的属性连接.算法 2 中给出了计算过程说明.

步骤 2.2. 节点属性连接分割.

在本步操作中,根据属性分布矩阵得出属性相关度,为所有需要分割的节点生成两个新节点,并根据属性相关度将原节点的属性分割成两组,分别赋予两个新节点.

步骤 2.3. 节点社交关系分割.

在本步操作中,将原节点与其他节点的社交连接进行分割,分别由新节点继承.社交关系继承时,需考虑当前两个新节点与原社交邻居节点分别的共同属性个数,由共同属性多的节点继承原有社交连接,共同属性少的

新节点与原邻居节点不再具有社交连接.

步骤 2.2 和步骤 2.3 为节点分割算法的实际分割过程,算法 3 中给出了具体说明.其中,单个节点分割的复杂度为 $O(A+E)$,A 为单个节点的最大属性个数,E 为单个节点的最大社交连接个数.而整个属性匿名过程包括 N 个相关节点的节点分割,因此,整个过程的复杂度为 $O(N\times(E+A))$.

算法 2. 局部属性分布矩阵生成 *Attmatrix*.

输入:局部子图 *Su*;

输出:局部属性分布矩阵 M_{su}.

1) $n\leftarrow$*distinctatt*(*Su*); //*Su* 中属性的个数
2) $M_{Su}\leftarrow$*Matrix*(n); //建立属性分布矩阵
3) For each User-node v in *Su* //处理单个节点的属性及属性连接
4) *Addatt*(v.*attributes*,M_{Su},1);
5) End For;
6) For each edge e in *Su* //处理由边连接形成的属性连接
7) $M_{Su}\leftarrow$*Addatt*($e.node_1$.*attributes*,$e.node_2$.*attribtues*,1/max($e.node_1$.*degree*,$e.node_2$.*degree*));
8) End For;
9) Return M_{Su};

算法 3. 节点分割 *NodeAnatomy*.

输入:属性-社交网络局部子图 *Su*,局部区域内属性相关性矩阵 C,匿名属性集合 P,相关度阈值α,已匿名节点集合 *Anonymizednodes*;

输出:新的局部子图.

1) For each User-node v in *Su* //节点属性连接分割
2) IF (v in *Anonymizednodes*) //优先级高节点的分割结果
3) ; ///不作分割,完全继承
4) ELSE
5) IF exist v.*attribute* $a\rightarrow|C(a,P)|>\alpha$ && $|v.attribute|>1$
6) //如果该节点某属性与某敏感属性相关度较高,且多于 1 个属性,则可以进行分割
7) $v_1,v_2\leftarrow$new *node*(); //为当前节点生成对应的新节点;
8) For each attribute b of v //分割节点的属性关系
9) IF max$|C(b,v_1.attributes)|<\alpha$ && max$|C(b,v_2.attributes)|<\alpha$ //与两点相关性均不强
10) *Distributebynum*(b,v_1,v_2); //将 b 分配到属性较少的节点
11) IF max$|C(b,v_1.attributes)|>$max$|C(b,v_2.attributes)|$ //与点 v_1 相关性更强
12) *Distributebycor*($b,v_1,v_2,C(b,v_1.attributes)$) //根据相关性的正负,分配属性
13) ELSE //与点 v_2 相关性更强
14) *Distributebycor*($b,v_1,v_2,C(b,v_2.attributes)$) //根据相关性的正负,分配属性
15) End For
16) IF ($v_1.attributes$!=*null* && $v_2.attributes$!=*null*)
17) *AttributeEdges*(v_1,v_2); $v.newnode\leftarrow v_1,v_2$;
18) ELSE
19) IF (exists v.*attribute* a in p) //包含敏感属性,强制分割
20) *Randomize*($v.attributes,v_1,v_2$); *AttributeEdges*(v_1,v_2); $v.newnode\leftarrow v_1,v_2$;
21) ELSE
22) *Remove*(v_1); *Remove*(v_2); //不作分割,完全继承

```
23)          ELSE
24)              ;                                          //不作分割,完全继承
25) End For
26) For each User-node v in Su                              //节点社交连接分割
27)    IF exist EU⟨v,v′⟩ in Su
28)       int a=|v.newnode|+|v′.nernode|
29)          If (a!=0)
30)             Socialedge(v,v′);                           //按共同属性多少,分别建立新连接
31)             removeedge(v,v′);                           //移除原社交连接
32)    //若子图范围外某节点与被分割节点具有社交连接,该节点与两个新节点中共同属性多的连接;
33)    IF ((v.newnode!=null) && (EU⟨v,v′⟩ not in Su))
34)    Outeredgeupdate(v1,v2,v′);
35) End For;
36) Return Su′;
```

2.3 方案分析

2.3.1 隐私保护程度分析

本节分析可能的隐私属性重识别攻击方法,以及本算法抵抗此类攻击的能力.

