软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(3):576−590 [doi: 10.13328/j.cnki.jos.004511]

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

社会网络数据发布隐私保护技术综述*

刘向宇, 王 斌, 杨晓春

(东北大学 信息科学与工程学院,辽宁 沈阳 110819)

通讯作者: 王斌, E-mail: binwang@mail.neu.edu.cn, http://www.neu.edu.cn

摘 要: 对社会网络隐私保护的研究现状与进展进行了阐述.首先介绍了社会网络隐私保护问题的研究背景,进而从社会网络中的隐私、攻击者背景知识、社会网络数据隐私保护技术、数据可用性与实验测评等方面对当前研究工作进行了细致的分类归纳和分析,指出了当前社会网络隐私保护的不足以及不同隐私保护技术间的对比和优缺点,并对未来需要深入研究的方向进行了展望.对社会网络数据隐私保护研究的主流方法和前沿进展进行了概括、比较和分析.

关键词: 社会网络;隐私保护;数据发布

中图法分类号: TP301 **文献标识码**: A

中文引用格式: 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

英文引用格式: 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). http://www.jos.org.cn/1000-9825/4511.htm

Survey on Privacy Preserving Techniques for Publishing Social Network Data

LIU Xiang-Yu, WANG Bin, YANG Xiao-Chun

(School of Information Science and Engineering, Northeastern University, Shenyang 110819, China)

Corresponding author: WANG Bin, E-mail: binwang@mail.neu.edu.cn, http://www.neu.edu.cn

Abstract: This paper surveys the state of the art of privacy preserving for publishing social network data. First, the research background of privacy preserving for social network data is presented. Then, four important aspects of privacy preserving for social network data are summarized and analyzed in detail. The discussed topics includes privacy of social network, adversaries' background knowledge, privacy preserving technologies of social network, and data utility and experimental analysis. In addition, this paper points out the defects of privacy preserving in social networks and provides the comparisons of privacy preserving technologies. Finally, some potential future research directions are introduced. This paper aims to offer a deep insight into the mainstream methods and recent progress in this field, making detailed comparison and analysis.

Key words: social network; privacy preserving; data publishing

随着网络技术以及社交网站的迅速发展,例如 Facebook、MySpace、人人网等,通过社交网站进行交友、联系和互动的用户群体数量迅速增加.以 Facebook 为例,其用户总数在 2013 年 1 月突破 10 亿,约占世界人口的 14%.由于社会网络的繁荣和广泛应用,越来越多的研究学者和开发人员将其科学研究和应用开发的注意力集中到社会网络这种虚拟世界当中.社会网络分析已经成为社会学、地理学、经济学、信息学等诸多学科的研究热点.

* 基金项目: 国家重点基础研究发展计划(973)(2012CB316201); 国家自然科学基金(61173031, 61272178); 教育部基本科研业务费(N120504001, N110404015)

收稿时间: 2012-02-22; 定稿时间: 2013-04-19; jos 在线出版时间: 2013-11-28

CNKI 网络优先出版: 2013-11-28 14:39, http://www.cnki.net/kcms/detail/11.2560.TP.20131128.1439.001.html

基于社会网络数据进行数据挖掘和分析潜在模式比传统关系数据更加科学、效果更好,社会网络分析又称为链接挖掘(link mining)[1].通过对社会网络进行链接挖掘可以获得实体更丰富(如某实体在整个网络中的重要性)、更准确(如预测某实体所属类别)的信息.因此,亟待发布和共享更多的社会网络数据,为数据挖掘和模式分析提供更丰富的数据来源.然而,发布和共享社会网络数据会导致隐私泄露,并且社会网络中的隐私信息类型广泛,潜在隐私泄露方式更加多样化.例如:在电话网络中,Ada 和 Bob 之间频繁的电话和短信联系可能被视为敏感关系,因为他们不希望别人得知他们之间的亲密关系;在医疗网络中,某人与肺癌医生之间的联系可能被其视为敏感信息.大量研究工作为关系数据提供隐私保护,其中,文献[2,3]首先提出 *K*-匿名隐私保护模型,继而出现了一系列基于 *K*-匿名模型的关系数据隐私保护技术[4−14].但是,关系数据隐私保护技术[2−12]不能为社会网络数据提供隐私保护,这是因为关系数据隐私保护模型仅考虑攻击者将关系数据中每条记录的属性值作为背景知识进行隐私攻击,忽略了社会网络中结点之间的关系、社会网络图结构、结点在图中的结构和位置重要性等均可作为攻击者的背景知识进行隐私攻击.文献[13,15]基于真实数据,通过实验证明了社会网络面临很大的隐私攻击和泄露的威胁.可以看出,关系数据只是社会网络数据中结点之间相互独立时的特例,因此,关系数据隐私保护技术不能够满足社会网络数据的隐私保护要求,需要基于社会网络数据的特点研究相应的数据隐私保护技术.本文对近年来社会网络数据隐私保护研究工作[13−42]进行了归纳总结,指出了当前社会网络隐私保护的不足以及不同隐私保护技术间的对比和优缺点.

本文第 1 节介绍社会网络中涉及的隐私信息,基于当前研究工作,指出其社会网络隐私保护的不足.第 2 节对攻击者的背景知识进行总结归纳和分类,包括社会网络图结构信息、结点信息、边信息、预测模型等均可被攻击者作为背景知识进行隐私攻击.第 3 节分别从隐私保护方法、动态性、并行性等方面介绍当前社会网络隐私保护技术,并指出不同隐私保护技术的优缺点.第 4 节归纳常用的社会网络隐私保护技术的实验评测指标,其中包括结点数据可用性、边数据可用性、图结构及性质、图查询、执行效率等方面.第 5 节展望未来研究趋势.第 6 节总结全文.

1 社会网络中的隐私信息

在社会网络中,组成社会网络的各个元素均可能涉及到隐私信息,包括结点、边、图性质等.在本文中,社会网络隐私分类为结点隐私、边隐私、图性质隐私,表 1 给出了具体的分类结果以及为每种隐私提供保护的研究参考文献.

Table 1 Privacy information in social networks

表 1 社会网络中的隐私信息

社会网络隐私	结点隐私	结点存在性 结点再识别 结点属性值 结点图结构	文献 [13,16,17,19,22,24,25,29,30,37−39,41] [17,33−35]
	边隐私	边存在性 边再识别	[14,21,40] [18,21,23,30,32,41]
		边权重 边属性值	[20,36] [14,17]
	图性质隐私		

1.1 社会网络中的结点隐私

在社会网络中,每个结点代表了社会中的真实个体,而与结点相关的任何信息均有可能成为隐私.本文将结点隐私具体分类为结点存在性、结点再识别、结点属性值、结点图结构等隐私信息.

- 结点存在性

所谓结点存在性,是指某个人是否以结点的形式出现在某个社会网络中.在某些情况下,某些人会将自己出现在某特定社会网络视为隐私信息.如果某人将此视为隐私信息,发布数据时应防止攻击者结合背景知识推测

出该人存在此社会网络中.例如,传染病传播网络对于研究公共健康和疾病传播途径等方面具有很大价值,然而在发布传染病传播网络数据的同时,如果攻击者能够推断出某攻击目标存在于此传染病传播网络中,则导致了该攻击目标隐私信息的泄露.从表 1 中可以看出,目前针对保护结点存在性隐私信息的研究工作尚属空白.

- 结点再识别

在发布社会网络数据时,为了保护网络中实体的隐私信息,通常将所有结点的身份信息删除,使得攻击者不能识别和推测出攻击目标在社会网络中的准确位置.但是攻击者可以基于与攻击目标相关的背景知识对社会网络中的结点进行匹配识别[13,16,17,19,22,24,25,29,31,37−39,41],从而准确地或者以一定概率识别攻击目标在社会网络中的位置.在社会网络中,攻击者基于背景知识对攻击目标的位置进行匹配识别的过程称为结点再识别.

