软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2019,30(4):1106−1120 [doi: 10.13328/j.cnki.jos.005368]

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

社交网络下的不确定图隐私保护算法*

吴振强[1,2], 胡 静[2], 田堉攀[2], 史武超[2], 颜 军[2]

[1](现代教学技术教育部重点实验室(陕西师范大学),陕西 西安 710062)

[2](陕西师范大学 计算机科学学院,陕西 西安 710119)

通讯作者: 吴振强, E-mail: zqiangwu@snnu.edu.cn

摘 要: 社交网络平台的快速普及使得社交网络中的个人隐私泄露问题愈发受到用户的关心,传统的数据隐私保护方法无法满足用户数量巨大、关系复杂的社交网络隐私保护需求.图修改技术是针对社交网络数据的隐私保护所提出的一系列隐私保护措施,其中不确定图是将确定图转化为概率图的一种隐私保护方法.主要研究了不确定图中边概率赋值算法,提出了基于差分隐私的不确定图边概率赋值算法,该算法具有双重隐私保障,适合社交网络隐私保护要求高的场景.同时提出了基于三元闭包的不确定图边概率分配算法,该算法在实现隐私保护的同时保持了较高的数据效用,适合简单的社交网络隐私保护场景.分析与比较表明:与(k,ε)-混淆算法相比,基于差分隐私的不确定图边概率赋值算法可以实现较高的隐私保护效果,基于三元闭包的不确定图边概率分配算法具有较高的数据效用性.最后,为了衡量网络结构的失真程度,提出了基于网络结构熵的数据效用性度量算法,该算法能够度量不确定图与原始图结构的相似程度.

关键词: 社交网络;不确定图;差分隐私;三元闭包;网络结构熵

中图法分类号: TP309

中文引用格式: 吴振强,胡静,田堉攀,史武超,颜军.社交网络下的不确定图隐私保护算法.软件学报,2019,30(4):1106−1120. http://www.jos.org.cn/1000-9825/5368.htm

英文引用格式: Wu ZQ, Hu J, Tain YP, Shi WC, Yan J. Privacy preserving algorithms of uncertain graphs in social networks. Ruan Jian Xue Bao/Journal of Software, 2019,30(4):1106−1120 (in Chinese). http://www.jos.org.cn/1000-9825/5368.htm

Privacy Preserving Algorithms of Uncertain Graphs in Social Networks

WU Zhen-Qiang[1,2], HU Jing[2], TIAN Yu-Pan[2], SHI Wu-Chao[2], YAN Jun[2]

[1](Key Laboratory of Modern Teching Technology of Ministry of Education (Shaanxi Normal University), Xi'an 710062, China)

[2](School of Computer Science, Shaanxi Normal University, Xi'an 710119, China)

Abstract: The rapid popularization of the social network platform is causing growing concern among users of personal privacy disclosure in social networks, and due to the characters of social network which have the large number of users and with complicated relationships, the traditional privacy preserving method cannot be applied to the social network privacy protection which have a number of users and complicated. Graph modification technique is a series of privacy preserving methods proposed for the privacy preserving of social network data. Uncertain graph is a privacy preserving method, which converting a deterministic graph into a probability graph. In this study, the edge probability assignment algorithm is mainly focused on in the uncertain graph, and an algorithm for assigning the edge probability assignment is proposed based on differential privacy. The algorithm has a double privacy protection, which is suitable for social networks with high privacy requirements. Meanwhile, a different algorithm of uncertain graphs' edge probability assignment is

* 基金项目: 国家自然科学基金(61602290, 61173190); 中央高校基本科研业务费专项资金(GK201704017, GK201501008)

Foundation item: National Natural Science Foundation of China (61602290, 61173190); Fundamental Research Funds for the Central Universities (GK201704017, GK201501008)

本文由"面向隐私保护的新型技术与密码算法专题"特约编辑禹勇教授推荐.

收稿时间: 2017-06-01; 修改时间: 2017-07-13; 采用时间: 2017-08-22

presented based on the triadic closure, which achieves privacy preserve while maintains high data utility and suitable for simple social networks. The analysis and comparison show that the algorithm for assigning the edge probabilities of uncertain graph based on differential privacy can achieve a higher privacy preserving which was compared with obfuscation algorithm, and the algorithm of uncertain graphs' edge probability assignment based on triadic closure has higher data utility. Finally, in order to measure the distortion of the network structure, a data utility measure is proposed based on network structure entropy. The algorithm can measure the similarity between the uncertain graph and the original structure.

Key words: social network; uncertain graph; differential privacy; triadic closure; network structure entropy

1 引 言

随着互联网技术的日益成熟和快速普及,越来越多的用户通过社交网络平台进行沟通并共享信息,导致社交网络积累了大量的用户行为数据.这些数据中包含了大量的个人敏感信息,如身份信息、社会关系信息等,这些敏感信息容易引起不法分子的关注并成为安全攻击的目标.例如 2016 年 5 月 19 日,美国社交网站 LinkedIn 宣布,有一个叫“peace”的黑客组织在黑市上以 5 比特币的售价公开销售 1.67 亿个领英用户登录信息,其中有 1.17 亿包括电子邮件和密码;同年 6 月初,同样是代号为“peace”的黑客称已经拿到了全球第二大的社交网站 MySpace 的 3.6 亿用户账号以及 4.27 亿密码,并且在暗网上以 6 个比特币的价格公开出售[1],使得用户的身心和财产安全遭到严重侵犯,除此之外,一些不法分子甚至利用这些账号信息在用户的亲友之间行骗.因此,如何在方便用户使用社交网络的同时,保护好用户的个人敏感信息不被泄露成为了急需解决的社会问题.

