软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(8):1922−1933 [doi: 10.13328/j.cnki.jos.005093]

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

时空数据发布中的隐式隐私保护*

王 璐, 孟小峰, 郭胜娜

(中国人民大学 信息学院,北京 100872)

通讯作者: 孟小峰, E-mail: xfmeng@ruc.edu.cn

摘 要: 随着大数据时代的到来,大量的用户位置信息被隐式地收集.虽然这些隐式收集到的时空数据在疾病传播、路线推荐等科学、社会领域中发挥了重要的作用,但它们与用户主动发布的时空数据相互参照引起了大数据时代时空数据发布中新的个人隐私泄露问题.现有的位置隐私保护机制由于没有考虑隐式收集的时空数据与用户主动发布的位置数据可以相互参照的事实,不能有效保护用户的隐私.首次定义并研究了隐式收集的时空数据中的隐私保护问题,提出了基于发现-消除的隐私保护框架.特别地,提出了基于前缀过滤的嵌套循环算法用于发现隐式收集的时空数据中可能泄露用户隐私的记录,并提出基于频繁移动对象的假数据添加方法消除这些记录.此外,还分别提出了更高效的反先验算法和基于图的假数据添加算法.最后,在若干真实数据集上对提出的算法进行了充分实验,证实了这些算法有较高的保护效果和性能.

关键词: 隐式隐私;时空数据;隐私保护

中图法分类号: TP311

中文引用格式: 王璐,孟小峰,郭胜娜.时空数据发布中的隐式隐私保护.软件学报,2016,27(8):1922−1933. http://www.jos.org.cn/1000-9825/5093.htm

英文引用格式: Wang L, Meng XF, Guo SN. Preservation of implicit privacy in spatio-temporal data publication. Ruan Jian Xue Bao/Journal of Software, 2016,27(8):1922−1933 (in Chinese). http://www.jos.org.cn/1000-9825/5093.htm

Preservation of Implicit Privacy in Spatio-Temporal Data Publication

WANG Lu, MENG Xiao-Feng, GUO Sheng-Na

(School of Information, Renmin University of China, Beijing 100872, China)

Abstract: In the emerging big data era, in addition to explicit publication of users' locations on geo-social networks, positioning system embedded in mobile phones implicitly records users' locations. Although such implicitly collected spatiotemporal data play an important role in a wide range of applications such as disease outbreak control and route recommendation for life science or smart city, they cause new serious privacy issues when cross-referencing with the explicitly published data from users. Existing location privacy preservation techniques fail to preserve the proposed implicit privacy because they ignore the cross-reference between implicitly and explicitly spatiotemporal data. To tackle this issue, this work for the first time investigates and defines the implicit privacy and proposes the discover and eliminate framework. In particular, this paper proposes prefix filtering based nest loop algorithm and frequent moving object based algorithm to generate dummy data to preserve the proposed implicit privacy. Further, it constructs an improved reverse a priori algorithm and graph based dummy data generation algorithm respectively to make the solution more practical. The results of extensive experiments on real world datasets demonstrate the effectiveness and efficiency of the proposed methods.

Key words: implicit privacy; spatio-temporal data; privacy preserving

* 基金项目: 国家自然科学基金(61532010, 61379050, 91224008); 国家高技术研究发展计划(863)(2013AA013204); 高等学校博士学科点专项科研基金(20130004130001); 中国人民大学科学研究基金(11XNL010)

Foundation item: National Natural Science Foundation of China (61532010, 61379050, 91224008); National High-Tech R&D Program of China (863) (2013AA013204); Research Fund for the Doctoral Program of Higher Education of China (20130004130001); Research Funds of Renmin University of China (11XNL010)

收稿时间: 2015-12-19; 修改时间: 2016-04-14; 采用时间: 2016-06-02

大数据时代,随着定位技术的发展,基于位置的服务越来越普及,用户的时空数据通过各种各样的服务发布出来.在用户通过签到等移动社交网络服务主动发布自己的时空行为的同时,大量记录人们行为的时空数据在人们使用手机进行通话、接发短信息时被移动通信商隐式收集[1].由于手机与移动通信商间的时空数据由手机的信号基站自动收集,这些隐式收集的数据具有数据量大、蕴含人类行为的特征,并且在疾病传播[2,3]、贫困消除[4]、城市规划[5]等重大科学社会问题以及路线推荐[6]、乘车出行[7]等重要生活应用中发挥了关键作用.然而,这些隐式收集到的时空数据通过与用户主动发布的时空数据相互参照,会暴露用户有关个人身份、行动目的、健康状况、兴趣爱好等多方面的敏感隐私信息[8,9].近年来,随着个人隐私观念的增强以及数据发布使用中法律法规的健全,这些隐式收集到的时空数据在供科学研究与数据挖掘前需要首先消除其中可能暴露用户隐私的记录.

