理性隐私保护模型及应用（长摘要）

数学与统计学院 应用数学专业 丁红发

指导教师：向淑文 教授

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。近年来主要的各类重大隐私泄露事件，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。

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和基于非密码学的方案两大类，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。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；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；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。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强度的量化。**第二、隐私分析与推断**。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断，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、动态性的及自适应的隐私保护，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体系，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。

面对上述隐私领域的主要科学问题挑战，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，展开隐私度量、隐私分析、隐私保护，以及隐私保护与数据效用平衡方面研究，旨在能够深入探究隐私基础理论，提高对隐私泄露及隐私保护机理的理解，以提出能够动态、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。

本文以国家自然科学基金项目《理性隐私计算及隐私风险可控技术研究》为支撑，主要聚焦在以信息论通信模型及其扩展工具研究隐私度量的基础性框架模型，能够对隐私定义、隐私分析攻击模型和隐私保护机制进行量化；以概率推断为工具建立序列型隐私数据的属性隐私分析敌手模型，并针对真实数据进行分析推断攻击，量化敌手隐私分析攻击强度；因风险访问控制模型为基础，定义并量化风险隐私，设计动态自适应访问控制模型；以博弈论为工具，刻画访问控制隐私保护机制参与者的隐私和数据效用需求的理性行为和有限理性行为，设计能动态平衡隐私保护和数据效用关系的理性风险访问控制隐私保护机制。研究体系如图1所示，具体取得了如下6方面的成果。



图1 论文研究体系及结构安排

1. **基于信息熵的隐私通信模型及度量方法**

利用信息论的相关工具，如熵、互信息等来对匿名隐私、成员关系隐私和属性隐私进行形式化定义和度量的研究较多，但是大部分是集中在位置匿名、轨迹匿名、数据集匿名、数据集成员属性、训练集关系隐私、社交网络匿名和属性等方面的研究。在隐私量化方面，缺乏对隐私定义、隐私分析、隐私保护等统一的量化方法。

本文基于Shannon信息论的通信模型框架提出了几种隐私保护信息通信模型，对不含敌手的隐私保护、含敌手的隐私保护、多隐私保护源的隐私保护等不同情境提出了相应的模型进行建模，以满足对隐私度量、隐私保护机制效果度量和敌手隐私分析强度度量等需求。在所提出的度量模型中，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；基于该假设，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，形成了以信息论为核心的隐私度量方法体系；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强度的量化，为隐私泄露的量化提供了一种支撑，对整个隐私保护过程中的保护机制、敌手能力都提供了量化方法。

在构建的隐私保护通信模型中，将含敌手的隐私保护机制构建了如图2所示的通信模型，并定义了信源熵、攻击条件熵、互信息量等来量化隐私。

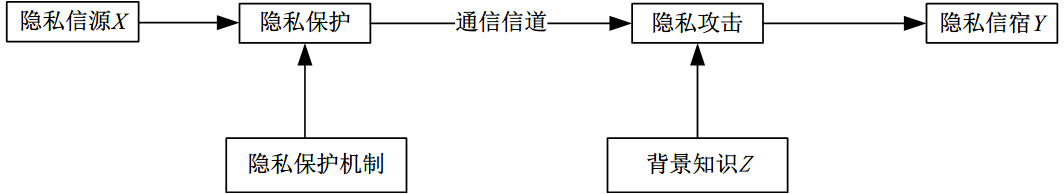
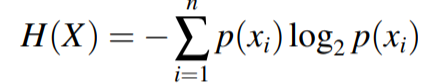


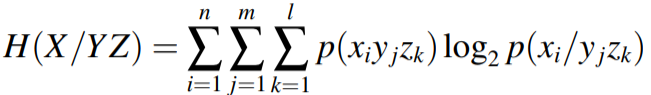
图2 含敌手背景知识攻击的单隐私信源隐私保护通信模型

针对该模型，隐私信源熵*H*(*X*)为



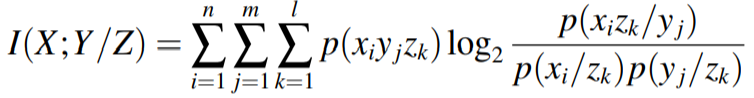
*H*(*X*)用于刻画隐私信源的平均隐私信息量，也是隐私信源的隐私不确定程度，*H*(*X*)越大，隐私泄露就可能越小，从而它亦可以用于衡量隐私的保护程度，在没有外部条件影响时，该值是一个确定的值。