社交网络由多种角色构成,它们之间的关系错综复杂且相互影响.传统数据的隐私保护方法已经不适用于社交网络中的隐私信息保护,对社交网络的研究不仅要关注每个社会角色的属性,同时也要关注这些角色之间的关系.为了准确描述社交网络中实体之间的关系,可以将社交网络抽象成图(graph)结构,每个社会角色用节点表示,角色之间的关系通常用边来表示,角色之间关联程度可以通过对边进行加权重来表示[2].因此,对社交网络的研究可以转化为对图结构的研究.由于社交网络中包含许多个人的敏感信息,将其抽象为图结构后,图中的节点和边也会涉及到个人隐私信息,因此,通过研究图结构的隐私保护来实现社交网络隐私保护的目的.目前,社交网络隐私保护方法可分为 5 类.(1) 标识符替换法.标识符替换是采用伪名的思想,用合成的标识符来替换掉身份证号或名字等唯一标识属性,这种方法实现简单但容易遭受背景信息攻击.Backstrom[3]等人提出在发布真实图数据之前用合成标识符替换可识别属性的隐私保护技术,但是这种方法容易受到背景知识的攻击,攻击者可以根据顶点的结构特征推断出顶点的身份信息.(2) 差分隐私法.差分隐私[4]是一种具有严格可证明的数学定义和可量化评估的隐私保护模型,将差分隐私应用到社交网络图结构中,可以提供一种严格、有效的量化评估方法.Hay 等人[5]基于差分隐私技术提出图结构中节点和边差分隐私,其中边的差分隐私相对较为简单,因此应用更为广泛.Day 等人[6]基于聚类和累积直方图提出了两种在节点差分隐私下发布度分布的方法.Karwa 等人[7]提出了一种严格、高效的边差分隐私保护方法来发布实用性高的网络数据统计信息.Task 等人[8]提出了图结构中出度差分隐私,实现简单,但相较于节点保护能力较弱.(3) 图聚类方法.图的聚类[9]是将多个节点或边聚合成一个超级节点集合,对外只公布超级节点的统计信息,从而保护了节点内用户的隐私.Hay 等人[10]针对背景信息攻击提出一种基于节点聚类实现网络数据匿名化的隐私保护方法.考虑到社交网络中的关系对应图结构中的边,这些连接关系也会泄露个人隐私信息.Zheleva 和 Getoor[11]针对关系链识别攻击提出了一种基于边聚类的隐私保护方法.但聚类在一定程度上会降低原社交网络图数据的实用性,因此在满足隐私保护的前提下,为了尽可能地保持原图数据的实用性,引入边/顶点修改技术.(4) 边/顶点修改技术是对图的局部进行修改来实现隐私保护.针对社交网络中角色关系复杂多样的特点,可以通过随机修改图中部分边来实现社交网络图的匿名化,进而实现隐私保护的目的.随机边修改技术主要包括随机增添和删除边、随机旋转边和随机交换边 3 种.Ying 等人[12]研究了不同的随机方法对于顶点之间关系的影响,提出了保护原始图特征谱的算法.攻击者依据原图中节点的度信息,仍可以重新定义原图数据信息.针对链接攻击,Liu 和 Terzi[13]提出了 k-度匿名概念.k-度匿名模型是 k-匿名[14]在图上的应用与拓展.设图 $G=(V,E)$中的任意一个节点度数至少与 $k-1$ 个节点的度数相同,则该图为 k-度匿

名图.(5) 不确定图方法.不确定图方法是向原社交网络图中的部分节点之间随机添加或删除小概率边来注入不确定性.通过对数据更小的扰动变化既可以实现隐私保护,又可以保持原数据效用性.2012 年,Boldi 等人[15]首次提出了(k,ε)-混淆算法,该方法在抗顶点身份攻击的同时保证了图结构数据的最小化失真;2013 年,Mittal 等人[16]提出了基于随机游走的不确定图数据发布方法,该方法可防止链接攻击;2014 年,Nguyen 等人[17]提出了方差最大化的方法来衡量隐私与效用之间的关系;2015 年,Nguyen 等人[18]在前期工作的基础上又提出了基于不确定邻接矩阵 UAM 的通用匿名模型,并将 UAM 引入到方差最大化算法中.为了提高不确定图方法的隐私保护效果,胡静[19]利用边差分隐私实现了对原图基于任何背景知识攻击的隐私保护,并且经过差分隐私的后置处理方法,提高了隐私保护的数据效用.

目前,不确定图已成为一种新的隐私保护方法,它的主要原理是将不确定性注入到社交网络图的边中,发布经混淆后的不确定图来达到隐私保护的目的.该方法通过为图中的边分配概率值来实现隐私保护,同时对原图数据改变较小,一定程度上保持了较高的原数据效用,相较于完全去除或添加边,保护效果更好.

本文针对社交网络图的隐私保护,提出了两种不确定图隐私保护算法:基于差分隐私的不确定图边概率赋值算法和基于三元闭包的不确定图边概率分配算法.为了对现有算法和本文提出的算法进行统一的隐私分析,我们采用边熵来衡量算法的隐私保护程度.同时,针对社交网络的度量问题,提出了一种基于网络结构熵的图结构数据效用性的度量算法,该算法与 k 度匿名模型相比,能够更加精确地反映出网络结构隐私性的变化过程.

本文第 1 节介绍背景和相关工作.第 2 节介绍本文算法的相关基础知识.第 3 节基于不确定图中边概率的赋值提出两种社交网络图隐私保护算法:基于差分隐私的不确定图边概率赋值算法和三元闭包的不确定图边概率分配算法.第 4 节对本文提出的两种算法分别从数据的隐私性和效用性方面进行实验验证分析.第 5 节提出一种基于网络结构熵的图结构数据效用性度量算法.第 6 节总结全文并展望下一步工作.

2 相关基础知识

2.1 差分隐私的基本定义和相关概念

定义 1(差分隐私). 存在两个相邻数据集 D,D'和算法 K,$K(D)$表示算法 K 在数据集 D 上的输出集合,O 是算法 K 所有输出值的集合,若算法 K 在数据集 D 和 D'上任意输出结果为满足下面不等式(1):

$$\Pr[K(D)\in O]\leqslant e^{\varepsilon}\times\Pr[K(D')\in O] \tag{1}$$

则算法 K 满足ε-差分隐私.ε称为隐私保护预算,ε的取值决定了保护效果,ε取值的大小与保护效果成正比,与数据失真程度成反比.差分隐私以其严格的数学定义为隐私的评价提供了理论依据.

差分隐私实现机制包括:指数机制、拉普拉斯机制和高斯机制.其中,指数机制一般应用于非数值类数据,拉普拉斯机制和高斯机制适用于数值型数据的隐私保护.

定义 2(邻近图). 给定图 $G_1=(V_1,E_1)$,$G_2=(V_2,E_2)$,若在 G_1,G_2 中有 $|V_1\oplus V_2|+|E_1\oplus E_2|=1$,则称 G_1、G_2 为邻近图.

本文中,由于 $V_1=V_2$,只要 $|E_1\oplus E_2|=1$,即 E_1 与 E_2 的汉明距离为 1,我们就称 G_1、G_2 为邻近图,如图 1 所示,图 1(a)和图 1(b)所示为邻近图.

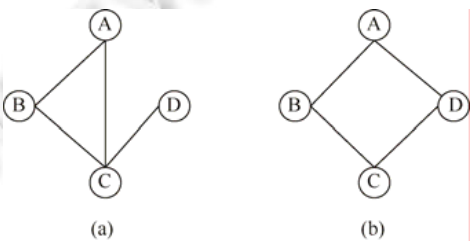

Fig.1 The example of neighboring graphs

图 1 邻近图示例

定义 3(敏感度). 给定一个函数 $f:G\to G''$,其中,G、G'具有相同的顶点集合,函数 f 的全局敏感度为

$$\Delta f = \max_{G_1,G_2} \| f(G_1) - f(G_2) \|_1 \tag{2}$$

其中,G_1、G_2是邻近图,G''为经过随机算法后的输出图,f 是查询函数,表示对于 G_1、G_2 中的边 e_i,查询边 e_i 是否存在于 G_1 和 G_2 中,图 1(a)和图 1(b)中的Δf=2.

