数据隐私保护体系



* **数据资源格式划分**

数据资源按格式可以分为结构化、半结构化、非结构化数据。

结构化数据，即行数据,存储在数据库里,可以用二维表结构来逻辑表达实现的数据。

非结构化数据，包括所有格式的办公文档、文本、图片、XML、HTML、各类报表、图像和音频/视频信息等等

半结构化数据是介于完全结构化数据（如关系型数据库、面向对象数据库中的数据）和完全无结构的数据（如声音、图像文件等）之间的数据，HTML文档就属于半结构化数据。它一般是自描述的，数据的结构和内容混在一起，没有明显的区分。

而具体的数据数据格式又可以划分为整型、字符串和对于身份证、电话号码、电子邮箱等具有特定特征的数据。

* **数据敏感程度分级**

数据按内容敏感程度可以分为公开、半公开和敏感数据，而敏感数据在国家、企业和个人等不同层面又有不同的体现形式。在个人层面，如身份证、手机号、邮政编码等信息都属于隐私数据；对于相同的敏感数据，个人和团体对其可能又有不同的敏感等级划分。

* **数据发布阶段隐私保护**

在隐私数据的发布阶段，需要对敏感数据进行识别和脱敏，如通过自然语言处理技术对文本类的敏感数据进行识别，通过匿名算法等来对要发布的数据进行脱敏；传统的隐私发布技术主要针对结构化的记录表进行脱敏处理，针对云平台中存储的数据信息，研究识别其中的记录表及相应的文本中的敏感信息的方法，同时针对不同的敏感信息，研究遵循不同匿名原则的脱敏算法，分别对识别的记录表及文本类敏感信息进行脱敏处理。

* **数据存储阶段隐私保护**

在数据的存储阶段，数据存储方一般为云存储平台，大数据的存储者和拥有者是分离的，云存储服务商并不能保证是完全可信的。用户的数据面临着被不可信的第三方偷窃数据或者篡改数据的风险。加密方法是解决该问题的传统思路，但是，由于大数据的查询、统计、分析和计算等操作也需要在云端进行，为传统加密技术带来了新的挑战。同态加密技术、混合加密技术、基于BLS短签名POR模型、DPDP、Knox等方法，是针对数据存储时防止隐私泄露而采取的一些方法。

针对云存储场景的大量分析和统计发现存在大量冗余数据的具体情景，数据去重技术被广泛应用进来检测并识别云存储系统中的共有重复数据以节省存储成本和网络带宽。

然而基于数据去重的云存储系统面临着数据隐私保护风险及其解决方法带来的性能问题，例如数据去重加密方法的机密性与时间开销，文件上传和文件访问中的隐私保护与计算开销，以及共享数据的可用性保障与存储开销等问题。因此，如何构建安全且高效的基于数据去重的云存储系统是目前亟待解决的问题。

* **数据传输阶段隐私保护**

从数据传输路径的角度来看，对传输过程的安全保护主要涉及传输消息的源发送者身份、发送者位置、发送者节点标识、接受者位置、接受者身份等隐私信息。此外对于消息数据的访问及使用也应当通过访问控制、认证权限管理等方法进行相应的保护。一般的消息内容保护以及访问认证保护，通过传统的成熟的密码学认证方法都可以实现。以密码学和通信认证协议为基础的对称、非对称加密方法经过多方实用验证和研究，能够对用户数据隐私提供坚实的保障，为用户数据隐私和通信加密提供了有利的技术支持。

而对于云平台上的多租户数据分享传输的问题，云服务上仅仅存储用户的密文数据，用户查看密文数据需要通过访问代理的解密。实际应用场景下，身处不同访问代理之间的用户存在分享密文的需求；用户自身也可能由于工作原因从某一访问代理迁移到另一访问代理，用户希望访问原访问代理加密的数据。 因此访问代理之间密文数据的分享是基于访问代理架构的密文搜索亟待解决的新问题。解决这一问题的关键是解决访问代理之间密钥的分享。

* **特定情境下的数据隐私保护**

与特定情境相关的数据，比如移动终端产生的位置数据、轨迹数据等也存在着隐私泄露的风险。随着智能终端的普及和无线通信技术的发展，基于位置的服务已渗入到人们的日常生活当中。这些服务在给人们的日常生活带来便利的同时，也带来隐私泄漏的风险。

针对轨迹数据的推理攻击不仅可分析出目标用户的家庭住址、工作地点等敏感位置信息，甚至可推测出用户的生活习惯、健康状态、宗教信仰等隐私信息。轨迹隐私能否得到妥善保护已成为制约移动互联网发展的瓶颈问题。

根据用户的查询请求在到达 LBS 服务器之前变换方式的不同， 当前的轨迹隐私保护方法可分为以下 3 类，分别是: 基于 k 匿名泛化的方法、基于噪声数据的方法、基于动态假名的方法。

其中文献《移动社交网络的位置隐私保护方法》针对用户的敏感兴趣点泄露问题，提出了一种情景感知的隐私保护方法。该方法将位置信息、社交关系、个人信息引入到知识构建算法中以计算兴趣点间的相关性，并利用该相关性及时空情景实时判断发布当前位置是否会泄露用户隐私，进而实现了隐私保护与服务可用性间的平衡。最后通过仿真实验验证了该方法的有效性。

文献《F-Seeker: 基于重匿名的粒度化好友搜索架构》针对社交网络中好友检索服务的隐私保护问题，本文提出一种基于重匿名技术的粒度化好友搜索架构F-Seeker。对用户发布的位置信息采用增强的 k 匿名策略—( k，m，e) -匿名，用以防止“好奇”的搜索服务提供方对用户隐私的推测。在处理好友搜索服务过程中，服务提供方根据粒度化的可视策略对数据实施重匿名，实现了对用户位置信息粒度化的访问控制。

* **隐私泄露后的处理**

除了上述从不同角度来减少和预防隐私泄露风险以外，还应该考虑隐私泄漏后的溯源问题，分析个人数据的可溯源性及溯源路径。很多知识产权保护技术可以应用于个人数据产品溯源标识体系。目前，有多种先进技术用于知识产权保护，如加密技术、认证技术、数字水印、电子签名等，可在这些技术基础上开发一套溯源技术体系。

从隐私对抗的角度来看，模型反演攻击可以利用去噪自编码网络还原训练数据集中的原始信息，利 用 生 成 式 对 抗 网 络 ( generative adversarial networks， GAN) 可以生成与训练数据集相近的数据。