攻击者可依据节点邻居个数或属性个数进行用户的重识别,进而实现隐私属性的重识别.本算法的匿名结果可以有效地扩大目标节点的范围,抵抗此类攻击.由于本算法对具有隐私属性的节点或者具有其他与隐私属性密切相关属性的节点均进行了分割,部分节点可能被多次分割,导致新生成节点的邻居个数和属性个数均与原节点不同,攻击者只能确定目标节点邻居个数或属性个数的范围,而这一范围远远大于真实数据的分布范围.根据攻击者背景知识的不同,下面分别分析其攻击能力以及本算法的抵抗能力.

(1) 攻击者了解目标节点的邻居个数为 K.

在节点分割前,攻击者可以将目标范围缩小至具有 K 个邻居的用户节点集合,但在节点分割之后,根据目标节点可能所处位置的不同和可能具有属性的不同,使得目标范围包含以下集合:

(a) 如果目标节点不具有隐私属性,未与其他被分割节点关联,那么所有满足以下条件的用户节点均为可能的目标节点:

$$\{u \mid \{a \mid a \in u.attributes \text{ and } a \in P\} = \varphi \text{ and}$$
$$\{v \mid \langle u,v\rangle \in EU \text{ and } \{a \mid a \in v.attributes \text{ and } a \in P\} \neq \varphi\} = \varphi \text{ and } \mid EU(u) \mid = k\};$$

(b) 如果节点具有隐私属性,并且经过一次节点分割,那么所有满足以下条件的两个节点的组合,均可能还原为目标节点:

$$\{(u_1,u_2) \mid Distance(u_1,u_2) \geqslant 2 \text{ and } u_1.attributes \cap u_2.attributes = \varphi \text{ and}$$
$$\{a \mid a \in u_1.attributes \cup u_2.attributes \text{ and } a \in P\} \neq \varphi \text{ and } \mid EU(u_1) \mid + \mid EU(u_2) \mid \geqslant k\};$$

(c) 如果节点具有隐私属性,并且经历过多次节点分割,那么所有满足以下条件的节点组合,均为可能的目标节点:

$$\{(u_1,u_2,...,u_i,...,u_m) \mid Distance(u_i,u_j) \geqslant 2 \text{ and } \bigcap u_i.attributes = \varphi \text{ and}$$
$$\{a \mid a \in \bigcup u_i.attributes \text{ and } a \in P\} \neq \varphi \text{ and } \sum \mid EU(u_i) \mid \geqslant k\};$$

(d) 如果目标节点不具有隐私属性,但与被分割节点关联,或者具有相关度高的属性,经过一次或多次节点分割,那么所有满足以下条件的节点组合均为可能的目标节点:

$$\{(u_1,u_2,...,u_i,...,u_m) \mid Distance(u_i,u_j) \geqslant 2 \text{ and } \bigcap u_i.attributes = \varphi \text{ and}$$
$$\{a \mid a \in \bigcup u_i.attributes \text{ and } a \in P\} = \varphi \text{ and } \sum \mid EU(u_i) \mid \geqslant k\}.$$

由以上 4 个集合可以看出:即使攻击者很容易确定集合 a 中的节点,集合 b,c,d 中的节点数量却仍然难以进行有效缩减,因此,单独依据节点邻居个数进行攻击是无效的.

同理,单独依据属性个数 J,攻击者也难以对攻击目标范围进行有效限定.

(2) 攻击者同时掌握目标节点的社交连接数目 K 和属性连接数目 J.

攻击者可根据两者的组合发起隐私属性重识别攻击.由于存在多次节点分割的可能性,攻击者仍旧无法缩减以上分析中集合 c 和集合 d 的范围.因此,此类攻击对本算法效果有限.

(3) 攻击者掌握目标节点邻居区域结构的背景知识,了解目标节点连接的其他节点的度数,能够主动与目标节点建立连接关系,发起主动攻击或被动攻击.