定义 4(边差分隐私). 给定一种随机算法 M,$Range(M)$为算法 M 的取值范围,若算法 M 在邻近图 G_1、G_2上的输出结果 $S(S\in Range(M))$满足:

$$\Pr[M(G_1)\in S] \leqslant e^{\varepsilon} \times \Pr[M(G_2)\in S] \tag{3}$$

则称算法 M 满足ε-差分隐私,其中,ε表示隐私保护程度,Pr[·]表示算法 M 的随机性.

为了实现边差分隐私,通过拉普拉斯分布产生的噪声值加入到真实输出值实现边数据的扰动,从而实现了差分隐私保护.

定义 5(Laplace 机制). 对于函数 $f:G\to G'$,若算法 M 的输出结果满足下列等式,则称算法 M 是满足ε-差分隐私的.

$$M(G) = f(G) + Lap(\Delta f/\varepsilon) \tag{4}$$

其中,$Lap(\Delta f/\varepsilon)$是添加到查询结果中的期望值为 0,位置参数是$\Delta f/\varepsilon$的拉普拉斯分布的噪声.同时,噪声值的大小与敏感度$\Delta f$ 和隐私预算ε有关.这种通过添加服从拉普拉斯的噪声值实现差分隐私的机制称为 Laplace 机制.

2.2 不确定图

定义 6(不确定图). 给定一个图 $G=(V,E)$,如果映射 $p:V_p\to[0,1]$是边集中每条边存在的概率函数,那么图 $G'=(V,p)$是关于图 G 的不确定图,其中,V_p 表示集合 V 中所有可能的顶点对,即 $V_p=\{(v_i,v_j)|1\leqslant i<j\leqslant n\}$,相应地,$V_p=n(n-1)/2$.

2.3 三元闭包

定义 7(邻近潜在边). Leskovec 等人[20]指出,真实图中节点之间的链接概率随着它们层级之间相对距离的增加而减小,这些边的存在可以减少基于最短路径的统计.Vázquez[21]提出最近邻居可以解释邻居间的聚集系数、平均度和度分布的能量规律,这些属性与对社交网络图的观察结果完全一致.

三元闭包的基本原则:如果两个人有共同的朋友,这两个人将来成为朋友的可能性就会增加.例如,节点 B 和节点 C 具有共同的朋友 A,则 B 和 C 成为朋友的概率会增加.类似地,节点 A 和节点 E 之间也会产生关联边,如图 2 所示.将三元闭包理论引入到社交网络研究中,可以形成三角形网络结构,利用三角形结构特性将原图转变为不确定图.

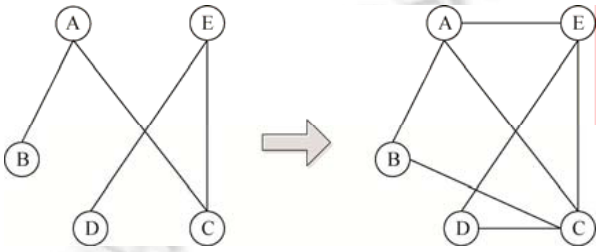

Fig.2 The theory of triadic closure

图 2 三元闭包理论模型

2.4 网络结构熵

熵表示系统的混乱程度.网络结构熵是对整个网络结构是否有序的度量.网络结构熵 *Entropy* 定义如下:

$$Entropy = -\sum_{i=1}^{n} I_i \cdot \ln I_i \tag{5}$$

$$I_i = \frac{d_i}{\sum_{j=1}^{n} d_j} \tag{6}$$

其中,d_i 为节点 v_i 的度.当网络结构为全连接图时,其值最大,反之,无边相连的孤立节点图,其值最小.

2.5 *k*-度匿名

k-度匿名模型是 *k*-匿名在图上的应用与拓展,由 Liu 等人[13]首次提出,相关概念定义如下.

(1) *k*-匿名向量.如果一个向量中的每个值在该向量中出现的次数至少是 *k* 次,则这个向量被称作 *k*-匿名向量.例如向量 v=[4,4,3,3,3],则该向量为 2-匿名向量.

(2) *k*-度匿名图.G=(V,E)表示一个图,如果这个图的度序列所组成的向量是一个 *k*-匿名向量,则该图为 *k*-度匿名图.

3 基于不确定图的隐私保护算法

本节主要介绍两种基于不确定图边概率赋值的隐私保护算法,并分别对其给出详细介绍.基于差分隐私的不确定图边概率赋值算法(简称差分隐私法)提供了严格可证明的隐私保护,并在一定程度上保护了数据效用.基于三元闭包的不确定图边概率分配算法(简称三元闭包法)不仅实现了隐私保护的目的,而且有约束的分配边概率可以保持较高的数据效用.

3.1 差分隐私法

下面首先将不确定图的构造过程抽象为一个模型,然后在该模型下提出一种基于差分隐私技术实现不确定图边概率赋值的隐私保护(uncertain graph based on differential privacy,简称 UGDP)模型,并对该模型进行说明.然后给出 UGDP 算法的伪代码并进行分析.最后,证明了 UGDP 算法满足差分隐私.

1) 不确定图模型

Fig.3 Uncertain graph construction model of UGDP algrothm

图 3 UGDP 算法的不确定图构造模型

该模型(如图 3 所示)包括 3 个部分,第 1 部分和第 3 部分为算法的输入和输出,其中,算法的输入为原始数据图,输出为不确定图,中间部分是该模型的主要环节,研究者可以根据不同的需求提出不同的隐私保护算法来实现不确定图的隐私保护.在该模型下,我们提出了基于差分隐私的不确定图边概率赋值算法 UGDP.

2) UGDP 算法

UGDP 算法是利用差分隐私技术来实现不确定图边概率赋值的隐私保护算法,算法的执行框如图 4 所示,主要由以下 4 步组成.

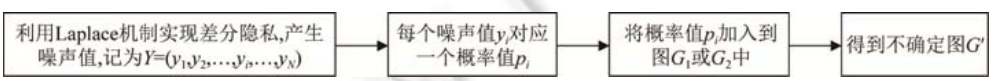

Fig.4 The UGDP algorithm

图 4 UGDP 算法

(1) 利用 Laplace 机制进行加噪,产生的噪声表示为 Y=(y_1,y_2,…,y_i,…,y_N),将产生的噪声加入图 G_1 或 G_2 中,得到噪声图.

(2) 为了构造不确定图,每个噪声值 y_i 对应一个概率值 p_i,概率值 p_i=Pr[y_i]=F(y_i).

(3) 将概率值 p_i 加入到图 G_1 或 G_2 中构成不确定图,将概率值 p_i 作为顶点和顶点之间存在边的概率.

$$F(y_i) = \int_{-\infty}^{y_i} g(x)\mathrm{d}x \tag{7}$$

其中,g(x)是服从期望值μ、位置参数 b 的拉普拉斯分布.

$$g(x)=\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \tag{8}$$

(4) 在进行数据发布时,为了更好地保护图中的隐私信息,将邻近不确定 G'作为发布图.

算法详细描述如算法 1 所示.

算法 1. UGDP 算法.