例如,图 1(b)是图 1(a)删除身份信息后的发布数据,如果攻击者掌握了 Ada 的 1-邻居子图(如图 1(c)所示),则可以推断出图 1(b)中的结点 6 是 Ada,从而准确地识别出 Ada 在社会网络中的位置,导致 Ada 隐私信息泄露.

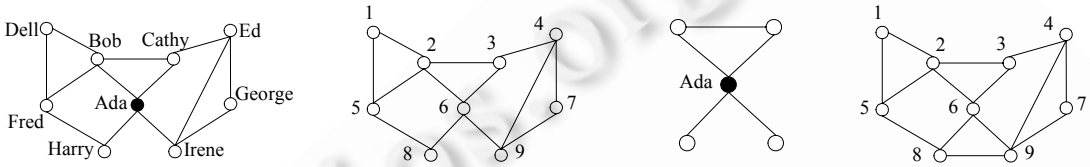


Fig.1 Neighborhood subgraph

图 1 结点邻居图

- 结点属性值

社会网络中的每个结点具有属性值,这些属性值描述了社会中每个人的真实信息,其中某些属性信息会涉及到个人隐私,例如收入信息、医疗记录中的患病信息等.发布社会网络数据时,结点之间的相互关系使得攻击者具有更多的背景知识推测目标结点的敏感属性信息.例如在家族遗传病史社会网络中,即使删除了某个重要结点的疾病信息,但是攻击者还可以基于其亲戚患有遗传疾病的情况,推测该目标结点可能患有的疾病.文献[17]提出采用结点 *K*-匿名的方法来保护结点的敏感属性值,而文献[33−35]显示了基于社会网络基本常识即可准确地推测出大部分结点的敏感属性信息.

- 结点图结构

不仅结点的某些属性值是敏感的,结点在社会网络中的图结构性质在某些情况下也被视为敏感和隐私,例如结点的度、两个结点间的最短距离、结点到社会网络中某个社区中心的距离等.例如在商品供货网络中,每个结点的入度和出度分别表示其供货渠道数目和销售渠道的数目,这些信息属于需要保护的敏感信息而防止其被竞争对手获得.表 1 所示了目前尚无相关工作针对保护结点的图结构隐私信息进行深入研究.

1.2 社会网络中的边隐私

在社会网络中,一条边表示其两端结点具有某种关系,结点由于相互间具有各种关系从而形成庞大的网络图.在某些情况下,边相关信息可能是敏感的,例如两点之间是否具有某种关系、参与某种敏感关系的结点信息、边权重、边的相关属性等.本文将边隐私具体分类为边存在性、边再识别、边权重、边属性值等隐私信息.

- 边存在性

所谓边存在性,是指社会网络中的两个指定结点是否具有某种关系.如果某两个结点的边是敏感的,简单地将此两个目标结点的敏感边删除并不能很好地保护隐私信息,攻击者可以通过背景知识推测两个目标结点是否具有敏感边.文献[14]假设攻击者采用 noisy-or 概率模型并基于现有结点之间的边连接来计算目标结点间具有敏感关系的概率,从而对可能被删除的敏感边进行恢复.在文献[40]中,通过实验验证了在真实社会网络数据上采用链接推演技术可以高概率地预测两个目标结点之间是否具有边连接.

- 边再识别

对于社会网络中的某条边,识别该边两端结点的过程称为边再识别.在社会网络中,每条边的两端连接着社会网络中的两个结点,表明两个结点所代表的个人具有某种关系,该关系可能被视为敏感信息.例如在异性交友网络中,两个结点之间的边表示了两个结点所代表的个人曾经具有男女朋友关系,显然,此种关系可能涉及个人隐私.文献[18,23,30,41]研究了如何使边再识别概率小于指定阈值.文献[32]同样将两结点之间的边连接视为隐私信息,并提出技术保证在不得知结点之间边连接情况的同时,较准确地计算任意两点之间的最短路径长度.

- 边权重

在不同应用背景中,社会网络中的边具有权重.在电子邮件通信网络中,边权重可以表示两个人之间收发电子邮件数目;在商业网络中,边权重可以表示两个商业公司之间的贸易额.类似商业公司之间的贸易额等边权重信息可能被视为敏感信息.在文献[20]中,研究了在防止边权重值泄露的同时保持某些重要结点间的最短路径不变;而文献[36]提出的技术在对边权重提供隐私保护的同时保证线性图性质不变.

- 边属性值

与结点属性值相似,社会网络中的边也可以具有属性值,例如边上的标签可以表示边两端结点的关系类型.边的敏感属性值对于边的两端结点所代表的个人来说属于隐私信息.文献[14,17]研究了在社会网络中,如何防止攻击者基于背景知识推测出边的敏感属性值.

1.3 社会网络中的图性质隐私

很多图性质是社会网络分析的重要评估标准,例如中间性(结点位于其他结点连接路径上的度)、中心性(结点与其他结点具有关系的数目)、路径长度(网络中两结点间的最短距离)、可达性(任意结点与其他结点联通的度)等.某些结点的图性质亦被视为个人隐私信息,目前尚无相关工作对结点图性质提供隐私保护.

对社会网络中的隐私信息进行分类归纳意义重大,因为社会网络中,不同类型隐私信息泄露均会威胁到个人隐私信息安全,只有对社会网络中的隐私信息做好辨识和分类工作,才能对不同隐私信息提出相应保护技术.从表 1 可以看出,社会网络中很多方面的隐私信息需要深入研究来为其提供保护.

2 攻击者背景知识

由于社会网络蕴含的信息具有多样化的特点,攻击者可以采用多种类型知识发动隐私攻击,对进行社会网络隐私保护提出很大挑战.本文将攻击者背景知识分类为社会网络图结构、结点信息、边信息、预测模型等方面,表 2 给出了具体分类结果以及每项研究工作所涉及的参考文献.

Table 2 Classification of adversaries' background knowledge

表 2 攻击者背景知识分类

攻击者背景知识	社会网络图结构	结点邻居图	Ref.[17,22,24,25,29]
		子图	Ref.[13,24,25,29]
		图查询	Ref.[19,24,25,29]
	结点信息	属性值	Ref.[17,23,30]
		度	Ref.[16,19,24,25,29,30,41]
	边信息	连接关系	Ref.[17,18,23]
		属性值	Ref.[30,37−39]
	预测模型	基于邻居	Ref.[14,[33,35,40]
		基于兴趣组	Ref.[33,34]

2.1 社会网络图结构

攻击者可以将结点间连接情况,即社会网络图结构,作为背景知识来进行隐私攻击.社会网络图结构可具体分类为结点邻居图[17,22,24,25,29]、社会网络子图[13,24,25,29]、图查询[19,24,25,29]等方面,为攻击者提供图结构背景知识.

- 结点邻居图

在社会网络中,将距离结点 u 长度 d 之内的所有结点称为 u 的 d-邻居结点,u 的 d-邻居结点及其相互之间的

边构成的子图称为结点 u 的 d-邻居子图.结点邻居图是一种常见的图结构背景知识[17,22,24,25,29].

图 1 给出采用 1-邻居子图进行隐私攻击的实例,例如:图 1(c)显示了 Ada 的 1-邻居子图,而图 1(b)中只有结点 6 的 1-邻居子图与 Ada 相同,因此,攻击者可以在图 1(b)中唯一识别出结点 6 是 Ada,从而导致 Ada 隐私泄露.

- 子图

在社会网络图中,攻击者可以将具有特殊连接模式的子图[13,24,25,29]作为背景知识,从而为其进行隐私攻击提供结构唯一性的识别标记.

文献[13]针对结构唯一性子图导致隐私泄露的可行性进行了研究:在发布社会网络数据前,攻击者嵌入具有结构唯一性的子图,并建立该子图与目标结点之间的连接,当匿名化的社会网络数据发布后,攻击者首先识别嵌入子图,然后基于嵌入子图和目标结点之间的联系来识别目标结点.通过实验显示,嵌入由 7 个结点构建的特殊子图平均可以识别出 70 个目标结点.