隐私攻击条件熵*H*(*X/YZ*)为



*H*(*X/YZ*)反映了攻击者在获得隐私信宿消息*Y*和背景知识*Z*后，关于*X* 仍存在的不确定度，它实际了可以作为在具有攻击分析的情况下隐私信息的不确定度。

隐私攻击平均互信息 *I*(*X*;*Y*/*Z*)为



*I*(*X*;*Y*/*Z*)反映了得到 *Z* 的条件下， *X* 和 *Y* 之间的平均互信息量，即接收方获得的 隐私信息量，即可以刻画具有背景知识攻击下的隐私泄露程度。

1. **独立序列型数据属性隐私推断模型**

近年来，由于数据种类繁多、数量庞大且应用需求多样化，越来越多的数据被以集中或分布式的形式共享、开放，造成了大量的隐私泄露，这些泄露又成为敌手进行隐私分析的背景知识，增加了数据共享的隐私泄露风险，对数据隐私泄露的潜在威胁量化，对数据隐私保护机制设计都提出了高的要求。特别是需要对隐私泄露的原理进行进一步研究，以帮助更好地度量隐私、理解隐私泄露机理，并设计更好的隐私保护方法。目前针对匿名方法的去匿名性分析研究较多，针对社交网络的用户偏好、个人信息等属性隐私的分析研究较多，但是对新型序列化数据的属性隐私，如时间序列的位置隐私、基因序列的基因座敏感值隐私较少，此类数据在很多共享应用场景（如疾病诊断、车联网导航）中需要非匿名化，需要对其敏感的属性隐私（特定基因座的基因型，特定行车位置）进行保护。

本文针对基因序列数据的属性隐私提出了一种基于概率推断的隐私分析模型。该模型通过对单条敏感数据记录属性值存在的相互关联关系进行分析，构建目标属性值推断的敌手模型。在提出的敌手模型基础上，分别提出了两种不同的基因序列属性隐私分析方法。第一种主要基于Monte Carlo-Markov抽样和隐Markov推断算法，建立了目标基因序列的“抽样解析”——“单倍体属性值概率推断”——“二倍体合成”三个步骤的属性隐私推断模型；第二种方法应用卷积神经网络构建概率推断算法，改进了单倍体属性值推断过程，实现了大规模序列型数据的属性推断目标。所提出的方法针对不存在亲属关系的群体基因序列数据共享场景，在本文提出的隐私度量模型基础上，定义了序列型数据属性隐私和量化方法，并应用于分析属性隐私泄露情况，通过量化隐私泄露量和敌手获取隐私量等信息，提高对序列型数据属性隐私的认识和理解。实验表明，本文提出的方法比现有基因序列属性隐私分析模型和算法更优，敌手对属性隐私的错误率、不确定度降低，敌手获得隐私信息量都比已有的工作更优。

1. **关联序列型数据属性隐私推断模型**

随着不同机构和个人更加容易获取基因组数据，且这些敏感数据被广泛地应用于医疗、保险、寻亲及社交等场景，对数据安全和隐私的担忧也在不断加剧。为了证实在序列型数据属性隐私方面，存在个人共享基因数据也会大量泄漏他人属性隐私的问题，为了进一步分析家族成员基因序列数据共享会造成他人基因序列属性隐私泄露的机理，需要对相互关联的基因序列型数据进行隐私分析。

本文利用因子图和置信传播算法针对亲属间的基因序列属性隐私建立分析推断敌手模型和分析算法。该模型考虑了单核苷酸多态性间高阶相关性，利用公开DNA参照数据集和全基因组关联研究(GWAS)目录数据，提高了推断攻击模型的属性隐私分析强度。该模型的敌手隐私分析强度通过本文所提出的隐私度量框架，对基因序列属性隐私进行了定义，并将隐私损失量作为评价指标进行了隐私分析强度量化。实验结果表明，所提出的攻击更适合于高密度基因组数据隐私推断，且具有较少的错误率、不确定性和更多隐私损失，显著提高了属性隐私的隐私分析推断能力。

1. **隐私保护风险自适应访问控制模型**

在以数据为中心的开放系统中，数据往往通过云服务或其他集中式的方式按需提供数据共享、开放、应用服务，这些需求多样复杂产生了复杂的隐私泄露风险和威胁，需要动态化、细粒度、适应性的方案对数据提供访问控制模式的隐私保护。但目前基于传统的强制访问控制、基于角色访问控制以及新型的基于属性访问控制，都不能很好的解决该问题。