输入:原图 $G=(V,E)$,敏感度Δf;

输出:$G'=(V,p)$.

1: $b\leftarrow(\Delta f/\varepsilon)$

2: $y_i\leftarrow Lap(b),y_i\in Y$

3: $\Pr[y_i]\leftarrow F(y_i)$

4: $p_i\leftarrow\Pr[y_i],p_i\in p$

5: 添加概率 p_i 到边 e_i 上,且 $p_i\in p,e_i\in E$

6: Return $G'=(V,p)$

3) 隐私分析

定理. UGDP 算法满足差分隐私.

证明:令 $f(\cdot)$为函数 $f:G\to G'',G''$是算法输出的不确定图,G_1、G_2 为邻近图即图中边的汉明距离为 1. P_{G_1} 表示 $UGDP(G_1,f,\varepsilon)$的概率密度函数, P_{G_2} 表示 $UGDP(G_2,f,\varepsilon)$的概率密度函数.概率和噪声的关系为 $y_i\sim p_i$.G_3 是算法过程中得到的噪声图.因为噪声图到不确定图的过程符合差分隐私的后置处理技术,因此为了证明 UGDP 算法满足差分隐私,只需要证明原图到噪声图的过程满足差分隐私即可.具体证明过程如下.

$$\begin{aligned}\frac{P_{G_1}[G_3]}{P_{G_2}[G_3]}&=\frac{P_{G_1}[UGDP(G_1,f,\varepsilon)-f(G_1)]}{P_{G_2}[UGDP(G_2,f,\varepsilon)-f(G_2)]}\\&=\prod_{i=1}^{j}\left(\frac{\exp\left(-\frac{|f(G_1)_i-G_{3_i}|}{\Delta f/\varepsilon}\right)}{\exp\left(-\frac{|f(G_2)_i-G_{3_i}|}{\Delta f/\varepsilon}\right)}\right)\\&=\prod_{i=1}^{j}\exp\left(\frac{|f(G_2)_i-G_{3_i}|-|f(G_1)_i-G_{3_i}|}{\Delta f/\varepsilon}\right)\\&\leqslant\prod_{i=1}^{j}\exp\left(\varepsilon\cdot\frac{|f(G_2)_i-f(G_1)_i|}{\Delta f}\right)\\&=\exp\left(\varepsilon\cdot\frac{\|f(G_2)-f(G_1)\|_1}{\Delta f}\right)\leqslant\exp(\varepsilon).\end{aligned}$$ □

由于 UGDP 算法是在符合不确定图隐私的同时满足差分隐私,具有双重的隐私保护效果,因此,UGDP 算法具有较强的隐私保护的特征.在隐私性度量上我们采取差分隐私的度量方式,即用隐私预算ε来度量隐私保护程度.隐私预算ε越小,噪声的取值范围越广,噪声值对应的概率的取值也越广,在边混淆时达到的混淆程度越好,因此隐私保护效果越好.反之,隐私预算ε越大,噪声的取值范围越窄,噪声值对应的概率的取值也越窄,在边混淆时达到的混淆程度较差,因此隐私保护效果较差.

本文提出的 UGDP 算法与原有的不确定图算法相比,在满足不确定图的同时满足差分隐私的所有特征,尤其它是一种严格可证明的隐私保护算法,然而,该算法也存在某些局限性,如数据的低可用性问题.

3.2 三元闭包法

本节首先构建不确定图模型,并在该模型下提出一种基于三元闭包的不确定图边概率分配算法来实现图的匿名化,然后对该模型的 3 个主要流程详细描述并给出该算法的伪代码.该算法在实现隐私保护的同时保持

了原数据的效用性.

1) 基于三元闭包的不确定图算法模型

如图 5 所示,该模型包括 5 个部分,第 1 部分和第 5 部分为算法的输入和输出,其中,算法的输入为原始数据图,输出为不确定图,中间部分是该模型的主要环节,该环节分为 3 个步骤,具体内容详见后续的算法描述.

Fig.5 Uncertain graph based on triadic closure algorithm

图 5 基于三元闭包的不确定图构造模型

2) 算法描述

对于一个社交网络图 G,它的点集为 V,边集为 E.

(1) 加边

该过程(如图 6 所示)首先集中在收集图中所需的信息以获得节点集 V 和一个边集 E.我们随机选择点 $u,v\in V$ 且边 $(u,v)\notin E$,若两点之间的距离 $dis(u,v)$ 等于 2,则将边 (u,v) 加到图 G 中.如图 6 所示,最多可以增加 3 条边.

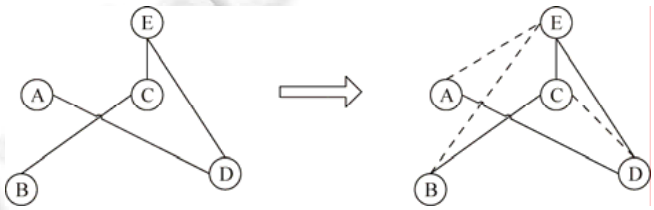

Fig.6 Add-edges

图 6 加边

(2) 确定三角形数量

当添加边 (u,v) 时,这个边与其附近的两个邻近边可以形成一个三角形,继续执行上一步,可以得到多个三角形.为了得到需要的三角形,在选择三角形时必须附加一些约束条件.如果两个三角形具有公共边,必须选择一个并放弃一个.最终得到一个三角形集合,该集合中任意两个三角形没有公共边.例如,在图 7 中,当添加边 AE 和边 BE 时,得到ΔAED 和ΔBEC.然后判断两个三角形是否具有公共边.如图 7 所示,这两个三角形没有公共边,则可以将边 AE 和边 BE 添加到图中.与边 AE 和边 BE 相反,边 CD 被放弃,因为包括边 AE 的ΔAED 与ΔDEC 具有相同边 ED.

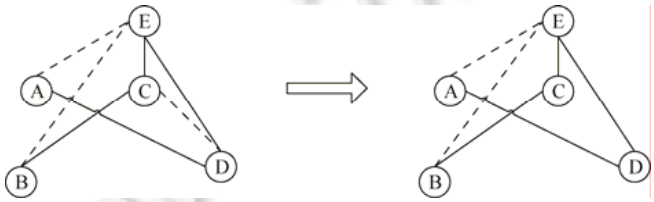

Fig.7 Determine the triangle

图 7 确定三角形数量

详细的描述如算法 2 和算法 3 所示.

算法 2. 加边和选取三角形.

输入:原图 $G=(V,E_G)$,节点数 V,边数 E_G,添加的边数 e_m;

输出:所加边的集合 S 和三角形集合 T,满足要求添加的总边数 m.

1: $S\leftarrow\varnothing,T\leftarrow\varnothing,i=0$

2: while $i<e_m$ do

3: 随机选点 u,v 且$(u,v)\notin E_G$,$dis(u,v)=2$

4: if $T(u,v)\cap T=\varnothing$

5: $E_G\leftarrow E_G\cup(u,v)$

6: $S\leftarrow S\cup(u,v)$,$T\leftarrow T\cup(u,v)$

7: $i=i+1$

8: Return $S,T,m=|T|$

算法 3. 注入不确定.