- 图查询

在社会网络中可以执行多种图查询,而针对某些结点或者边的图查询结果具有唯一性,从而为攻击者提供了进行隐私攻击的背景知识[19,24,25,29].

例如:对于结点 v,定义查询 $Q(v)$为 v 的所有邻居结点度的升序序列.在图 1(a)中,Q(Fred)=[2,2,4].如果攻击者将 Fred 的朋友的度信息作为背景知识,则可以在图 1(b)中识别出结点 5 即是 Fred,因为只有结点 5 的度序列与 Fred 相同.文献[19]评估了不同图查询作为背景知识的隐私攻击能力;而文献[24,25,29]虽然没有定义可导致隐私泄露的图查询,但其提供的隐私保护技术可以防御部分[29]或者全部[24,25]图查询导致的隐私泄露.

2.2 结点信息

对于某些社会网络隐私攻击,尤其是结点隐私攻击,攻击者会将结点自身的一些相关信息作为背景知识.

- 结点属性值

社会网络中结点的属性值可以分类为标识属性和敏感属性.标识属性为攻击者提供了结点识别的背景知识,例如年龄、性别、籍贯、学历等,攻击者可以将网络中的结点标识属性值和其掌握的实体属性值进行链接匹配,从而识别结点的真实身份[17].文献[17,30]研究了如何防范基于结点属性值的结点再识别隐私攻击,而文献[23]侧重研究攻击者基于结点属性值进行边再识别隐私攻击.

- 结点度

在社会网络中,结点度表示了该结点所代表的实体与社会中的其他实体之间的关系数目,在现实中,攻击者很容易收集到目标的度信息,并作为背景知识进行结点再识别[16,19,24,25,29,30]、边再识别[30,41]等隐私攻击.

图 2 描述了如何基于结点度进行结点再识别攻击.图 2(a)只有结点 A 的度为 2,其他结点的度均为 1,因此,当攻击者掌握 A 的度为 2 的背景知识时,可以很容易地识别出 A 在社会网络中的位置.相似地,攻击者可以基于目标结点的度进行边再识别攻击[30,41].在文献[41]中,假设攻击者背景知识为互为邻居的两个结点的度,例如图 2(a)中结点 C,D 的度对(1,1).由于图 2(a)中具有度对(1,1)的边只有一条,所以攻击者识别出边 CD 的成功概率为 100%.

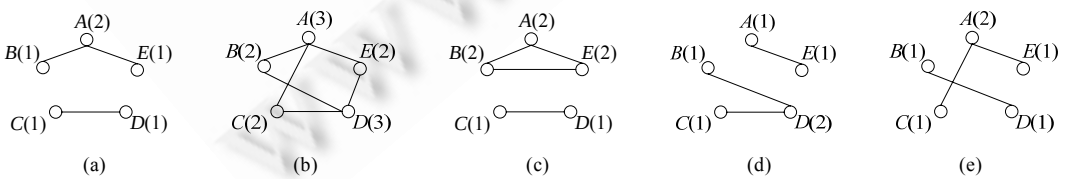


Fig.2 Privacy attack based on degrees

图 2 基于结点度的隐私攻击

2.3 边信息

社会网络中,连接结点的边是其重要的组成部分,攻击者可以将边的相关信息作为背景知识,包括边连接关

系[17,18,23]、边属性值[30,37−39]等.

- 连接关系

如果攻击者事先掌握了某些目标的边连接关系,则可以根据这些连接关系进行推演,从而获得隐私信息.文献[17,18,23]研究了连接关系可能导致的隐私泄露.参照表示基于连接关系的隐私攻击的图 3,如果朋友关系被视为敏感关系,则可以基于图 3(a)中 u_1 和 u_2 与结点 $friend_1$ 的连接关系推断出 u_1 和 u_2 具有朋友关系的隐私信息.

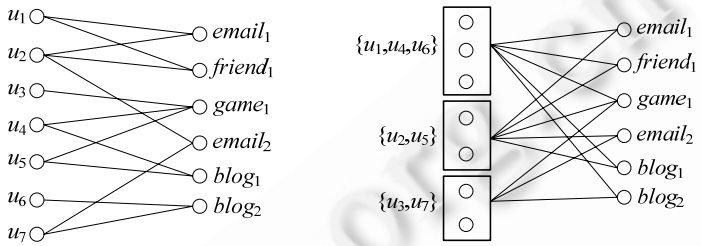


Fig.3 Privacy attack based on relations

图 3 基于连接关系的隐私攻击

- 边属性值

边上的属性值(标签、权重等)可以为攻击者提供隐私攻击的背景知识.例如在朋友网络中,边标签表示朋友之间的联系方式,可以是电话、短信、电子邮件等.如果攻击者知道某目标基本上仅采用电子邮件与其他朋友联系,基于此背景知识,攻击者能够以很大概率在社会网络中识别出这个目标结点.在加权社会网络图中,边权重可以作为攻击者的背景知识.文献[37−39]研究了加权图中目标结点与其他结点相连接的边权重信息如何导致身份泄露.对于结点 v,将与 v 相连接的边权重按照降序排序得到的序列定义为结点 v 的权重包,记作 $\boldsymbol{w}_v$.例如,图 4(a)中结点 A 权重包为 $\boldsymbol{w}_A=[w_{AB},w_{AD}]=[2,1]$.如果攻击者掌握了结点 A 的权重包信息,则可以识别出图 4(b)中的结点 1 即为 A,从而导致了身份泄露.

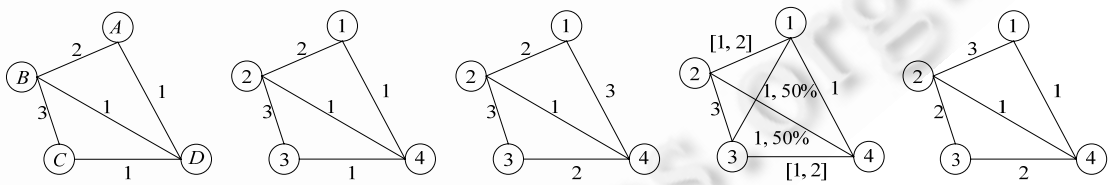


Fig.4 Anonymization of weighted graphs

图 4 加权图匿名

2.4 预测模型

攻击者可以基于社会网络常识构建预测模型,从而推演目标的隐私信息.当前,社会网络中的预测模型主要分为两类:基于邻居的预测模型[14,35,40]和基于兴趣组的预测模型[33,34].

- 基于邻居的预测模型

所谓物以类聚,人以群分,在社会网络中,此种现象尤为明显.一般情况下,具有朋友关系的实体具有相同或相似的属性值,攻击者可以根据邻居属性值来推断目标的敏感属性值.在文献[35]中,研究了采用贝叶斯网络来推演目标的敏感属性值.相似地,可以通过链接推演技术来预测和恢复社会网络中的敏感关系.很多链接推演技术均是基于社会人际交往常识,其中一项常识是:如果两个人具有很多共同朋友,则他们也很有可能是朋友.文献[40]评估了在真实数据集 Email-1 和 LiveJ-1[29]上采用链接推演技术预测敏感关系的可行性.在实验测试中,如果两个结点的共同邻居数目大于阈值 δ,则认为两者在图中具有边连接.实验结果表明:当 δ 增大时,正确预测

率逐渐增高;当δ=20 时,Email-1 和 LiveJ-1 数据集上的正确预测率分别达到了 91.06%和 66.5%.可以看出:攻击者可以凭借链接推演技术,以较高的概率推断出社会网络中的敏感关系.