本文针对云环境中共享、应用涉及隐私或敏感信息数据的场景研究面向隐私保护的访问控制模型。在XACML上扩展提出了一种基于风险的自适应访问控制模型，以动态化地在访问控制过程中保护数据隐私，约束隐私侵犯行为，激励诚实访问行为。首先，根据风险访问控制场景的隐私保护需求提出了面向隐私保护的风险访问控制敌手模型；其次，该模型在标准的XACML框架进行了扩展，新增了策略风险评估、会话控制和风险消减服务三个组件，增强了策略执行、策略访问和策略信息组件。在新增的组件中，以Shannon信息熵作为工具，在提出的隐私度量模型基础上，提出了基于风险的隐私定义和量化方法，对用户的访问控制请求风险和用户自身的风险类型结合，提出了访问请求类型判别方法；通过风险隐私量化及基于信用卡模型的激励机制，实现访问行为风险阈值的动态调整，考虑了用户短期访问行为和长期访问行为的影响。对比和分析表明，所提出的模型和方法较现有的工作更加动态化，且实现了隐私保护，易用性更好。

1. **基于扩展式博弈的理性隐私风险访问控制模型**

基于风险访问控制模型可很好的解决以数据为中心的开放系统中自适应数据隐私保护。但强制访问控制、基于角色访问控制、基于属性访问控制和已有的基于风险访问控制等模型，在平衡隐私保护需求与数据效用需求冲突方面，仍存在问题，特别是过度授权导致隐私泄露或授权不足导致数据可用性不足的问题需要进一步解决。此外，还需要对基于风险访问模型中对隐私保护的能力和方法进一步提升。

针对上述需求和问题，本文在所提出的风险自适应访问控制模型的基础上，进一步运用Shannon自信息和博弈论，提出了基于风险适应性的理性访问控制模型以实现数据共享场景中的保护隐私和数据应用需求间的平衡。在定义了隐私风险和隐私侵犯访问的概念之后，提出了基于博弈论的风险访问控制模型框架和工作流程。此外，还进一步利用Shannon信息的定义提出了量化访问请求隐私风险和用户隐私风险值的计算公式，强化了访问控制请求对数据隐私的刻画；以提出的理性风险访问控制模型、访问请求隐私 风险和用户隐私风险为基础，提出了多轮二人博弈来刻画面向隐私保护的风险访问控制中访问者与数据服务提供者的“隐私保护-数据服务”冲突与合作关系，进一步提出并分析了博弈效用函数及其二人博弈过程。分析表明，在基于隐私风险访问控制的每一轮博弈中都存在子博弈精炼纳什均衡，可以通过限制侵犯隐私的访问请求来保护隐私，实现隐私保护与数据访问效用间的平衡。分析和对比表明，该方法比已有的工作更有优势，需要更少的辅助信息，提供更多的风险适应性和隐私保护强度。

1. **基于演化博弈的理性隐私风险访问控制模型**

社交网络、医疗信息系统等以数据为中心的大规模用户(访问者)开放信息系统，亟需能够保护隐私的细粒度自适应访问控制模型，且需实现数据隐私保护需求和数据效用需求的平衡。现有基于理性的访问控制模型难以满足适应性保护隐私的需求，且博弈参与者的完全理性假设太强，不符合实际场景。基于风险访问控制能够实现细粒度的访问控制隐私保护目标，但如何进一步放松参与者完全理性的假设，并实现隐私保护与数据效用关系的动态平衡，仍需要进一步研究。

本文针对这些问题和需求，在提出的风险自适应访问控制模型和完全理性隐私风险访问控制模型的基础上，进一步提出一种面向隐私保护的有限理性风险自适应访问控制模型。新提出的模型包含了新的隐私风险量化模块和演化博隐私弈决策模块。该模型首先基于信息量对访问请求的数据集隐私信息量进行量化，构造了访问请求隐私风险函数和用户隐私风险函数；其次，基于演化博弈在有限理性假设下构建多参与者的访问控制演化博弈模型，利用复制动态方程分析了访问控制参与者的动态策略选择和演化稳定状态形成机理，提出了隐私风险访问控制博弈演化稳定策略的选取方法。仿真实验和对比表明，所提出的访问控制模型能够有效动态自适应地保护敏感信息资源系统中的隐私信息，具有更好的隐私风险适应性，有限理性参与者的动态演化访问策略选取更加符合实际场景。