输入:原图 $G=(V,E_G)$,节点数 V,边数 E_G,预期添加的边数 e_p,满足要求添加的总边数 m,$m\geqslant e_p$,筛选得到的边集合 S,三角形集合 T;

输出:不确定图 $G'(V,p)$.

1: $m=|T|$,$i=0$

2: while $i<e_p$ do

3: 随机选取$\Delta ABC\in T$

4: $p(A,B)=random(0.5,1)$

5: $p(A,C)=(2-p(A,B))/2$

6: $p(B,C)=2-(p(A,B)+p(A,C))$

7: $i=i+1$

8: Return $G'=(V,p)$

(3) 注入不确定性

对三角形集合中每个三角形的3条边随机分配概率,为了使原图的边概率保持不变,本文规定每个三角形3条边的概率和 $\sum_{i=1}^{3}p_i=2$, 且原图中不属于三角形的边的概率值为 1.通过向所有边注入概率值,可以将原图转换为不确定图.当所有的边完成概率赋值后,可以得到 $\sum_{i=1}^{|E|}p_i=|E_G|$, E_G 等于原图中边的个数.通过对边注入不确定性生成的不确定图如图 8 所示.

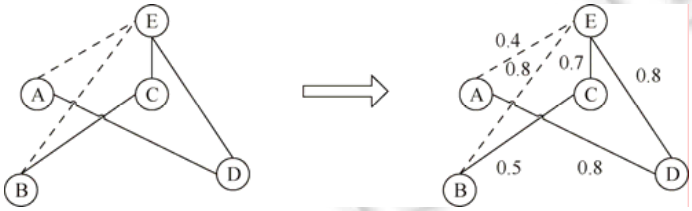

Fig.8 Injecting probability

图 8 注入不确定性

通过上述流程,我们在增加约束的条件下对原图 G 注入不确定性,得到不确定图 G'.

3) 算法分析

为了实现较高的隐私保护效果,将原图转换为不确定图,则攻击者按组合方法只能以比较低的概率从不确定图中恢复出原图,因此本算法能够实现对原图的概率性保护.在基于三元闭包的不确定图边概率分配算法中,首先基于三元闭包原理对原图加边,实现对原图的修改.然后,选择加边形成的三角形,对其三边分配一定的概率值得到一个不确定图.若具有概率值的三角形越多,则能够得到原图的概率越低,因此该算法能够对原图进行隐私保护.在转换过程中,对于三角形的概率分配进行限制,使得三边的概率之和等于 2,目的是确保加边前后的边概率之和保持不变.在这种约束条件下,得到的不确定图的边数概率之和等于原图的边数概率之和.同时,这种约束条件可以提高不确定图的数据效用.因此,基于三元闭包的不确定图算法既能实现隐私保护,又保持了较

高的数据效用性.

提出的基于三元闭包的不确定图边概率分配算法,通过加边构建三角形集合并对选择后的三角形集合中所有的边注入不确定性,将原社交网络图转化为不确定图,进而实现对社交网络的隐私保护.该算法在对社交网络隐私保护的同时保持了原网络数据效用,与其他已有算法相比,在处理简单的社交网络数据时,该算法运行效率更高.由于该算法是对网络图的局部进行扰动,所以对于更复杂的网络图还需要进一步加以探索.

4 分析与比较

为了对不同隐私保护方法的隐私效果进行评价,下面利用不确定图的边概率信息,引入边熵来衡量所发布不确定图的隐私保护程度,并以此为评价依据衡量变换前后隐私程度的变化情况.同时,为了验证算法对数据的效用性影响程度,又定义了一些图统计指标来说明本文算法的数据效用性.然后,利用 Python 语言及第三方功能包 NetworkX 对本文算法进行程序编程和仿真实验.实验数据分为两部分,一部分为真实数据集,一部分为合成数据集.真实数据集来自于 Karate 和 Dolphin 俱乐部,合成数据集分别为 200 和 500 个节点,连接概率为 0.2 的随机网络图.为了减少随机性,实验进行了 10 次模拟取平均值.

4.1 隐私性度量

信息熵是信息论中用于度量信息量的一个概念,一个系统越是有序,信息熵就越低;反之,一个系统越是混乱,信息熵就越高,在不确定图隐私度量中,由于不确定图的边具有很强的随机性,我们引入边熵 Ent_e 来衡量隐私保护效果.根据边的不确定程度,使用边熵来衡量不确定图的不确定性,即不确定图对原始图的隐私保护程度,不确定图的边熵定义如以下公式所示.

$$I_{e_i} = -p(e_i) \times \log_2 p(e_i) \tag{9}$$

其中,$e_i \in G'$,$p(e_i)$是该条边存在的概率.

$$Ent_e = \sum_{e \in G'} I_{e_i} \tag{10}$$

其中,Ent_e 为不确定图的边熵,Ent_e 值越大,表示不确定图的不确定程度越大,对应隐私保护方法的隐私保护效果越好.

4.2 数据效用性度量

根据文献[15,18]中提出的相关指标来度量算法的数据效用性,其中,NE 表示图中边的个数,AD 表示图中节点的平均度,DV 表示图中节点的度方差.

在图数据中,我们用 $d_1,d_2,\ldots,d_n$ 来表示图中节点度的序列.由于不确定图中每条边是以概率的形式出现的,因此我们不能直接用确定图中节点的度来表示不确定图中节点的度.在不确定图中,节点度的序列 $d_1,d_2,\ldots,d_n$ 是一些随机变量.我们利用节点的期望度来表示不确定图中节点的度,也就是说,对于任意的节点 $v \in V$,节点 v 的期望度是与它相连的边的概率之和,见等式(11).

$$d_v = \sum p(i,j) \tag{11}$$

在公式(11)中我们规定 $i=v$ 或者 $j=v$ 且 $i \neq j$.

原始图中 NE、AD 的计算见公式(12)、公式(13),不确定图中 NE'、AD'的计算见公式(15)、公式(16),其中,DV 在原始图与不确定图中的计算方式一致,见公式(14).

$$NE = \frac{1}{2}\sum\nolimits_{v \in V} d_v \tag{12}$$

$$AD = \frac{1}{n}\sum\nolimits_{v \in V} d_v \tag{13}$$

$$DV = \frac{1}{n}\sum\nolimits_{v \in V} (d_v - AD)^2 \tag{14}$$

$$NE=\frac{1}{2}\sum_{v\in V}\sum_{u\in V\setminus v}p(u,v)=\sum_{e\in V_2}p(e) \tag{15}$$

$$AD=\frac{1}{n}\sum_{v\in V}\sum_{u\in V\setminus v}p(u,v)=\frac{2}{n}\sum_{e\in V_2}p(e) \tag{16}$$

4.3 隐私性分析

为了对本文所提出的方法和已有方法的隐私性进行统一度量,我们利用边熵的变化情况来说明隐私保护效果.表 1 为 UGDP 算法边熵的变化情况,表 2 为三元闭包法中边熵的变化情况,表 3 为(k,ε)-混淆算法中边熵的变化情况.