在社会网络中,结点之间具有不同的关系.基于常识可以知道,各种关系之间不是相互独立而是相关的.例如,具有同学关系的两个人是朋友的概率比没有任何关系的两个人是朋友的概率大.在文献[14]中,研究了通过非敏感关系边采用 noisy-or 概率预测模型[26]来预测敏感关系.e_{ij}^s=1 表示结点 i 和 j 具有敏感关系 s,如果边 $e_k(k=1,\ldots,n)$的影响参数是λ_k,e_k 对于 e_{ij}^s 的影响是相互独立的,并且所有观察边对于 e_{ij}^s 的影响参数是λ_0,则基于 noisy-or 概率模型得到结点 i 和 j 具有敏感关系 s 的概率为

$$P(e_{ij}^s=1)=P(e_{ij}^s=1\mid e_1,\ldots,e_n)=1-\prod_{k=0}^{n}(1-\lambda_k) \tag{1}$$

基于公式(1)计算的概率,即可对 i 和 j 是否存在敏感关系进行推测.

- 基于兴趣组的预测模型

在社会网络中,实体加入不同的兴趣组,比如在豆瓣网中,每个用户可以凭借自己爱好加入诸如摄影、影视等方面的兴趣组.利用实体之间的朋友关系、加入兴趣组情况,可以对实体的隐私属性进行推测[33,34].其基本思想是:参加相同兴趣组的两个实体具有相同属性值的概率较大;参加相同兴趣组的数目越多,则两个实体具有相同属性值的概率越大.

在文献[34]中,基于实体参加兴趣组的情况,采用贝叶斯法则来推测未知属性值.由于每个兴趣组中组员属性值分布不同,即每个兴趣组对属性值的预测能力不同,文献[34]提出了兴趣组细化的贝叶斯分类器,可以较高概率地预测未知属性值.在文献[33]中提出的预测模型中,不仅考虑了实体参与兴趣组情况,也结合了实体之间的朋友关系,其属性值预测准确率高于文献[34]中的预测模型.

3 社会网络数据隐私保护技术

针对不同背景知识可能导致的隐私泄露,提出了相应的社会网络隐私保护技术.本节分别从隐私保护方法、动态性、并行性等方面介绍当前社会网络隐私保护技术,并指出不同隐私保护技术的优缺点.

表 3 给出了当前社会网络隐私保护技术的具体分类结果.

Table 3 Privacy preserving techniques in social network

表 3 社会网络隐私保护技术

文献	隐私保护方法					动态性	并行性
	结点 K-匿名	子图 K-匿名		数据扰乱			
	结点聚类	加伪点加伪边	删除边概括	数值	图结构增、删边交换端点	推演控制	
16		√					
17	√						
18	√						
19	√						
20				√			
21					√ √		
22		√	√				
23	√						
24		√	√			√	
25		√ √					
29		√				√	
30		√ √	√				
31		√					
32							√
36				√			
37		√	√				
38		√					
39		√					
40						√	
41		√					
42	√					√	

3.1 隐私保护方法

社会网络隐私保护方法主要分为结点 *K*-匿名、子图 *K*-匿名、数据扰乱、推演控制这 4 种.

- 结点 *K*-匿名[17−19,23,42]和子图 *K*-匿名[16,22,24,25,29−31,37−39,41]的主要思想是:攻击者基于目标背景知识在匿名化社会网络数据中进行匹配识别时,至少有 *K* 个候选符合,即目标的隐私泄露概率小于 1/*K*;
- 数据扰乱[20,21,36]的主要思想是:对社会网络进行随机化修改,使得攻击者不能准确地推测出原始真实数据,数据扰乱方法具体分为数值扰乱[20,36]和图结构扰乱[21];
- 推演控制[40]的主要思想是:对于不同隐私预测模型,通过对社会网络进行针对性地修改,使得攻击者采用预测模型不能推演出隐私信息,从而起到保护社会网络隐私的目的.

3.1.1 结点 *K*-匿名

所谓结点 *K*-匿名,是指通过将社会网络中所有结点聚类成若干超点,其中每个超点至少包含 *K* 个结点,由于在超点中结点相互之间不可区分,因此在该社会网络中,受结点再识别攻击而导致隐私泄露的概率小于 1/*K*.

图 5 显示了结点聚类与可能社会网络,图 5(b)给出了图 5(a)的一个结点聚类图,每个超点记录了其内部结点间边连接数目,两个超点之间边的数目等于端点分别为两个超点内部结点的边的数目.

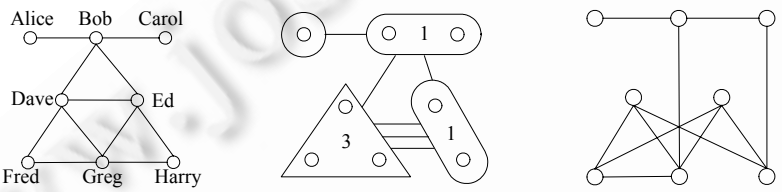


Fig.5 Clustering vertices and possible social networks

图 5 结点聚类与可能社会网络

显然,结点聚类成超点导致了边两端结点的信息损失,增加了图结构不确定性,降低了数据可用性.假设匿名图 *G* 的超点集为 *V*,则 *G* 的可能社会网络数目 *W*(*G*)可以通过公式(2)计算得到,其中,*d*(*X*,*X*)表示超点 *X* 内的边数目,*d*(*X*,*Y*)表示超点 *X* 和 *Y* 之间的边数目.例如,图 5(b)表示了 960 个可能社会网络,图 5(c)为图 5(a)的一个可能社会网络.

$$|W(G)|=\prod_{X\in V}\binom{\frac{1}{2}|X|(|X|-1)}{d(X,X)}\prod_{X,Y\in V}\binom{|X||Y|}{d(X,Y)} \tag{2}$$

在文献[19]中,研究如何通过结点聚类实现结点 *K*-匿名的同时最小化|*W*(*G*)|,其提出的技术主要基于模拟退火思想.文献[17]在文献[19]基础上做了改进,与文献[19]中研究简单社会网络不同,文献[17]假设社会网络中的每个结点具有属性信息,通过结点聚类生成超点时,每个超点内所有结点的属性信息还需要进行匿名化处理使得属性值相等,因此不仅会造成图结构信息损失,也会造成结点属性值的信息损失.文献[17]提出一种贪心聚类方法来实现复杂社会网络的结点 *K*-匿名.由于文献[17]提出的匿名算法需要数据发布者通过设定权重来决定图匿名过程侧重于减少图结构信息损失还是结点属性信息损失,而两者的数据可用性难以量化,使得在实际应用中无法设定所需的权重,导致文献[17]中方法的实用性较差.文献[18,23]采用结点 *K*-匿名来隐藏二部图社会网络中的敏感关系.图 3(b)给出了基于图 3(a)进行结点 *K*-匿名化后的二部图社会网络数据.结点 *K*-匿名隐私保护能力强,具有很好的通用性,可以防止多种类型隐私泄露.然而,结点 *K*-匿名在提供强隐私保护的同时,导致了图数据可用性降低,并且结点 *K*-匿名的执行效率低,不适用于大型社会网络数据.

3.1.2 子图 *K*-匿名

所谓子图 *K*-匿名,是指当攻击者将目标所在的特定子图作为背景知识进行隐私攻击时,社会网络中至少有

*K*个子图可作为候选,则目标子图导致隐私泄露的概率小于 1/*K*.与目标相关的标识性子图均可作为攻击者的背景知识,例如结点的度、邻居图等.通过在社会网络中加伪点、加伪边、删除边、概括等,可实现子图 *K*-匿名.

文献[16]研究了攻击者将结点的度作为背景知识时如何进行子图 *K*-匿名,提出了采用 *K*-度匿名图来防止此类攻击.所谓 *K*-度匿名图,是指对于该图中的任意结点,至少有 *K*–1 个结点与该点的度相同.例如,图 2(b)、图 2(c)均为 2-度匿名图.文献[16]通过采用动态规划的方法实现了通过加入最少数目的边来生成 *K*-度匿名图.在文献[30]中,每个结点可以设定所需的隐私保护级别:第 1 级别为防止结点标签导致身份泄露;第 2 级别为防止结点标签、度导致身份泄露;第 3 级别为防止结点标签、度、结点边标签导致身份泄露.对于需要第 2 级别和第 3 级别隐私保护的结点,文献[30]对其进行 *K*-度匿名化操作.文献[31]在生成 *K*-度匿名图的同时,最小化社团结构信息损失.由于攻击者能够获得比结点度更复杂的目标子图背景知识,因此 *K*-度匿名的隐私保护能力较差.