由于节点的多次分割,即使攻击者具有匿名前目标节点的邻居区域结构背景知识,也无法确定目标节点匿名后的局部区域状况,因而无法实施目标节点的去匿名化.对于基于度数等价类的攻击方案,由于形成连接关系的两个节点都可能进行多次节点分割,因此其攻击目标范围更大,攻击能力更弱.

(4) 攻击者掌握目标节点具有的若干属性.

由于攻击者无法确定节点分割后该节点原有属性的分割状态,也无法确定该节点可能的分割次数,因此无法有效确定匿名后属性的组合情况,也就无法缩小攻击目标范围,无法进行有效的去匿名化.

结论:上述分析表明,本算法能够抵抗基于社交连接数目或属性数目的单独/组合攻击,也能够抵抗基于邻居结构和基于属性组合的攻击.

2.3.2 数据可用性分析

考虑到属性-社交网络兼具社交连接关系和属性连接关系的特征,本文对于数据可用性的分析分为两个部分:(1) 对社交结构的扰动;(2) 对属性分布的扰动.

在社交结构扰动方面,本文主要衡量了以下参数[15,16]的变化:(1) 社交连接度数分布,用于展示匿名前后节点度数分布的变化;(2) 接近中心势(closeness centralization)[17],用于统计匿名前后图中节点接近程度的变化;(3) 集聚系数(clustering coefficients)[18],用于统计匿名前后图结构的聚集特征变化.这 3 项参数的变动能够充分描述图结构数据的变化程度,其变动越小,数据可用性保持得就越好.

在社交结构扰动方面,本算法在匿名过程中保持了未分割节点间的社交连接数目和分布、分割节点与未分割节点间的社交连接数目,但对于分割节点间的连接数目和分布的扰动相对较大.但考虑到系统的容量,被分割节点占整体节点比例相对较少,因此对系统社交结构造成的改变有限.而且,被分割节点能够分别继承原节点的社交结构,系统特征能够得到较好的保持.

其次,本算法对于属性分布的扰动,也可通过系统全局属性分布的相同参数来衡量.但对于属性连接来说,属性间的相关度还可通过不同属性间连接的次数来表现.因此,我们将增加了边的权重分布特征作为数据可用性的衡量标准.

在属性分布扰动方面,由于扰动集中于少量敏感属性和与敏感属性相关的属性,因此对整个系统的结构并不造成很大破坏.而且在匿名过程中考虑到了属性间的相关性原则,因此,在原始属性分布图中相关性越强的属性间的连接关系受到的扰动越小,属性分布特征也能够得到保持.

但由于本算法仅考虑局部属性间的相关性,而局部属性的相关性可能并不能反映全局属性的特征,因此,本算法可能造成局部范围内的属性分布特征偏离全局属性分布特征.

3 实验结果与分析

实验分别采用了 Google+的数据集和 Last.FM 的数据集.其中,Google+数据集为 2011 年 9 月的 Google+系统的快照,我们随机选取其中 SEP4(每个用户都至少具有 4 个属性取值)的数据进行实验,其中共有 5 200 个用户节点,9 539 个属性节点,7 422 条社交连接,24 690 条属性连接.Last.FM 数据集中具有 1 892 个用户节点,17 632 个属性节点,12 717 条社交连接,92 834 条属性连接.

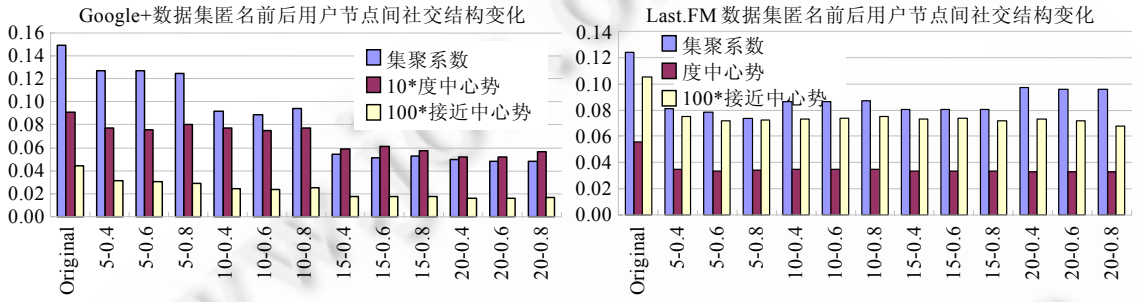
本文在以上两个数据集上分别随机选取了 5%,10%,15%,20%的属性取值作为匿名属性集加以匿名.针对数

