论文编号：2015010008


貴州大學

# 2019届博士学位论文

# 理性隐私保护模型及应用

学科专业：应用数学

研究方向：密码学与数据安全

中国·贵州·贵阳

2019年 10月

# 目　录

# 摘　　要

TBC

**关键词：** 隐私保护，博弈论，隐私量化，隐私推测，基于风险访问控制

# Abstract

TBC

**Keywords:** Privacy preserving, Game Theory, Privacy quantification, Privacy inference, Risk adaptable based access control

# 第一章　绪论

## 1.1　研究背景及意义

互联网、移动互联网和物联网快速发展，以及5G技术的不断推进和商用推广，社交网络、位置服务、医疗健康、生物基因、工业控制等海量数据被主动或被动采集、传输、存储、流转、分析并应用。海量数据的产生和应用推动了云计算、大数据和边缘计算等新兴产业和技术的爆发式增长，并产生了智慧医疗、智慧交通、智慧政府、智慧城市等不同的应用，极大地丰富了人们的物质和精神生活。同样，数据海量化增长、网络跨域泛在、计算云端化、应用多样复杂化等新的变化为安全和隐私带来了巨大挑战，大量的病毒、漏洞、攻击和数据关联分析，致使隐私严重泄漏，引发了人们极大的担忧。表1.1展示了近年来主要的隐私泄露事件，充分表明了隐私泄露已经成为网络空间的重要威胁。在此背景下，深入的理解隐私并保护隐私变得尤为重要。

表 1.1: 近年来主要隐私泄露事件简况

| 时间 | 事件 | 影响 | 原因 |
|---|---|---|---|
| 2017年7月 | 韩国加密货币交易所客户数据泄露 | 3万个人用户数据被盗并遭受电话诈骗 | 黑客入侵攻击 |
| 2017年10月 | 全球11个国家41个凯悦酒店数据泄露 | 数据量不详，涵盖信用卡姓名、卡号、到期日期、验证码等 | 通过恶意软件进行黑客入侵 |
| 2017年10月 | 马来西亚超过总人口的手机用户信息泄露 | 4620万人用户地址、身份证号、手机识别卡信息泄露 | 不详 |
| 2017年10月 | 埃森哲服务器大量敏感信息泄露 | 19亿敏感的密码和解密密钥泄露 | 操作失误将数据放到未保护的云服务上 |
| 2017年10月 | 南非史上最大规模数据泄露 | 3160万人个人资料被公之于众 | 数据在未保护的服务器上导致黑客窃取 |
| 2018年3月 | Facebook用户数据泄露 | 5千万用户数据泄露，影响美国大选 | 越权采集并分析用户喜好、性格、行为特点、政治倾向 |
| 2018年8月 | 华住集团数据泄露 | 5亿条、140G华住旗下酒店的用户数据泄露 | 不详 |
| 2018年8月 | 谷歌采集设备、地图、搜索位置信息 | 全球超20亿用户数据被越权采集 | 谷歌公司故意采集 |

由于90%以上的数据被提供公共服务的政府、社会组织和企业所采集、存储，为了使数据发挥更大的价值，往往需要对包含大量隐私信息的数据进行共享、开放、交

换和分析处理；同时很多信息服务也是基于个人隐私信息与服务质量的交换，如网站注册服务、公共WIFI接入、云存储、智能手机导航、信息搜索与广告推送、在线信用卡支付、RFID应用等。这些场景中由于法律法规要求和个人意愿，需要对隐私信息进行保护，同时服务提供方、数据利用方或恶意第三方希望获取更多的隐私敏感信息，以提供更好的服务、获取更大数据价值，得到更好的数据效用，两个目标同时存在且相互冲突，需要均衡解决。

关于隐私的研究，自2006年 $k$ 匿名模型[1]被提出以后逐步变成系统化的研究，隐私研究发展为基于密码学的方案[2-3]和基于非密码学的方案[1,4-7]两大类，这些方案被大规模应用于以数据为中心的开放、复杂、跨域场景中，如云存储、社交网络、基于位置服务、物联网、边缘计算、数据挖掘、机器学习、医疗健康等。众多应用场景中，隐私保护目标和数据利用目标天然矛盾，如何平衡二者的关系是核心问题之一。在这两类隐私研究中，基于密码学的方案通常利用可证明安全理论定义密码学意义上的隐私保护目标，设计对应的密码学方案，如同态加密、可搜索加密、属性密码方案等实现隐私保护目标[2-3]；基于非密码学的方案主要是定义了匿名性设计达到匿名化效果的算法来实现用户的身份匿名隐私保护[1,4-5]，通过定义邻近数据集的查询结果不可区分性，设计加噪的方法达到这种不可区分性来实现属性值的隐私保护[6]，通过定义数据动态隐私，设计自适应的风险的细粒度访问控制实现隐私数据不被非授权用户访问[7]。其中，基于密码学的方案具有严格的理论方法支撑，能够达到预期的隐私保护目标，但是这些隐私定义是密码学意义上安全性定义，隐私保护方案设计也依赖公钥密码，其计算高度复杂导致效率低下，且难以采用折中的措施实现隐私保护效果和数据效用的平衡；基于非密码学的方案通过概率或信息论定义匿名性和不可区分性意义上的隐私，并设计泛化匿名或加噪的方式实现匿名或属性值隐私保护，效率高且有利于平衡隐私保护效果和数据效用。目前，以数据为中心的开放应用场景多样化，特别是数据开放共享应用中，大规模的个人隐私需要在保证数据可用的前提下得到实用性的隐私保护，研究基于非密码学的方案可以达到这一目标，平衡隐私保护与数据效用，具有重要的现实意义。