边熵是用来度量图的不确定性程度,边熵的值越大表明图的不确定程度越大.表 1 表示的是 UGDP 算法中边熵的变化情况,随着节点个数的增加,边熵也对应增大,表明整个图的不确定程度在加大,即隐私保护效果越好.由于 UGDP 算法是符合差分隐私的,根据差分隐私的特性可知,隐私预算ε越小,则隐私保护程度越高.但是,在利用UGDP算法将确定图转化为不确定图时,边概率的赋值服从拉普拉斯分布,具有一定的随机性,因此,边熵的变化也具有随机性,与差分隐私的隐私保护程度不一定完全相符,也就是说,隐私预算越小,边熵的值不一定越大.

Table 1 The changes of edge-entropy in UGDP algorithm

表 1 UGDP 算法中边熵的变化情况

隐私预算 / 节点个数	ε		
	ε=0.01	ε=0.1	ε=1
34(Karate)	28.59	27.78	27.41
61(Dolphin)	58.97	57.86	57.65
200	2 128.27	2 122.26	2 112.45
500	8 927.23	8 912.71	8 990.46

基于三元闭包的不确定图边概率分配算法在生成不确定图的过程中,对原图进行加边,然后对选择后的三角形的边赋予概率值.在表 2 中,随着边数的增加,得到的不确定图的边熵也在增加,表现为同步增长趋势.这种趋势反映了图隐私保护的效果越来越好,同时证明了三元闭包法能够实现隐私保护的目的 m 为不确定图的生成过程中总共添加的边数,c 为调节加边数的因子,实际添加的边数为 $m\cdot c$.

Table 2 The changes of edge-entropy in triadic closure algorithm

表 2 三元闭包法中边熵的变化情况

增加的边数 / 节点个数	m		
	c=0.2	c=0.5	c=0.1
34(Karate)	5.71	13.66	28.31
61(Dolphin)	6.46	16.83	33.49
200	40.31	102.50	204.41
500	106.94	266.94	535.68

(k,ε)-混淆算法将确定图转化为不确定图时可以保持较好的数据效用性[18],为了把该算法达到的隐私保护效果与本文提出的算法进行直观的对比,下面采用边熵来进行隐私性度量.表 3 为(k,ε)-混淆算法中边熵的变化情况,由于篇幅和不同参数组合具有的多种可能性,下面只列出(k,ε)-混淆算法的一种情形来说明隐私保护效果,同时隐私度量的主要参数即混淆级别 k 的取值分别为 10 和 20,其他参数分别为ε=0.1,c=1,q=0.01.

Table 3 The changes of edge-entropy in (k,ε)-obfuscation algorithm

表 3 (k,ε)-混淆算法中边熵的变化情况

混淆级别 / 节点个数	k	
	k=10	k=20
34(Karate)	9.92	8.46
61(Dolphin)	17.09	17.65
200	626.74	619.24
500	2 535.95	2 541.95

通过表 1 和表 3 的对比,如果利用边熵来度量算法的隐私性,则 UGDP 算法在隐私保护程度上优于(k,ε)-混淆算法,可以达到较好的隐私保护效果.通过表 2 和表 3 的对比,(k,ε)-混淆算法优于三元闭包方法,可以达到较好的隐私保护效果.

4.4 数据效用性分析

在隐私保护的同时,我们通过 *NE*、*AD*、*DV* 对数据效用性进行了测量,见表 4~表 7.表 4 和表 5 列出 UGDP 算法的数据效用性.在表 4 中,对原始数据图的数据效用及ε=0.1 时 UGDP 算法生成的不确定图的数据效用进行了对比,可以看出,UGDP 算法的数据效用性较低.同时,表 5 表示的是不同隐私预算ε下的数据效用性的对比,我们得出隐私预算ε越大,隐私保护程度越差,数据效用性相对较好的结论.

表 6 列出的是三元闭包法的数据效用性.在表 6 中,通过三元闭包法得到的不确定图与原图比较可以得出,在度量指标 *NE*、*AD* 中,数据效用性保持不变,这说明,通过该方法得到的不确定图具有较好的数据效用性.同时,*DV* 在网络中明显增加,这说明,对原图进行修改时节点的度发生了变化.

Table 4 Comparison of data utility between uncertain graphs generated by UGDP algorithm when ε=0.1 and original graph

表 4 原始图与ε=0.1 时 UGDP 生成的不确定图的数据效用性比较

度量指标 / 节点个数	原始图			不确定图中ε=0.1		
	NE	*AD*	*DV*	*NE*′	*AD*′	*DV*
34(Karate)	78	4.0	14.94	39.83	2.34	3.50
61(Dolphin)	159	5.0	8.61	78.42	2.53	2.80
200	5 901	59	43.59	2 967.04	29.67	26.59
500	24 844	99	73.26	12 400.86	49.60	38.33

Table 5 Comparison of data utility when ε=0.01 and ε=1 in uncertain graphs generated by UGDP algorithm

表 5 ε=0.01 与ε=1 时 UGDP 算法生成的不确定图的数据效用性比较

度量指标 / 节点个数	不确定图中ε=0.01			不确定图中ε=1		
	NE	*AD*	*DV*	*NE*′	*AD*′	*DV*
34(Karate)	39.32	2.31	6.13	36.81	2.17	5.08
61(Dolphin)	78.32	2.53	2.81	78.19	2.52	2.16
200	2 959.71	29.60	30.50	2 963.22	29.63	24.49
500	12 467.18	49.87	45.90	12 392.10	49.57	42.54

Table 6 Comparison of data utility between uncertain graphs generated by triadic closure algorithm and original graph

表 6 原始图与三元闭包法生成的不确定图的数据效用性比较

度量指标 / 节点个数	原始图			不确定图		
	NE	*AD*	*DV*	*NE*′	*AD*′	*DV*
34(Karate)	78	4.58	14.94	78	4.58	11.97
61(Dolphin)	159	5.12	8.61	159	5.12	4.81
200	5 901	59.01	43.59	5 901	59.01	128.39
500	24 844	99.38	73.26	24 844	99.38	356.12

表 7 列出了不同 *k* 值情况下(k,ε)-混淆算法的数据效用性.通过度量指标 *NE*、*AD*、*DV* 在原始图和不确定图中的对比,我们可以看出,该算法具有较好的数据效用性.

Table 7 Data utility comparison for different *k* values in (k,ε)-obfuscation algorithm

表 7 (k,ε)-混淆算法中不同 *k* 值的数据效用性对比

度量指标 / 节点个数	(k,ε)-混淆 k=10			(k,ε)-混淆 k=20		
	NE	*AD*	*DV*	*NE*′	*AD*′	*DV*
34(Karate)	70.20	4.13	10.47	71.61	4.21	12.39
61(Dolphin)	145.22	4.68	6.79	145.35	4.77	6.99
200	5 405.66	54.05	48.06	5 416.10	54.16	50.13
500	22 781.41	91.13	78.38	22 793.89	91.18	83.39

通过不同算法的数据效用性比较,我们可以得出以下结论:UGDP 算法的数据效用性较差,三元闭包法的数据效用性最好,(k,ε)-混淆算法的数据效用性介于这两种算法之间.