当攻击者将目标的 1-邻居图作为背景知识时,文献[22]提出的匿名化方法使得对于任意结点的 1-邻居图,至少有 *K*–1 个结点的 1-邻居图与其同构.在匿名过程中,需要加入伪边和概括结点标签.图 1(d)给出了 *K*=2 时图 1(a)的匿名图.可以看出:具有唯一邻居图的结点 6 在匿名后与结点 2 和结点 9 具有相同的邻居图,因此攻击者基于结点 6 的 1-邻居图获得其真实身份的概率小于 1/2.为了防范攻击者将任意目标子图作为背景知识,文献[25]提出将社会网络匿名化为 *K*-对称图.所谓 *K*-对称图,是指对于图中的任意结点 *v*,在图中至少存在 *K*–1 个结点与 *v* 是结构对等的.例如,图 6(b)是图 6(a)的 2-对称图,其中添加的伪点 v'_3 与 v_3 结构对等,因此攻击者识别出 v_3 的真实身份概率为 1/2.显然,为了构建 *K*-对称图,匿名化时需要在社会网络数据中加入伪点和伪边.*K*-对称图的潜在隐私威胁是:当攻击者获知 *K*-对称图生成算法时,可以将发布的 *K*-对称图中结构对等的结点进行合并,还原出部分原始图,从而导致隐私泄露.文献[29]提出 *K*-自同构来进行隐私保护.所谓 *K*-自同构,是指图自身存在着 *K* 个同构映射.*K*-自同构能够阻止结点再识别隐私攻击,但是不能防范敏感关系隐私攻击.为了能够同时保护结点和边隐私,文献[24]提出 *K*-同构隐私保护模型.所谓 *K*-同构,是指社会网络图分为 *K* 个子图,子图之间相互同构.为了实现 *K*-同构,首先需要将社会网络图分割为 *K* 个包含相同数目结点的子图,然后通过加入伪边和删除边的方法使得 *K* 个子图同构,增删边的数目和图数据可用性的大小主要取决于图分割策略.例如,图 6(d)为图 6(c)的 4-同构图.*K*-同构虽然能够很好地保护结点和边隐私,但是同构子图之间的边连接会被删除,导致位于同构子图之间的图模式会受到影响.

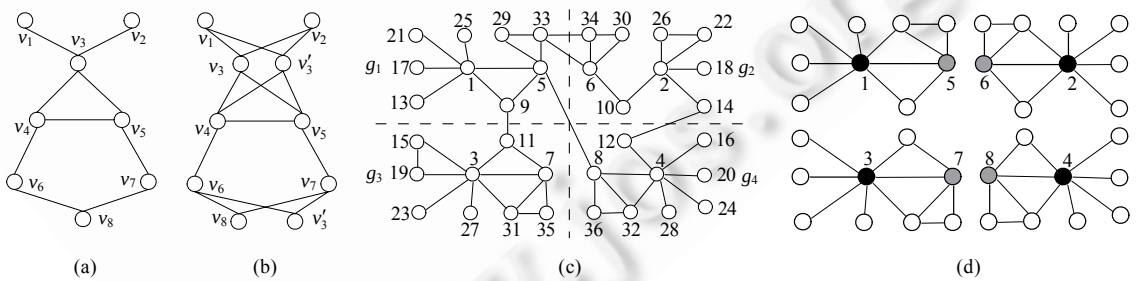


Fig.6 *K*-Symmetry graph and *K*-Isomorphism graph

图 6 *K*-对称图和 *K*-同构图

目标与邻居连接边上的属性值序列可以作为隐私攻击的背景知识[30,37−39].在文献[30]中,通过加入伪点、伪边、设置边标签来弱化邻居连接边标签序列的标识作用.文献[37–39]研究了如何阻止权重包导致的隐私泄露.文献[38]提出的加权图匿名化方法可以使得对于任意结点的权重包,至少有 *K*–1 个其他结点的权重包与其距离小于预先设定阈值,而不是完全相同;而文献[39]提出的 HA(histogram anonymization)算法使得至少有 *K*–1 个其他结点的权重包与其相同.对图 4(a)中的 *G* 采用 HA 算法匿名化后得到图 4(c)中的 G_{HA}.显然,HA 算法对于边及权重值的修改导致了图数据的统计特性的改变.例如,对于图查询 *Q*:SELECT COUNT($\forall edge \in G$) WHERE $edge.weight \geqslant 2$,在 G_{HA} 中 *Q* 的查询结果是 4,而 *Q* 在原图 *G* 中的查询结果是 2,两者具有一定的偏差.为了提高

匿名加权图的数据可用性,文献[37]提出边权重概括技术(记作 GA 算法)来实现 K-可能图.所谓 K-可能图,是指对于任意结点的权重包,在 K-可能图中可以找到至少 K 个为其概括实例的权重包.例如在图 4(d)的 G_{GA} 中,w_{12}=[1,2]表示边 e_{12} 的权重值位于区间[1,2],边 e_{13} 上的 50%表示 e_{13} 的存在概率,因此,G_{GA} 中权重包 $\boldsymbol{w}_1$ 可以表示为[([1,2],100%),(1,50%),(1,100%)].在图 G_{GA} 中,结点 1 可被推测为图 4(a)中 G 的 A,因为 A 的权重包 $\boldsymbol{w}_A$ 是 $\boldsymbol{w}_1$ 的一个可能实例;相似的,$\boldsymbol{w}_A$ 也是 $\boldsymbol{w}_4$ 的一个可能实例;因此,A 被准确识别的概率小于等于 1/2.在 G_{GA} 中执行查询 Q 时,边 e_{23} 符合 Q 的查询条件,边 e_{12} 和 e_{34} 可能符合 Q 的查询条件,其他边不符合 Q 的查询条件,则 Q 的查询结果为区间[1,3],包含了 Q 在原图 G 上的查询结果 2.实验结果证明,GA 算法很好地保持了加权图数据可用性.

3.1.3 数据扰乱

图数据扰乱隐私保护方法的基本思想是:通过对社会网络图进行随机化修改,使得攻击者不能准确推测出原始真实数据,从而起到保护社会网络数据隐私的作用[18,20,21,36].本文分别从数值扰乱[20,36]和图结构扰乱[18,21]等方面介绍基于数据扰乱思想的社会网络隐私保护技术.

- 数值扰乱

社会网络中可以记录大量的数值信息,通过对数值信息进行随机化的扰乱和修改,可以使得攻击者不能猜测出原始真实数值.目前,数值扰乱[20,36]方法主要用于为加权图中的边权重提供隐私保护.

文献[20]研究了通过扰乱技术保护社会网络边权重隐私的同时,降低扰乱噪声对于社会网络中两点间的最短路径序列及最短路径大小的影响.对于动态社会网络,文献[20]提出在边权重中加入高斯噪声进行扰乱:

$$w_i^* = w_i(1 - x_i) \tag{3}$$

其中,$_i$ 和 w_i^* 分别表示边 i 的初始权重、扰乱后权重;x_i 表示加入的高斯噪声,x_i 服从高斯分布 $N(0,\delta^2)$.例如,图 4(e)显示了对图 4(a)采用服从 $N(0,0.15^2)$分布的高斯噪声扰乱边权重后的社会网络.在静态社会网络中,为了保持指定结点对的最短路径序列及其大小不变,文献[20]通过将社会网络数据中的边分类从而提出贪心扰乱算法,侧重扰乱其他边的权重.文献[20]中的扰乱方法可以高效率地在边权重中加入噪声,并保证指定结点对的最短路径及其大小不变,但是噪声对于边权重的影响不大,仍然会泄露边权重隐私,而且不能保证所有结点间的最短路径保持不变,数据可用性不高.文献[36]提出线性规划模型构建方法,在对边权重进行扰乱的同时,保持加权图的线性图性质,例如最短路径、K-最近邻等.与文献[20]中仅能保持指定结点对的最短路径相比,文献[36]中的方法较大程度地保证了加权图的数据可用性.