为了确保这些个人敏感信息不被泄露,大量针对时空数据隐私保护的工作致力于匿名化可能暴露个人敏感信息的时空数据.例如,位置与轨迹数据上的 k 匿名等方法将用户位于特定时间范围和空间区域的位置记录泛化为相同的空间区域,从而让攻击者不能识别出某个时间范围与空间区域中的特定用户[2].但是,这些方法没有考虑到攻击者可以参照用户主动发布的时空数据从隐式收集到的时空数据中找到暴露用户隐私的记录,因此,经过这些方法保护的时空数据集依然会泄露用户隐私的数据.

例如,假设攻击者看到 Alice 在社交网络上主动广播了自己 5 月 1 日去海南、10 月 1 日去哈尔滨的两次旅行,如果手机运营商收集到的时空数据集中在这两天分别去过海南和哈尔滨的记录都具有相同的用户 ID,那么攻击者就可以推断这条记录中的用户 ID 在现实中对应用户 Alice,从而根据数据库中同样 ID 的时空记录发现 Alice 去过的其他地方,并进而推断出她的兴趣爱好、行为习惯等隐私信息.

最新的研究[10]表明,现有的位置或轨迹隐私保护方法即使将待发布位置泛化为 15 平方公里,时间戳泛化为一个小时,95%以上的用户也可以被 4 个泛化后的时空记录唯一地确定.如果这些用户还通过签到服务主动发布了自己的若干行为,他们就很容易被攻击者识别出来.

在大数据时代,由于攻击者可以通过越来越多的位置社交网站同时收集用户主动发布的时空数据,隐式收集到的时空数据集更容易暴露用户隐私,这对保护这些时空记录在效果、效率和效用这 3 个方面提出了严重挑战:(1) 对时空数据集的保护效果如何不受攻击者收集用户主动发布的时空数据的能力的增强而减弱;(2) 由于数据集中可能暴露用户隐私的记录个数随着时空数据集的增大呈指数级增长,如何高效发现待保护的记录;(3) 为了发布时空数据而对时空点集进行保护时,如何保证较高的数据效用.

为了应对上述挑战,本文首先在隐式收集到的数据集上定义与用户主动发布数据无关的隐式隐私,以保证无论攻击者收集到多少用户主动发布的数据,经过隐私保护后的隐式收集到的时空数据集都不会泄露额外的信息.随后,本文提出发现和消除两个步骤来保护隐式收集到的时空数据.为了高效寻找可能暴露用户隐私的记录,本文提出基于前缀过滤的嵌套循环算法,并针对其可扩展性差的缺陷提出高效的反先验算法.为了避免这些记录暴露用户的隐私,我们设计了基于频繁移动对象的假数据添加算法,并针对其数据扭曲程度高的缺陷提出了基于图的假数据添加算法,以确保经过保护后的数据集具有高数据效用.

(1) 首次提出并定义时空数据中的隐式隐私,并提出保护该隐私的框架.我们首次定义了隐式收集到的时空数据中的(ε,k)隐私问题,并提出了发现-消除框架和相应算法确保经过保护的数据集不因为攻击者收集到更多用户主动发布的时空数据而泄露额外信息.

(2) 高效的发现算法.我们针对框架中的发现步骤提出了基于前缀过滤的嵌套循环的算法,并针对其扩展性差的缺陷设计了基于反先验算法的优化算法,高效地寻找暴露隐式隐私的时空点集合.

(3) 高效的保护算法.我们针对框架中的消除步骤提出了基于频繁移动对象的假数据添加算法,并针对其数据效用差的缺陷设计了基于图的假数据添加方法,大幅提高了数据效用.

(4) 真实数据集上的充分实验.我们使用两个真实数据集验证了我们方法的效果和效率.

本文第 1 节主要介绍时空数据发布中隐私保护的相关技术.第 2 节介绍隐式隐私问题及其发现-消除保护框架.第 3 节介绍相应的发现,消除算法.第 4 节是实验效果展示.