4.5 算法整体分析

数据的隐私性与效用性本身是一对矛盾体,如果要达到高的隐私保护程度,通常就需要牺牲数据的效用性.在第 4.3 节和第 4.4 节中通过与现有的(k,ε)-混淆算法进行对比可以看出,本文提出的 UGDP 算法具有较高的隐私保护性,但是这种高隐私保护程度是通过牺牲数据的效用性而得到的.同时,三元闭包方法具有高数据效用性,然而该方法的隐私保护性偏低.

不同的场景可能需要不同的需求,若需要达到高隐私保护程度或者高数据效用性,则可以根据不同的需求来选择不同的隐私保护算法.

5 基于网络结构熵的度量算法

本节首先介绍网络结构熵可作为隐私度量的评价依据来度量图的隐私性,其次使用网络结构熵来衡量不确定图算法对原始图的破坏程度.

5.1 网络结构熵的隐私性度量

相关匿名算法[22]提出对隐私进行量化可以检验隐私保护算法的优劣,因此对图结构的隐私进行度量具有重要的理论意义和应用价值.网络结构熵是对整个结构是否有序的度量,自然可以用于研究图结构隐私度量.

如图 9 所示,4 个节点的连通网络结构的节点间连接方式不同,它们的度序列向量分别是 $\vec{v}_a=[1,2,1,1]$, $\vec{v}_b=[1,3,2,2]$, $\vec{v}_c=[2,3,3,2]$, $\vec{v}_d=[3,3,3,3]$.根据 k 度匿名图模型,图 9(a)和图 9(b)都是 1 度匿名图,图 9(c)是 2 度匿名图,图 9(d)是 4 度匿名图.毫无疑问,图 9(d)达到最高匿名级别,图 9(c)达到中等匿名级别,图 9(a)和图 9(b)达到相同匿名级别但具有不同的结构.可以看出,从 1 度匿名到 2 度匿名存在一个逐渐变化的过程,但 k 度匿名图模型并不能准确地反映出这个变化过程.

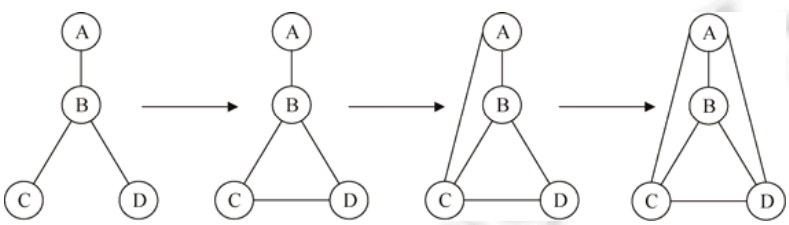

Fig.9 Four kinds of graphs with the same nodes but different linking relationships

图 9 4 种具有相同节点但不同链接关系的图结构

k-度匿名图模型的核心思想是度序列向量的匿名性.在一个 k 匿名向量中,在没有任何辅助信息的前提下,任何一个元素被识别出来的概率不超过 $1/k$.因此,许多图修改算法都通过加减边或者节点的方法来使图的度序列分布得更加均匀,而网络静态特征是用来描述网络特征的微观数量分布统计或宏观数量平均值统计.为了更加细致地描绘度序列的变化过程,本文采取网络静态特征之一的网络结构熵作为隐私度量指标来度量图的隐私性变化情况.

度序列分布反映了图的“形状”信息,而熵反映了图的“形状”是否有规律.图的形状越有规则,随机性就越小,因此熵就越小;若熵越大,图的度序列分布也越均匀.根据式(5)和式(6),在 k 邻近图中网络熵取最大值 $Entroy_{\max}=\log_2 n$,在星型图中其取得最小值 $Entropy_{\min}=\log_2 4(n-1)/2$,其中,$n$ 表示图中所有节点的个数.

在图 9(a)到图 9(d)的网络结构熵依次为 1.242、1.321、1.366 和 1.386.因此,隐私程度越高,网络结构熵越大.尽管图 9(a)和图 9(b)在 k-度匿名下的隐私程度相同,但网络结构熵仍然可以区分这两个不同结构

的隐私变化.

5.2 图结构数据效用性度量算法

网络结构熵的变化还可以用来衡量网络结构的变化,在衡量不确定图算法对原始图结构的破坏程度时,我们提出了基于网络结构熵的数据效用性度量算法.利用该算法可以判断原始图和不确定图结构熵的变化情况,从而可以衡量图结构的失真程度.不确定图的结构熵与原始图的结构熵越接近,说明不确定图对原始图的改变程度较小,较好地保持了原图的数据效用性.

结构熵是衡量网络结构的重要指标,可以精确、简洁地度量复杂网络的非同质性,当网络受损分裂成几个随机网或者多数节点在非连通的情况下,网络结构熵会变小.从表 8 中我们可以看出,原始图的结构熵与经过差分隐私法及三元闭包算法得到的不确定图的结构熵的差值不大,且不确定图中的网络结构熵相对于原始网络图来说较小,两者之间的误差值固定在 0.2 之内.表 8 只列出了节点个数小于等于 500 时网络结构熵的变化情况.随着节点个数的增加,两者网络结构熵的差值在逐渐减小.这说明,本文的算法在大数据场景下很好地保持了网络的结构特性.因此,与原始图的结构熵相比,4 种数据集在不同算法下得到的不确定图的结构熵与原始图的结构熵基本保持不变,从而说明我们提出的不确定图算法可以很好地保持原始图的结构特征.

详细描述如算法 4 所示.

算法 4. 图结构数据效用性度量算法.

输入:$G=(V,E)$或 $G'=(V,p)$;

输出:Entropy.