- 图结构扰乱

通过随机进行图数据扰乱和修改,可以阻止攻击者获知原始图结构,从而保护社会网络数据隐私[21].图扰乱的主要方法是随机添加、删除边和交换边端点等.例如,参照图 2,图 2(a)随机添加(B,D)、删除边(B,A)后得到图 2(d),随机交换边(B,A),(C,D)的端点 A 和 D 得到图 2(e).

很多图性质均与图谱相关,例如平均最短路径、社团结构、传递性等.为了保持图性质和图数据可用性,文献[21]研究了如何进行图扰乱的同时保持图谱基本不变.文献[21]指出,图谱主要由两个参数所决定:(1) 图邻接矩阵最大特征值λ_1;(2) 图拉普拉斯矩阵次最小特征值μ_2.通过研究图修改操作对于λ_1和μ_2的影响,文献[21]提出的随机图扰乱技术总是选择保持λ_1 和μ_2 基本不变的图修改操作执行,从而保持图谱不变.虽然扰乱图在一定程度上保护了图数据隐私,但是存在明显缺陷:(1) 与 K-匿名图提供量化隐私保护不同(隐私泄露概率不大于 1/K),图随机扰乱方法无法保证量化隐私保护,扰乱图中仍然存在隐私泄露威胁;(2) 采用文献[21]提出的图扰乱方法的必要条件是社会网络图是连通的,然而实际社会网络图一般不具有连通性,因此需要加入新边将图中的独立部分连接起来,使其具有连通性,引入了图噪声;(3) 图特征值的计算代价很高,然而一次随机边修改操作需要多次图特征值的计算,导致边修改操作计算代价高,图扰乱需要大量的边修改操作,因此,文献[21]的图扰乱算法的实际应用性不高.

3.1.4 推演控制

所谓推演控制,是指对于不同隐私预测和推演模型,针对性地修改社会网络,使得攻击者采用预测模型不能推演出隐私信息,起到保护社会网络隐私的目的.在第 2.4 节中分别给出了基于邻居的预测模型[14,35,40]和基于兴

趣组的预测模型[33,34],然而,只有文献[40]给出相应的推演控制技术来防止隐私泄露,因此,社会网络推演控制技术需要引起更多的关注.

在文献[40]中,首先提出了基于共同邻居数目的敏感关系预测模型,并定义了两种链接推演攻击:单步链接推演攻击和级联链接推演攻击.单步链接推演是指对于图上的所有无边连接的结点对执行链接推演操作;级联链接推演,是指在图上执行多次单步链接推演操作.其次,为了阻止链接推演攻击,提出了一种基于链接世系溯源的防推演机制来切断敏感链接的推演路径,在保护社会网络中敏感关系的同时,保持了图数据可用性.推演控制技术能够有效地防止特定预测模型导致的隐私泄露,由于其针对性地修改社会网络,可以保持图数据的高可用性.但是推演控制技术的隐私保护能力有限,对于图数据隐私保护不具有通用性.

3.2 动态性

当社会网络静止不变时,称该社会网络是静态社会网络;当社会网络是不断发展和变化时,该社会网络是动态社会网络,具有动态性.社会网络的动态性具体表现在:(1) 不断有新结点加入社会网络中,原有结点从社会网络中退出,即结点的添加和删除;(2) 社会网络中,两个无关系的实体建立新的边连接,网络中的某条边会被两个端点(实体)删除,即边的添加和删除.当前,社会网络隐私保护技术主要面向静态社会网络,而在现实中,几乎所有的社会网络都是动态的、不是静止不变的.从表 3 中可以看出,仅有少数隐私保护技术考虑了社会网络的动态性[24,29,42].

在文献[24,29]中,除了给出面向社会网络的子图 K-匿名技术,还研究了如何防止社会网络动态性和多次发布可能导致的隐私泄露.文献[24,29]提出采用随机结点 ID 编码来保证动态发布匿名社会网络的安全性,其基本思想是:每次发布匿名社会网络时,同一结点被赋予不同的ID,从而阻止了攻击者基于网络变化信息获得数据隐私.显然,结点 ID 的重新编码不利于观察和分析网络的变化趋势,降低了动态网络的数据可用性.

与文献[24,29]中研究如何防止网络动态性导致隐私泄露不同,文献[42]研究了如何利用网络动态性来提高图匿名算法的执行效率.对于动态社会网络,当发布最新版本图数据时,基本做法是在当前图数据上重新运行图匿名化算法,当数据发布比较频繁时,则导致较低的执行效率.为了提高动态社会网络匿名算法的执行效率,文献[42]提出采用动态网络预测技术来预测网络的变化趋势,例如某些无边连接的结点对在未来可能会增加边连接等,用于指导当前图数据的匿名化过程,目的在于减少未来图匿名化的重新计算量和负担,从而提高了动态发布过程中图匿名化的执行效率.

3.3 并行性

随着网络技术的发展,社会网络数据的数量和规模都在不断地增长,呈现海量化趋势.对于海量社会网络数据,采用并行算法进行分析和处理,是提高效率的有效途径.从表 3 中可以看出,目前,仅文献[32]研究了云环境中社会网络隐私保护,而其他研究工作主要面向单工作站的社会网络隐私保护技术,不适用于海量社会网络数据.

在文献[32]中,研究了在云环境中进行最短路径查询的同时保护图数据隐私,其研究目标是:攻击者不能推演出加权图中每个结点的邻居,数据查询者可以得到任意两点间的近似最短路径.文献[32]中提出的云环境中隐私保护最短路径查询技术的基本思想是:将加权图 G 转换为链接图 G_l 和外包图集 G_o,其中,G_l 相当于加权图的索引;而 G_o 中的每个外包图符合“1-邻居-d-半径”安全要求,即攻击者无法获知任意结点的邻居以及距离小于 d 的结点对,对于输入的最短路径查询,基于 G_l 找到相应的外包图进行最短路径的求解;每个外包图存储在云环境的节点中,记录了符合安全条件的结点对之间的最短路径距离,通过采用三角不等式来逐步求精最短路径查询结果.由于在构建链接图 G_l 和外包图集 G_o 时需要计算大量结点对之间的最短路径,使得大型加权图的分割和预计算的工作量很大,当网络动态变化时,链接图 G_l 和外包图集 G_o 均需要重新计算,而文献[32]没有考虑如何动态更新链接图 G_l 和外包图集 G_o.

4 数据可用性与实验评测

社会网络图匿名化会导致一定的信息损失,影响图数据可用性.不同社会网络隐私保护技术对图数据可用

性产生不同的影响,需要通过实验测试来分析和评估匿名化对数据的影响.本节归纳了常用的社会网络隐私保护技术的实验评测指标,其中包括结点数据可用性、边数据可用性、图结构及性质、图查询、执行效率等方面,具体结果见表 4.

Table 4 Influence on data utility and experimental analysis

表 4 数据可用性影响与实验评测

结点数据可用性	属性值	Ref.[14,17,22]
	增加结点数目	Ref.[25,43]
边数据可用性	属性值	Ref.[17,37−39]
	增加边数目	Ref.[22,25,41]
	删除边数目	Ref.[14,40,41]
图结构及性质	度分布	Ref.[14,16,17,19,24,25,29,30,37,41]
	最短路径	Ref.[16,19−21,24,25,29,31,36,37,39−41]
	传递性	Ref.[16,19,21,24,25,29,31,39−41]
	网络适应力	Ref.[19,25]
	传染性	Ref.[19]
	社区结构	Ref.[31]
	图谱	Ref.[21,39]
图查询		Ref.[18,22,23,30,32,42]
执行效率		Ref.[22,24,29,32,36,37,40,41]

在社会网络中,结点可能会具有表示类别的标签、与实体相关的属性值等信息,边可能会具有标签、权重等信息,而图匿名技术通常会对这些属性值进行修改、概括(generalization)等匿名化操作,因此需要评估图匿名化导致的结点[14,17,22]和边[17,37−39]的属性值信息的损失.如第 3 节中,添加和删除边是图匿名化中最基本的图修改操作,可以将图匿名化过程中的边增加[22,25,41]和删除[14,40,41]数目作为一种图信息损失的度量.特殊地,文献[25]通过加入结点来获得 K-对称图,因此不仅评测了边的添加数目,同时也测试了不同隐私要求下加入结点的数目.