隐私领域的研究主要有三方面科学问题。**第一、隐私定义与度量**。如何恰当形式化的定义隐私、并对隐私进行量化。特别是隐私量化，既包括对特定数据集中隐私量的量化，又包括在某种隐私分析攻击模型下，个人隐私潜在泄露量、隐私分析攻击后隐私泄露量评估，还包括某一隐私保护模型对数据集隐私保护能力的量化。**第二、隐私分析与推测**。在某一场景下针对保护后的隐私信息数据集进行隐私分析与推测，如何最大程度的获取更多隐私信息。**第三、隐私保护**。如何对某一场景下的隐私数据集进行有效隐私保护，如何在保护隐私的同时平衡隐私保护效果和数据效用。深入研究科学问题一和科学问题二有助于对隐私的理解和认识，能够对隐私泄露的机理进行深入剖析，能够对设计更好的隐私保护方案提供科学理论依据和评价方法，研究科学问

题三能够实现对数据隐私的预期性保护，如可量化的、动态性的、自适应的隐私保护，能够平衡隐私保护效果与数据效用间的关系。上述三个科学问题对基于非密码学的方案研究有重要的理论意义，能够有助于该领域完善其基础理论支撑，可在保证其实用性基础上提高隐私定义形式化及度量、隐私泄露机理、隐私保护方案的科学性。

本文主要针对数据开放共享场景下的基于非密码学隐私研究领域，展开隐私定义与度量、隐私分析与推测及隐私保护研究，旨在提出能够动态、自适应地对包含大量隐私信息的数据集进行隐私保护，并实现隐私保护与数据效用间的平衡。

## 1.2　研究现状

本节围绕本文的研究内容，就相关研究领域的现状进行梳理和分析，包括隐私度量、隐私分析与推测以及隐私保护三个方面，以更加深入的理解本文研究的背景。

### 1.2.1　隐私度量

早期对隐私的认知是法理上的“隐私权”，在技术上被定义为匿名性，即在一个匿名集中元素不能被唯一标识的状态。在匿名通信系统中，匿名性最初被量化为匿名数据集阶的自然对数 $A = log_2(N)$[8]，未考虑敌手获取的概率信息量。直到2002年，匿名性的量化引入了信息论[9]刻画敌手对匿名系统或匿名集的去匿名化攻击后获得的信息量，Serjantov 和Danezis[9]将匿名性定义为 $d = H(X)$，其中$H(X)$是攻击者对匿名集合中的元素进行去匿名分析后的信息熵；随后，利用正规熵[10]、相对熵[11]有不同的匿名性度量方法被提出，这些匿名性度量方法都未考虑敌手的背景知识，无法动态量化匿名性。2007年，Edman等[12]利用二分图邻接矩阵和信息熵对传输信息者和接收信息者的信息进行映射，量化匿名性的信息量。这些匿名性的量化仅针对匿名通信系统(如洋葱路由系统，Tor系统和Crowds系统)，更多此类方法见2009年Edman和Yener的综述[13]，但这些方法并不适用数据共享和应用中的匿名性度量。针对数据共享应用的隐私量化最早聚焦在数据库领域，2002年，Sweeney[1]将数据集中某一记录的匿名性量化为$d = 1/k$，其中 $k$ 是数据集中与该记录不可区分的记录数量；随后，该方法被扩展为 $l$ 多样性匿名[4]和 $t$ 邻近匿名[5]。针对数据集的匿名性定义被扩展到了基于位置服务[14]、社交网络[15]等应用场景，并用以不同形式的数据发布[16-17]。这些方法都是将匿名性量化为某一概率值，并不能对敌手去匿名化攻击获取的信息量进行量化，且无法根据敌手的背景知识进行动态量化。

林欣等[18]对位置$k$匿名无法在连续查询攻击下刻画匿名集中位置的匿名度，提出了匿名集查询结果信息熵的匿名度量化方法$AD(q) = 2^{H(q)}$，具有更好的适用性.