1: 如果输入的是确定图,则直接计算节点的度 d_v,转 4;否则,转 2;

2: $G'(V,p)\leftarrow G=(V,E)$;

3: 计算不确定图 $G'(V,p)$中节点 v 的度 d_v,其中,$d_v=\sum p(i,j)$;

4: 求整个网络的总度 d_{sum},$d_{sum}=\sum_{v\in V}d_v$;

5: 计算节点任意节点 $v\in V$ 的度占网络总度的概率 $p(v)=\dfrac{d_v}{d_{sum}}$;

6: 计算不确定图中网络结构熵 $Entropy=\sum -p(v)\times\log_2 p(v)$;

7: 返回 Entropy;

Table 8 The changes of network structure-entropy

表 8 网络结构熵的变化情况

比较对象 / 节点个数	原始图	不确定图			
		差分隐私法		三元闭包法	
		ε=0.1	ε=1	c=0.5	c=0.1
34(Karate)	4.70	4.58	4.60	4.73	4.71
61(Dolphin)	5.70	5.65	5.66	5.83	5.84
200	7.64	7.60	7.598	7.61	7.60
500	8.96	8.93	8.93	8.95	8.94

6 结 论

本文基于不确定图方法针对社交网络隐私保护提出两种不确定图算法:基于差分隐私的不确定图边概率赋值算法和基于三元闭包的不确定图边概率分配算法.差分隐私法将差分隐私与不确定图结合,不仅符合不确定图隐私保护方法,同时具有差分隐私的严格可证明特征,可以达到双重隐私保证,但其数据效用性有待提高.三元闭包法将三元闭包理论与不确定图结合实现了社交网络隐私保护,并在一定程度上保持了较高的数据效用,与(k,ε)-混淆算法相比,在处理简单的社交网络图数据时运行效率更高.因此,在下一步工作中,我们将探索三元闭包法在大规模社交网络图中的实证研究.同时,根据网络结构熵的特征,提出了一种基于网络结构熵的数据效用性度量算法,用来衡量对原始图结构的破坏程度,对于这方面内容的内在机理是未来重点探索的内容.

References:

[1] T1 life. Ten major data breaches in 2016: Social networking has become a harder-hit area (in Chinese). 2016. http://www.sohu.com/a/122021640_5266-42

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

[3] Backstrom L, Dwork C, Kleinberg J. Wherefore art thou r3579x: Anonymized social networks, hidden patterns, and structural steganography. In: Proc. of the 16th Int'l Conf. on World Wide Web. New York: ACM, 2007. 181–190. [doi: 10.1145/1242572.1242598]

[4] Dwork C. Differential privacy: A survey of results. In: Proc. of the Int'l Conf. on Theory and Applications of Models of Computation. Springer-Verlag, 2008. 1–19. [doi: 10.1007/978-3-540-79228-4_1]

[5] Hay M, Li C, Miklau G, *et al*. Accurate estimation of the degree distribution of private networks. In: Proc. of the IEEE Int'l Conf. on Data Mining. Washington: IEEE Computer Society, 2009. 169–178. [doi: 10.1109/ICDM.2009.11]

[6] Day WY, Li N, Lyu M. Publishing graph degree distribution with node differential privacy. In: Proc. of the Int'l Conf. on Management of Data. New York: ACM, 2016. 123–138. [doi: 10.1145/2882903.2926745]

[7] Karwa V, Raskhodnikova S, Smith A, *et al*. Private analysis of graph structure. ACM Trans. on Database Systems, 2014,39(3):1146–1157. [doi: 10.1145/2611523]

[8] Task C, Clifton C. A guide to differential privacy theory in social network analysis. In: Proc. of the IEEE Int'l Conf. on Advances in Social Networks Analysis and Mining. Washington: IEEE Computer Society, 2012. 411–417. [doi: 10.1109/ASONAM.2012.73]

[9] Stokes K, Torra V. On some clustering approaches for graphs. In: Proc. of the IEEE Int'l Conf. on Fuzzy Systems. IEEE, 2011. 409–415. [doi: 10.1109/FUZZY.2011.6007447]

[10] Hay M, Miklau G, Jensen D, *et al*. Resisting structural re-identification in anonymized social networks. VLDB Journal, 2010,19(6):797–823. [doi: 10.14778/1453856.1453873]

[11] Zheleva E, Getoor L. Preserving the privacy of sensitive relationships in graph data. In: Proc. of the ACM SIGKDD Int'l Conf. on Privacy, Security, and Trust in KDD. Berlin, Heidelberg: Springer-Verlag, 2007. 153–171. [doi: 10.1007/978-3-540-78478-4_9]

[12] Ying X, Wu X. Randomizing social networks: A spectrum preserving approach. In: Proc. of the SIAM Int'l Conf. on Data Mining. SIAM, 2008. 739–750. [doi: 10.1137/1.9781611972788.67]

[13] Liu K, Terzi E. Towards identity anonymization on graphs. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. New York: ACM, 2008. 93–106. [doi: 10.1145/1376616.1376629]

[14] Esmerdag E, Gursoy ME, Inan A, *et al*. Explode: An extensible platform for differentially private data analysis. In: Proc. of the 16th IEEE Int'l Conf. on Data Mining Workshops (ICDMW). IEEE, 2016. 1300–1303. [doi: 10.1109/ICDMW.2016.0189]

[15] Boldi P, Bonchi F, Gionis A, *et al*. Injecting uncertainty in graphs for identity obfuscation. Proc. of the VLDB Endowment, 2012, 5(11):1376–1387. [doi: 10.14778/2350229.2350254]

[16] Mittal P, Papamanthou C, Song D. Preserving link privacy in social network based systems. Computer Science, 2012.

[17] Nguyen HH, Imine A, Rusinowitch M. A maximum variance approach for graph anonymization. In: Proc. of the Int'l Symp. on Foundations and Practice of Security. Cham: Springer-Verlag, 2014,8930:49–64. [doi: 10.1007/978-3-319-17040-4]

[18] Nguyen HH, Imine A. Anonymizing social graphs via uncertainty semantics. In: Proc. of the 10th ACM Symp. on Information, Computer and Communications Security. New York: ACM, 2015. 495–506. [doi: 10.1145/2714576.2714584]

[19] Hu J. The research on preserving privacy methods in social networks based on uncertainty graph [MS. Thesis]. Xi'an: Shaanxi Normal University, 2018 (in Chinese with English abstract).

[20] Leskovec J, Kleinberg J, Faloutsos C. Graph evolution: Densification and shrinking diameters. ACM Trans. on Knowledge Discovery from Data, 2007,1(1):2. [doi: 10.1145/1217299.1217301]

[21] Vázquez A. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. Physical Review E Statistical Nonlinear & Soft Matter Physics, 2003,67(2):056104. [doi: 10.1103/PhysRevE.67.056104]

[22] Kelly DJ, Raines RA, Grimaila MR, *et al*. A survey of state-of-the-art in anonymity metrics. In: Proc. of the ACM Workshop on Network Data Anonymization. New York: ACM, 2008. 31–40. [doi: 10.1145/1456441.1456453]

附中文参考文献:

[1] T1 生活.2016 年十大数据泄露事件:社交网络成泄露重灾区.2016. http://www.sohu.com/a/122021640_526642

[2] 刘向宇,王斌,杨晓春.社会网络数据发布隐私保护技术综述.软件学报,2014,25(3):576–590. http://www.jos.org.cn/1000-9825/4511.htm [doi: 10.13328/j.cnki.jos.004511]

[19] 胡静.基于不确定图的社交网络隐私保护方法研究[硕士学位论文].西安:陕西师范大学,2018.

吴振强(1968－),男,陕西商洛人,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络科学,网络安全,隐私保护.

胡静(1993－),女,硕士,主要研究领域为隐私保护.

田堉攀(1994－),女,硕士,主要研究领域为隐私保护.

史武超(1991－),男,硕士,主要研究领域为隐私保护.

颜军(1974－),男,博士,讲师,主要研究领域为网络数据科学,隐私保护.