在图结构及性质的实验测评中,结点度分布[14,16,17,19,24,25,29,30,37,41]、最短路径[16,19−21,24,25,29,31,36,37,39−41]、传递性[16,19,21,24,25,29,31,39−41]是比较常见的图数据可用性度量标准.结点度分布是图中不同结点度的频率统计,是描述图状态的一种基本图性质.最短路径分布统计了图中两点间最短距离的分布,由于社会网络中结点数目巨大并且最短路径计算代价大,因此在实验测试中计算最短路径分布时,通常随机选取指定数目的结点对来进行计算.所谓传递性(又称聚集系数),是指一个结点所有邻居对中具有边连接的比例,即描述了“一个人的两个朋友也是朋友的概率”.当按照度由大自小删除图中结点时,网络适应力[19,25]表示图中最大社区所包含的结点数目,网络适应力描述了由于网络攻击导致部分结点通信不畅时网络的连通性.所谓传染性[19],是指对于一种假想疾病,当随机选择某个结点作为传染源时,在指定传染率下被传染结点的比例.为了测量社会网络图中社区结构变化大小,在文献[31]中定义了层次社区熵(hierarchical community entropy)来计算图匿名化所导致的社区结构的变化.图的很多性质均与图谱相关,可以通过图谱的变化[21,39]来衡量图匿名化对于图数据可用性的影响,在实验测试中,主要关注图邻接矩阵最大特征值λ_1和拉普拉斯矩阵次最小特征值μ_2的变化情况.

在社会网络图中进行图查询是一项重要应用,因此,图匿名化对于图查询[18,22,23,30,32,42]结果的影响也是社会网络隐私保护技术的一项重要测评指标.可以从多个方面评估图查询结果的变化,包括查询错误率、真实结果覆盖率等.对于执行效率,主要是测试图匿名化算法的执行时间[22,24,29,32,36,37,40,41].

5 未来研究趋势

社会网络隐私保护是一个新兴的研究方向,尚有许多值得深入探索的问题.在本文的最后,我们基于大量的调研和近几年来的研究经验,提出一些值得进一步挖掘的研究点,希望对本领域的其他研究者有所启发.

- 深入研究并行化社会网络隐私保护技术

当前,基于单工作站的社会网络分析和隐私保护技术不适合海量社会网络数据,例如,对于 Facebook 这种用

户数目达到上亿级别的社会网络,单工作站的社会网络算法的执行效率、数据处理能力均不能满足实际应用需求.因此,有必要研究并行化社会网络隐私保护技术.基于网络和并行计算思想的云计算技术使得进行社会网络海量数据的并行化分析和隐私保护成为了可能,例如,文献[32]初步尝试了云环境中的社会网络隐私保护研究.可以从两方面深入研究并行化社会网络隐私保护技术:1. 隐私保护的并行化社会网络分析;2. 并行化社会网络隐私保护算法.对于隐私保护的并行化社会网络分析,侧重研究并行化社会网络分析中如何防止隐私泄露;对于并行化社会网络隐私保护算法,侧重研究如何将现有的隐私保护技术和模型移植到并行计算环境中.不论对于哪种研究方向,并行化社会网络隐私保护技术都会面临无法载入海量数据、基于分割的图数据无法得到正确结果、数据处理效率非常低等诸多困难,需要深入研究并解决相应难点,实现社会网络隐私保护的并行化计算.

- 支持丰富数据应用的社会网络隐私保护

如前所述,在当前社会网络隐私保护研究中,并未指定发布数据的用途,而现实中发布的社会网络数据常被用于各种特定用途,例如进行社区中心发现、链接挖掘、可达性计算等.以前的研究工作并未基于数据发布用途来设计相应的隐私保护方法,而只是设计了通用的隐私保护方法,影响了发布数据的可用性.因此,有必要基于发布数据的用途实现社会网络隐私保护的定制化,从而提高发布数据的可用性.例如,文献[31]研究了保持图社区结构的图匿名化技术,开启了支持指定数据应用的社会网络隐私保护研究.图匿名过程包含了边添加和删除等操作,会对结点之间的可达性造成影响.如何在实现图匿名的同时减少结点间可达性的影响,是一个挑战性问题.

- 阻止社会网络预测模型导致的隐私泄露

在第 2.4 节中介绍了攻击者采用各种预测模型推演隐私信息,然而,目前只有文献[40]研究了如何阻止基于预测模型推演获得隐私信息.因此,有必要研究防范不同社会网络预测模型的隐私保护技术.例如,文献[40]仅研究了如何防范基于相似度的敏感链接推演攻击,没有对最大似然链接推演攻击和概率模型链接推演攻击给出隐私保护方法.隐私推演模型的复杂性和图中结点、边之间的高度相关性,对研究相应的隐私保护技术提出了挑战.

- 社会网络隐私保护模型亟待多样化

当前,社会网络隐私保护技术基本采用 *K*-匿名、数据扰乱和推演控制等隐私保护思想.由于隐私保护模型和方法缺乏多样性,从而导致隐私泄露威胁大、数据可用性低等缺点,亟待提出多样化的社会网络隐私保护模型.例如,相关工作已初步尝试将关系数据中的差分隐私移植到社会网络隐私保护中[43,44].然而,结点间的高度相关性以及大数据规模会导致图数据差分隐私的高复杂度,如何降低图差分隐私复杂度是一个挑战性问题.

6 总 结

本文在充分调研和深入分析的基础上,对社会网络隐私保护的研究进展进行了综述,分别从社会网络中的隐私、攻击者背景知识、社会网络数据隐私保护技术、数据可用性与实验评测等方面对现有研究工作进行了细致的分类归纳和分析,详细阐述了各种社会网络数据隐私保护的原理,指出了当前社会网络隐私保护存在的不足以及不同社会网络隐私保护技术间的对比和优缺点,最后对未来研究趋势进行了展望.

References:

[1] Getoor L, Diehl CP. Link mining: A survey. ACM SIGKDD Explorations Newsletter, 2005,7(2):3−12. [doi: 10.1145/1117454.1117456]

[2] Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information. In: Proc. of the 7th ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems. 1998. 188−202. [doi: 10.1145/275487.275508]

[3] Sweeney L. *K*-Anonymity: A model for protecting privacy. Int'l Journal on Uncertainty, Fuzziness and Knowledge-Based System, 2002,10(5):557−570. [doi: 10.1142/S0218488502001648]

[4] Wong RCW, Li J, Fu AWC, Wang K. (*a*,*k*)-Anonymity: An enhanced *k*-anonymity model for privacy-preserving data publishing. In: Proc. of the ACM 12th SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2006. 754−759.

[5] LeFevre K, DeWitt DJ, Ramakrishnan R. Mondrian multidimensional *k*-anonymity. In: Proc. of the 22nd Int'l Conf. on Data Engineering. 2006. 25−35. [doi: 10.1109/ICDE.2006.101]

[6] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. *L*-Diversity: Privacy beyond *k*-anonymity. In: Proc. of the 22nd IEEE Int'l Conf. on Data Engineering. 2006. [doi: 10.1109/ICDE.2006.1]

[7] Xiao XK, Tao YF. Anatomy: Simple and effective privacy preservation. In: Proc. of the 32nd Int'l Conf. on Very Large Data Bases. 2006. 139−150.