据发布者匿名需求的不同,匿名属性集的设计可以包含敏感属性值、无关属性值以及其他识别度较高的属性值等.实验分别考察了两个数据集在匿名前后社交结构和属性分布的变化情况.通过实验分析可以得知,本算法可以很好地满足数据可用性需求.另外,实验还考察了依据节点社交度数进行属性推测的错误率变化情况.通过匿名前后的数据对比可知:匿名后,攻击者通过社交结构确定攻击目标属性的能力大为降低,用户的隐私属性能够得到保护.

3.1 数据可用性分析

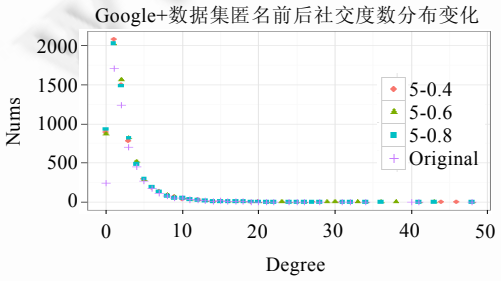
3.1.1 社交结构变化

图 3 分析了 Google+数据集和 Last.FM 数据集匿名前后用户的社交结构变化程度.其中,图 3-1 分析了匿名前后用户社交结构的集聚系数、度中心势、接近中心势的状况(注:X 轴坐标中整数部分表明匿名属性集的比例分别为 5%,10%,15%,20%,小数部分表明属性相关性阈值分别为 0.4,0.6,0.8,Original 项为数据集初始状态),图 3-2 分析了数据集匿名前后的社交度数分布变化(注:分类项中整数部分表明匿名属性集的比例为 5%,小数部分表明属性相关性阈值分别为 0.4,0.6,0.8,Original 项为数据集初始状态.由于 Last.FM 数据集中社交度数差异较大,其中 Original 集中、度数分布于(129,190)区间的 53 个节点未列入图中,度数为 0,1、个数超过 1 000 个的节点也未列入图中).

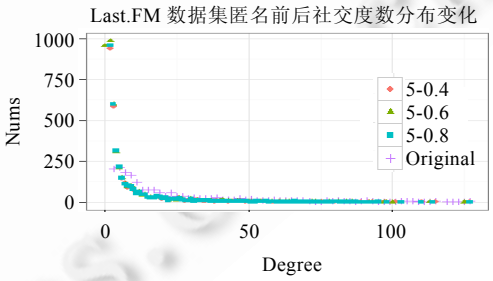


3-1(a) Google+数据集匿名前后用户社交结构变化

3-1(b) Last.FM 数据集匿名前后用户社交结构变化



3-2(a) Google+数据集匿名前后社交度数分布变化



3-2(b) Last.FM 数据集匿名前后社交度数分布变化

Fig.3 Social structure changes through data anonymization

图 3 数据匿名前后社交结构变化

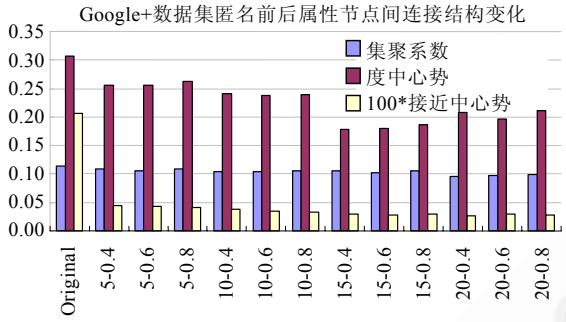
图 3-1 着重分析了匿名属性集的比例和属性相关性阈值的选取与匿名结果可用性的关系,图 3-1(a)与图 3-1(b)分别反映了 Google+数据集与 Last.FM 数据集的情况.由这两图分别可以看出:随着用户选取的匿名属性集比例的扩大,数据的可用性逐渐降低;但对于相同的匿名属性集比例,选取不同的阈值参数对匿名结果的影响不大.表明:在局部区域范围内,属性的相关性较明显;选取不同的阈值,对社交结构分割和属性分割的扰动较小.