### 1.2.2　隐私攻击与推测

对基于位置服务中用户的位置信息进行直接 $k$ 匿名保护的情况，林欣等提出了一

种连续查询攻击[18]，在不同 $k$ 匿名保护算法下的位置查询中成功区别出位置发送者。

### 1.2.3　隐私保护算法

鉴于基于密码学的隐私研究并非本文研究的聚焦点，尽管该领域亦有很多成果，本文也不再进行详述，可查阅基于属性密码[19]、可搜索加密[20-21]、同态密码[22]、安全多方计算[23-24]等领域的综述进一步了解。

## 1.3　有待解决的关键问题

## 1.4　本文工作

## 1.5　论文结构

# 第二章　基础知识

## 2.1　Shannon信息论及其扩展

### 2.1.1　熵

### 2.1.2　互信息

### 2.1.3　结构信息论

## 2.2　博弈论

### 2.2.1　博弈模型

### 2.2.2　策略博弈

### 2.2.3　扩展博弈

### 2.2.4　演化博弈

## 2.3　隐私定义及隐私保护

### 2.3.1　身份隐私

### 2.3.2　属性隐私

### 2.3.3　隐私保护模型

# 第三章　基于信息通信模型的隐私度量模型

## 3.1　引言

隐私保护的研究起步较早,但近年来突然受到产业界和学术界的广泛关注是因为大数据的不期而至.坦率地说,大数据的迅速发展让学术界始料未及,大数据的理论研究已经落后于产业需求,尤其是隐私保护成为大数据应用的主要瓶颈,移动网络、社交网络、基于位置服务等新型应用服务的推进,隐私问题更加突出.目前关于隐私保护有两个方向值得关注：一是研究隐私保护算法以更加有效的方式保护隐私;二是通过研究隐私泄露风险分析与评估,解决数据的可用性与隐私保护之间的的平衡.隐私保护算法目前主要集中在匿名方法,包括 $k$ 匿名、 $l$ 多样性匿名和 $t$ 接近匿名及其衍生的方法.隐私度量最早起源于相关匿名算法[1], 在匿名隐私保护算法的研究过程中,不时有学者关注隐私量化问题,尤其是在定位服务领域,位置匿名及轨迹匿名算法上已有不少隐私度量的相关研究[2,3],因此对于隐私保护算法来说,隐私度量仍需进一步深入研究.然而就目前来说,隐私泄露涉及因素众多,设计有效的隐私保护算法仍然是挑战性问题,但政府及企业数据开放共享中迫切的隐私保护需求,促使我们不得不在可用性与隐私泄露之间寻求一种平衡,要解决这个问题,隐私风险分析及评估不失为一种方法.风险分析依然涉及到隐私量化问题,也就是说量化风险评估不失为隐私保护一种可行的解决方案,量化隐私风险必然也涉及隐私度量问题.从这些分析来看,隐私度量的研究具有十分重要的理论意义和应用价值.

信息熵作为信息度量的有效工具,在通信领域已展现出其重要的贡献[4].隐私作为一种信息,自然可以考虑用熵来量化,为此,不少学者或多或少进行了探索,比如事件熵、匿名集合熵、条件熵等[5-7],但其研究还较为零散,更多是针对某一具体领域,如位置隐私保护领域,目前尚未形成统一的模型及体系,其应用范围也受到限制,特别是隐私是具有时空性的,与人的主观感受也有关系,不同的人对同一隐私的认同可能不同. 鉴于以上分析,本文旨在参考Shannon信息论的通信框架[8],提出几种隐私保护信息熵模型,包括隐私保护基本信息熵模型、含敌手攻击的隐私保护信息熵模型、带主观感受的信息熵模型和多隐私信源的隐私保护信息熵模型.在这些模型中,将信息拥有者假设为发送方,隐私谋取者假设为接收方,隐私的泄露渠道假设为通信信道;基于这样的假设,分别引入信息熵、平均互信息量、条件熵及条件互信息等来分别描述隐私保护系统信息源的隐私度量、隐私泄露度量、含背景知识的隐私度量及泄露度量;以此为基础,进一步提出了隐私保护方法的强度和敌手攻击能力的量化测评,力图为隐私泄露的量化风险评估提供一种理论支持.

# 第四章　基于结构信息论的隐私度量模型

# 第五章　相互独立的序列型数据的隐私属性推测模型及其应用

# 第六章　相互关联的序列型数据的隐私属性推测模型及其应用

# 第七章　面向隐私保护的风险自适应访问控制模型

# 第八章　理性的隐私风险访问控制模型及其分析

### 8.0.1　引言

### 8.0.2　基于风险访问控制模型

### 8.0.3　基于二人博弈的理性风险访问控制模型

### 8.0.4　基于演化博弈的理性风险访问控制模型

To be completed.