[8] Xiao X, Tao Y. Personalized privacy preservation. In: Proc. of the 2006 ACM SIGMOD Int'l Conf. on Management of Data. 2006. 229−240. [doi: 10.1145/1142473.1142500]

[9] Xu J, Wang W, Pei J, Wang X, Shi B, Fu AWC. Utility-Based anonymization using local recoding. In: Proc. of the 12th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2006. 785−790. [doi: 10.1145/1150402.1150504]

[10] Li N, Li T, Venkatasubramanian S. *t*-closeness: Privacy beyond *k*-anonymity and *l*-diversity. In: Proc. of the 23rd IEEE Int'l Conf. on Data Engineering. 2007. 106−115. [doi: 10.1109/ICDE.2007.367856]

[11] Wong RCW, Fu AWC, Wang K, Pei J. Minimality attack in privacy preserving data publishing. In: Proc. of the 33nd Int'l Conf. on Very Large Databases. 2007. 543−554.

[12] Tao Y, Xiao X, Li J, Zhang D. On anti-corruption privacy preserving publication. In: Proc. of the 24th Int'l Conf. on Data Engineering. 2008. 725−734. [doi: 10.1109/ICDE.2008.4497481]

[13] Backstrom L, Dwork C, Kleinberg J. Wherefore are thouR3579X?: Anonymized social networks, hiddern patterns and structural steganography. In: Proc. of the 16th Int'l Conf. on World Wide Web. 2007. 181−190.

[14] Zheleva E, Getoor L. Preserving the privacy of sensitive relationships in graph data. In: Proc. of the 1st ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD. 2007. 153−171. [doi: 10.1007/978-3-540-78478-4_9]

[15] Korolova A, Motwani R, Nabar SU, Xu Y. Link privacy in social networks. In: Proc. of the 24th Int'l Conf. on Data Engineering. 2008. 1355−1357. [doi: 10.1109/ICDE.2008.4497554]

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

[17] Campan A, Truta TM. A clustering approach for data and structural anonymity in social networks. In: Proc. of the 2nd ACM SIGKDD Workshop on Privacy, Security, and Trust in KDD. 2008. 33−54.

[18] Cormode G, Srivastava D, Yu T, Zhang Q. Anonymizing bipartite graph data using safe groupings. In: Proc. of the 34nd Int'l Conf. on Very Large Databases. ACM Press, 2008. 833−844.

[19] Hay M, Miklau G, Jensen D, Towsley D. Resisting structural identification in anonymized social networks. In: Proc. of the 34th Int'l Conf. on Very Large Databases. ACM Press, 2008. 102−114.

[20] Liu L, Wang J, Liu J, Zhang J. Privacy preserving in social networks against sensitive edge disclosure. Technical Report, CMIDA-HiPSCCS 006-08, Department of Computer Science, University of Kentucky, 2008.

[21] Ying X, Wu X. Randomizing social networks: A spectrum preserving approach. In: Proc. of the 2008 SIAM Int'l Conf. on Data Mining. 2008. 739−750.

[22] Zhou B, Pei J. Preserving privacy in social networks against neighborhood attacks. In: Proc. of the 24th IEEE Int'l Conf. on Data Engineering. 2008. 506−515. [doi: 10.1109/ICDE.2008.4497459]

[23] Bhagat S, Cormode G, Krishnamurthy B, Srivastava D. Class-Based graph anonymization for social network data. In: Proc. of the 35th Int'l Conf. on Very Large Databases. 2009. 766−777.

[24] Cheng J, Fu AWC, Liu J. *K*-Isomorphism: Privacy preserving network publication against structural attacks. In: Proc. of the 2010 ACM SIGMOD Int'l Conf. on Management of Data. 2010. 459−470. [doi: 10.1145/1807167.1807218]

[25] Wu W, Xiao Y, Wang W, He Z, Wang Z. *K*-Symmetry model for identity anonymization in social networks. In: Proc. of the 13th Int'l Conf. on Extending Database Technology. 2010. 111−122. [doi: 10.1145/1739041.1739058]

[26] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco: Morgan Kaufmann Publishers, 1988.

[27] Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. Nature, 2000,406(6794):378−382. [doi: 10.1038/35019019]

[28] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature, 1998,393(6684):440−442. [doi: 10.1038/30918]

[29] Zou L, Chen L, Ozsu MT. *K*-Automorphism: A general framework for privacy preserving network publication. In: Proc. of the 35th Int'l Conf. on Very Large Databases. 2009. 946−957.

[30] Yuan M, Chen L, Yu PS. Personalized privacy protection in social networks. In: Proc. of the 36th Int'l Conf. on Very Large Databases. 2010. 141−150.

[31] Wang Y, Xie L, Zheng B, Lee KCK. Utitily-Oritented *k*-anonymization on social networks. In: Proc. of the 16th Int'l Conf. on Database Systems for Advanced Applications. 2011. 78−92.

[32] Gao J, Xu JY, Jin R, Zhou J, Wang T, Yang D. Neighborhood-Privacy protected shortest distance computing in cloud. In: Proc. of the 2011 ACM SIGMOD Int'l Conf. on Management of Data. 2011. 409−420. [doi: 10.1145/1989323.1989367]

[33] 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. 2009. 531−540.

[34] Xu W, Zhou X, Li L. Inferring privacy information via social relations. In: Proc. of the 24th Int'l Conf. on Data Engineering Workshop. 2008. 525−530. [doi: 10.1109/ICDEW.2008.4498373]

[35] He J, Chu WW, Liu Z. Inferring privacy information from social networks. In: Proc. of the Intelligence and Security Informatics. 2006. 154−165. [doi: 10.1007/11760146_14]

[36] Das S, Egecioglu O, Abbadi AE. Anonymizing weighted social network graphs. In: Proc. of the 26th Int'l Conf. on Data Engineering. 2010. 904−907.

[37] Liu X, Yang X. A generalization based approach for anonymizing weighted social network graphs. In: Proc. of the 12th Int'l Conf. on Web-Age Information Management. 2011. 118−130.

[38] Yuan M, Chen L. Node protection in weighted social networks. In: Proc. of the 16th Int'l Conf. on Database Systems for Advanced Applications. 2011. 123−137. [doi: 10.1007/978-3-642-20149-3_11]

[39] Li Y, Shen H. Anonymizing graphs against weight-based attacks. In: Proc. of the 2010 IEEE Int'l Conf. on Data Mining Workshops. 2010. 491−498. [doi: 10.1109/ICDMW.2010.112]

[40] Liu X, Yang X. Protecting sensitive relationships against inference attacks in social networks. In: Proc. of the 17th Int'l Conf. on Database Systems for Advanced Applications. 2012. 335−350. [doi: 10.1007/978-3-642-29038-1_25]

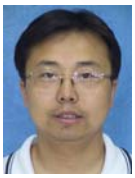
[41] Tai CH, Yu PS, Yang DN, Chen MS. Privacy-Preserving social network publication against friendship attacks. In: Proc. of the 17th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2011. 1262−1270. [doi: 10.1145/2020408.2020599]

[42] Bhagat S, Cormode G, Krishnamurthy B, Srivastava D. Prediction promotes privacy in dynamic social networks. In: Proc. of the 3rd Conf. on Online Social Networks. 2010. 6−6.

[43] Karwa V, Sofya R, Smith A, Yaroslavtsev G. Private analysis of graph structure. In: Proc. of the 37th Int'l Conf. on Very Large Databases. 2011. 1147−1157.

[44] Chen S, Zhou S. Recursive mechanism: Towards node differential privacy and unrestricted joins. In: Proc. of the 2013 Int'l Conf. on Management of Data. 2013. 653−664.

刘向宇(1981－),男,辽宁铁岭人,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: neulxy@gmail.com



王斌(1972－),男,博士,副教授,CCF 会员,主要研究领域为数据查询处理,数据质量管理.
E-mail: binwang@mail.neu.edu.cn

杨晓春(1973－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为数据库理论与技术.
E-mail: yangxc@mail.neu.edu.cn