由图 3-2 可以看出,本算法能够很好地保持度数分布特征.图 3-2(a)与图 3-2(b)分别反映了 Google+数据集与 Last.FM 数据集的情况.Google+数据集匿名前后,节点社交度数分布的趋势几乎不变.在 Last.FM 数据集中,

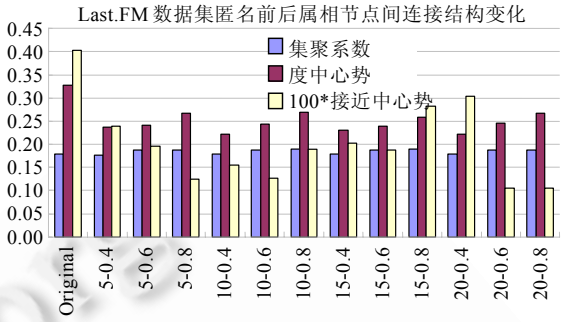
社交度数的分布趋势也能很好地保持一致性.但是由于本算法在节点分割时可能造成新生成的节点部分社交连接丢失,或者由于局部区域内分割的节点较多,造成属性较多的节点社交连接增加.因此,在本算法的匿名结果集中,度数较低和较高的节点比原始数据集都有所增加.但从图中可以看出,本算法的匿名结果仍能够忠实地反映数据集的节点分布趋势.

3.1.2 属性分布变化

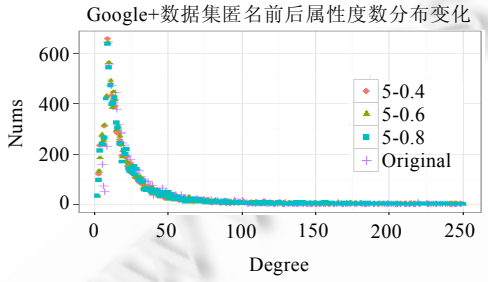
图 4 分析了 Google+数据集和 Last.FM 数据集匿名前后,用户的属性分布变化程度.



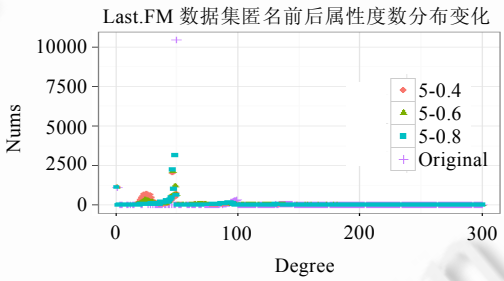
4-1(a) Google+数据集匿名前后属性连接结构变化



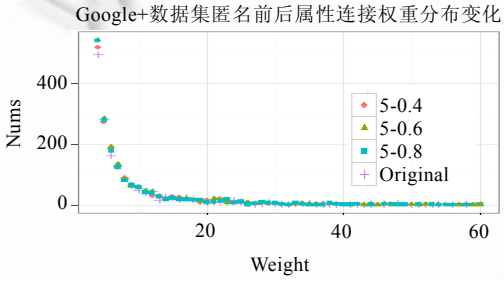
4-1(b) Last.FM 数据集匿名前后属性连接结构变化



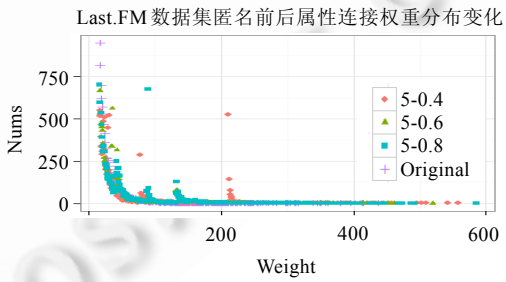
4-2(a) Google+数据集匿名前后属性节点度数变化



4-2(b) Last.FM 数据集匿名前后属性节点度数变化



4-3(a) Google+数据集匿名前后属性连接权重变化



4-3(b) Last.FM 数据集匿名前后属性连接权重变化

Fig.4 Attribute link distribution changes through data anonymization

图 4 数据匿名前后属性分布特征变化

图 4-1 分析了匿名前后,系统的属性分布图中,集聚系数、度中心势、接近中心势的状态变化情况;图 4-2 和图 4-3 分析了匿名前后,属性分布图中,属性节点的度数变化和属性间形成的连接关系的权重变化情况.