# 第九章　总结及展望

## 9.1　结论

## 9.2　展望

# 参考文献


[1] SWEENEY L. k-anonymity: A model for protecting privacy[J]. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5):557-570.

[2] NABEEL M, BERTINO E. Privacy preserving delegated access control in public clouds[J]. IEEE Trans. Knowl. Data Eng., 2014, 26(9):2268-2280.

[3] 黄刘生, 田苗苗, 黄河. 大数据隐私保护密码技术研究综述[J]. 软件学报, 2015, 26 (4):945-959.

[4] MACHANAVAJJHALA A, KIFER D, GEHRKE J, et al. *L*-diversity: Privacy beyond *k*-anonymity[J]. TKDD, 2007, 1(1):3.

[5] LI N, LI T, VENKATASUBRAMANIAN S. t-closeness: Privacy beyond k-anonymity and l-diversity[C]//ICDE. [S.l.]: IEEE Computer Society, 2007: 106-115.

[6] DWORK C. Differential privacy[C]//Lecture Notes in Computer Science: volume 4052 ICALP (2). [S.l.]: Springer, 2006: 1-12.

[7] ZHANG W, LI H, ZHANG M, et al. Privacy-aware risk-adaptive access control in health information systems using topic models[C]//SACMAT. [S.l.]: ACM, 2018: 61-67.

[8] REITER M K, RUBIN A D. Crowds: Anonymity for web transactions[J]. ACM Trans. Inf. Syst. Secur., 1998, 1(1):66-92.

[9] SERJANTOV A, DANEZIS G. Towards an information theoretic metric for anonymity[C]//Lecture Notes in Computer Science: volume 2482 Privacy Enhancing Technologies. [S.l.]: Springer, 2002: 41-53.

[10] DÍAZ C, SEYS S, CLAESSENS J, et al. Towards measuring anonymity[C]//Lecture Notes in Computer Science: volume 2482 Privacy Enhancing Technologies. [S.l.]: Springer, 2002: 54-68.

[11] DENG Y, PANG J, WU P. Measuring anonymity with relative entropy[C]//Lecture Notes in Computer Science: volume 4691 Formal Aspects in Security and Trust. [S.l.]: Springer, 2006: 65-79.

[12] EDMAN M, SIVRIKAYA F, YENER B. A combinatorial approach to measuring anonymity[C]//ISI. [S.l.]: IEEE, 2007: 356-363.

[13] EDMAN M, YENER B. On anonymity in an electronic society: A survey of anonymous communication systems[J]. ACM Comput. Surv., 2009, 42(1):5:1-5:35.

[14] NIU B, LI Q, ZHU X, et al. Achieving k-anonymity in privacy-aware location-based services[C]//INFOCOM. [S.l.]: IEEE, 2014: 754-762.

[15] CAMPAN A, TRUTA T M. Data and structural k-anonymity in social networks [C]//Lecture Notes in Computer Science: volume 5456 PinKDD. [S.l.]: Springer, 2008: 33-54.

[16] WONG R C, LI J, FU A W, et al. ($\alpha$, $k$)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing[C]//KDD. [S.l.]: ACM, 2006: 754-759.

[17] YING X, PAN K, WU X, et al. Comparisons of randomization and k-degree anonymization schemes for privacy preserving social network publishing[C]// SNAKDD. [S.l.]: ACM, 2009: 10.

[18] 林欣, 李善平, 杨朝晖. LBS中连续查询攻击算法及匿名性度量[J]. 软件学报, 2009, 20(4):1058-1068.

[19] EDEMACU K, PARK H K, JANG B, et al. Privacy provision in collaborative ehealth with attribute-based encryption: Survey, challenges and future directions[J]. IEEE Access, 2019, 7:89614-89636.

[20] BÖSCH C, HARTEL P H, JONKER W, et al. A survey of provably secure searchable encryption[J]. ACM Comput. Surv., 2014, 47(2):18:1-18:51.

[21] POH G S, CHIN J, YAU W, et al. Searchable symmetric encryption: Designs and challenges[J]. ACM Comput. Surv., 2017, 50(3):40:1-40:37.

[22] ACAR A, AKSU H, ULUAGAC A S, et al. A survey on homomorphic encryption schemes: Theory and implementation[J]. ACM Comput. Surv., 2018, 51(4):79:1-79:35.

[23] CRAMER R, DAMGÅRD I, NIELSEN J B. Secure multiparty computation and secret sharing[M]. [S.l.]: Cambridge University Press, 2015.

[24] DUGAN T M, ZOU X. A survey of secure multiparty computation protocols for privacy preserving genetic tests[C]//CHASE. [S.l.]: IEEE Computer Society, 2016: 173-182.