由图 4-1 可以看出:系统的属性分布特征中的度数分布和系统集聚特征都能够得到很好的保持.但是,属性分布的接近中心势迅速降低.结合接近中心势的定义分析可知,这表明系统中属性间的距离差异降低,属性间的连接分布更加均匀.

通过图 4-2 中属性节点度数的变化也可以看出:本算法的匿名结果与原始图的属性分布特征是相符的(注:Google+数据集中属性节点的度数最大可大于 1 000 度,但大于 250 度的属性节点很少出现,因此,图 4-2(a)中只显示度数小于 250 度的属性节点;同理,Last.FM 数据集(如图 4-2(b)所示)将观察范围集中于 300 度以内的属性节点).图 4-2 还说明,属性阈值的选取对属性分割的结果影响不大.这表明:在局部区域范围内,用户的属性相关性特征较明显.因此,按照属性相关性进行节点分割也是合理的.由图 4-2 可知,在 Last.FM 原始数据集中,由于绝大多数用户均具有 50 个属性,这使得 Original 数据项在度数为 50 的节点数目上超过了 10 000.但经过节点分割之后,度数为 20~30 或者接近 50 时均出现了一定程度的增加.

图 4-3 分析了属性节点间连接关系的权重在匿名前后的状态.由图中可以看出,属性连接关系的权重保持在原始权重周围较小范围内浮动,与权重分布的整体特征相符.其中,在 Last.FM 数据集中,由于权重较少的连接数大大超过权重大的连接数,因此在图中统计结果仅包含权重>15 的边.权重为 1~15 的边的统计结果如图 5 所示,反映了匿名后每种权重的边数与原始数据集中该权重的边数的差值比例.由图 5 可以看出:在匿名后的属性连接中,权重<15 的边的数量基本呈现降低的特征,因此,匿名后的权重<15 的边数和权重>15 的边数的差值能够部分降低.这表明,本算法的匿名结果使得具有相同权重的边数分布也趋向于平缓.

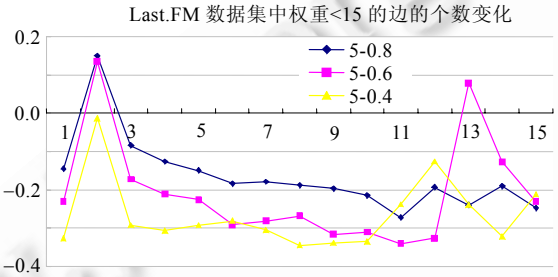


Fig.5 Attribute link changes for links with weight<15 in Last.FM dataset

图 5 Last.FM 数据集中属性连接权重<15 的连接变化程度

3.2 隐私保护程度分析

图 6 给出了 Google+数据集中攻击者通过节点度数和匿名结果中的属性分布,推测用户属性能力的情况.实验中分别选取了原始数据集中度数为 K 的用户群为攻击目标,统计该群组用户的部分属性(至少出现 1 次,且出现频繁程度位于前 50 的属性)在匿名前后占该群组用户的所有属性的比例.依据属性出现的频繁程度,推测攻击者可根据用户社交度数 K 确定攻击目标具有某属性的可能性.

从图 6 可以看出,本算法的匿名结果能够很好地保护用户稀疏群体中属性的匿名性.图 6 中,当用户社交度数分别为 K=15 和 K=10 时,匿名结果明显能够更好地降低属性泄露的风险.从图 3-2 所示的社交度数分布中可以得知:在社交度数越大的群组中,用户的数目越少.由于群组规模较小,其中的用户相对容易识别,因此,用户的身份暴露,进而隐私属性暴露的风险也就越大.本算法通过用户节点的分割,可以有效扩大分割后节点的分布范围,从而增大群组规模,实现用户的隐藏.而且在此过程中,群组中用户的属性规模也得到扩大,隐私属性所占比例降低,被推测的可能性进一步降低.

而当 K 较小时,同一群组中的用户规模较大,识别出用户的难度也很大,在原始数据集中,隐私属性泄露的风险就很小.本算法的匿名结果可提供进一步的扰动,增加群组中属性的种类,降低隐私属性所占比例和被推测的可能性.从图 6 中 K=3 的情况可以看出:当 K 较小时,属性扰动的范围较小;而且由于度数较高节点的分割,可能出现某些属性数量增加的情况,从而导致属性被推测概率上升;尤其是当匿名方案扰动的匿名属性集的比例较大,如图 6 所示的 20%时,扰动中涉及到的属性种类和用户数目增多,可能出现若干属性的特征增强的现象.

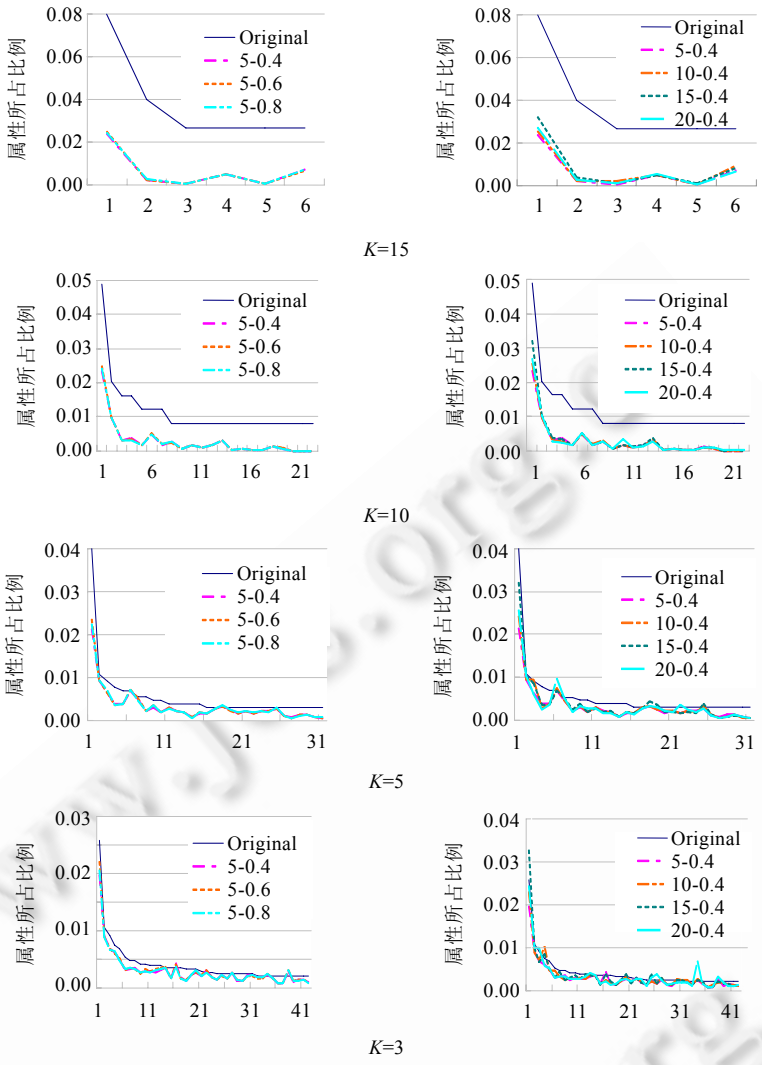


Fig.6 Attribute inferring possibility for users with different social degrees in Google+

图 6 Google+数据集中具有不同社交度数 K 的用户属性可推测程度

4 结束语

本文研究了属性-社交网络中用户隐私属性匿名的问题,并提出了一种基于节点分割的匿名算法.该算法通过对具有敏感属性取值的用户节点进行分割,既实现了用户身份的隐藏,也实现了隐私属性的匿名.同时,该算法能够保持属性间的相关性,减少不必要的信息损失.实验结果表明,该算法能够有效降低用户属性被推测概率,并保持较高的数据可用性.但是,该算法的匿名有效性依赖于局部区域内属性的相关性,因此存在一定的偏差.在下一步工作中,可以对此进行改进.

References:

[1] Anagnostopoulos A, Kumar R, Mahdian M. Influence and correlation in social networks. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM Press, 2008. 7–15. [doi: 10.1145/1401890.1401897]

[2] Mislove A, Viswanath B, Gummadi KP, Druschel P. You are who you know: Inferring user profiles in online social networks. In: Proc. of the 3rd ACM Int'l Conf. on Web Search and Data Mining. ACM Press, 2010. 251–260. [doi: 10.1145/1718487.1718519]

[3] Zheleva E, Getoor L. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In: Proc. of the 18th Int'l Conf. on World Wide Web. ACM Press, 2009. 531−540. [doi: 10.1145/1526709.1526781]

[4] Narayanan A, Shmatikov V. De-Anonymizing social networks. In: Proc. of the 2009 30th IEEE Symp. on Security and Privacy. IEEE, 2009. 173−187. [doi: 10.1109/SP.2009.22]

[5] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Proc. of the 2008. IEEE Symp. on Security and Privacy. IEEE, 2008. 111−125. [doi: 10.1109/SP.2008.33]

[6] Sarwar B, Karypis G, Konstan J, Riedl J. Analysis of recommendation algorithms for E-commerce. In: Proc. of the 2nd ACM Conf. on Electronic Commerce. ACM Press, 2000. 158−167. [doi: 10.1145/352871.352887]

[7] Deng AL, Zhu YY, Shi BL. A collaborative filtering recommendation algorithm based on item rating prediction. Ruan Jian Xue Bao/Journal of Software, 2003,14(9):1621−1628 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/14/1621.htm

[8] Yin Z, Gupta M, Weninger T, Han J. Linkrec: A unified framework for link recommen-dation with user attributes and graph structure. In: Rappa M, ed. Proc. of the 19th Int'l Conf. on World Wide Web. New York: ACM Press, 2010. 1211−1212.

[9] 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. 2012. http://arxiv.org/pdf/1112.3265.pdf

[10] Yang SH, Long B, Smola A, Sadagopan N, Zheng Z, Zha H. Like like alike—Joint friendship and interest propagation in social networks. In: Sadagopan S, Ramamritham K, Kumar A, Ravindra MP, Bertino E, Kumar R, eds. Proc. of the 20th Int'l Conf. on World Wide Web. New York: ACM Press, 2011. 537−546.

[11] Yuan MX, Chen L, Yu PS, Yu T. Protecting sensitive labels in social network data anonymization. IEEE Trans. on Knowledge and Data Engineering, 2013,25(3):633−647. [doi: 10.1109/TKDE.2011.259]

[12] Sun X, Sun L, Wang H. Extended *k*-anonymity models against sensitive attribute disclosure. Computer Communications, 2011, 34(4):526−535. [doi: 10.1016/j.comcom.2010.03.020]

[13] Zhou B, Pei J. The *k*-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. Knowledge and Information Systems, 2011,28(1):47−77. [doi: 10.1007/s10115-010-0311-2]

[14] Ford R, Truta TM, Campan A. P-Sensitive *k*-anonymity for social networks. In: Stahlbock R, Crone SF, Lessmann S, eds. Proc. of the DMIN. Las Vegas: CSREA Press, 2009. 403−409.

[15] Wasserman S, Faust K. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press, 1994.

[16] Fard AM, Wang K, Yu PS. Limiting link disclosure in social network analysis through subgraph-wise perturbation. In: Proc. of the 15th Int'l Conf. on Extending Database Technology. ACM Press, 2012. 109−119. [doi: 10.1145/2247596.2247610]

[17] Freeman LC. Centrality in social networks: Conceptual clarification. Social Networks, 1979,1(3):215−239.

[18] Barrat A, Barthelemy M, Pastor-Satorras R, Vespignan A. The architecture of complex weighted networks. Proc. of the National Academy of Sciences of the United States of America, 2004,101(11):3747−3752. [doi: 10.1073/pnas.0400087101]

附中文参考文献:

[7] 邓爱林,朱扬勇,施伯乐.基于项目评分预测的协同过滤推荐算法.软件学报,2003,14(9):1621−1628. http://www.jos.org.cn/1000-9825/14/1621.htm

付艳艳(1986－),女,山东烟台人,博士生,CCF 学生会员,主要研究领域为数据安全,隐私保护.
E-mail: fuyy@tca.iscas.ac.cn

冯登国(1965－),男,博士,研究员,博士生导师,CCF 高级会员,主要研究领域为信息安全和密码学,可信计算与信息保障.
E-mail: feng@tca.iscas.ac.cn

张敏(1975－),女,博士,副研究员,CCF 高级会员,主要研究领域为数据库安全理论与技术,可信计算与信息保障关键技术.
E-mail: mzhang@tca.iscas.ac.cn

陈开渠(1976－),男,高级工程师,主要研究领域为网络安全,云安全.
E-mail: chenkq@nsccsz.gov.cn