数据安全治理关键技术进展

王伟平
中国科学院信息工程研究所

目录

CONTENT

- 数据安全治理的主要任务
- 数据安全治理若干关键技术
- 总结与展望

大数据是一场技术革命，一场经济变革，也是一场国家治理的变革

2015年 8月

“数据已成为国家重要的基础性战略资源”

国务院印发《促进大数据发展行动纲要》

2015年 10月

首次提出“国家大数据战略”

中共十八届五中全会

2016年 9月

推动政务信息系统互联和公共数据共享

国务院印发
《政务信息资源共享管理暂行办法》

2016年 12月

对十三五时期大数据产业发展进行整体规划、全面部署

工信部印发
《大数据产业发展规划2016-2020年》

2017年 12月

推动实施国家大数据战略

习近平总书记在中央政治局第二次集体学习上的讲话

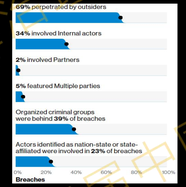
大数据时代，数据安全尤为重要

数据安全关系到国家安全

例：2017年哈罗德·托马斯·马丁三世窃取美国政府信息系统中超过50TB数据，包括至少5亿页政府机密文件

数据泄露日趋严重

71%的数据泄露以金钱为目的，25%涉及战略目的的间谍活动；69%的网络攻击来自外部人员，34%涉及内部人员的窃取



数据安全法规和标准逐渐完善

2017：《网络安全法》实施；
2019：《等保2.0》实施；
2020：《密码法》实施；
2020：《数据安全法》《个人信息保护法》征求意见

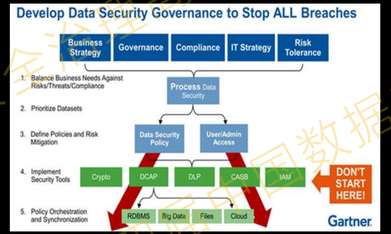
数据安全需要有系统化思维和建设框架

我国大数据安全建设仍存在安全防护有技术、少体系，安全监管缺手段、少抓手，数据管理有规划、缺落实等问题，数据安全治理将数据安全技术与数据安全管理融合，对于提升数据安全，促进数据共享利用具有重要意义

什么是数据安全治理？

▪Gartner数据安全治理理念

数据安全治理不仅是一套用工具组合的产品级解决方案，而是从决策层到技术层，从管理制度到工具支撑，自上而下贯穿整个组织架构的完整链条。



▪微软的DGPC理念

隐私、保密和合规性框架的数据治理计划，围绕人员、流程和技术三个核心能力领域组织。基于信息生命周期，从安全基础架构、身份和访问控制、信息保护、审计和报告四个技术领域提供保证，并遵循数据隐私和保密原则。

▪中国网信联盟数据安全治理委员会

“让数据使用更安全”为目的的安全体系构建的方法论。

《数据安全治理白皮2.0》

数据安全治理技术体系

•信息安全技术
数据安全能力成熟度模型

•数据生命周期安全过程域

数据生存周期安全过程域

数据采集安全	数据传输安全	数据存储安全	数据处理安全	数据交换安全	数据销毁安全
•PA04数据质量管理 •PA03 数据源鉴别及记录 •PA02数据采集安全管理 •PA01 数据分类分级	•PA05数据传输加密 •PA06网络可用性管理	•PA09数据备份和恢复 •PA08逻辑存储安全 •PA07存储媒体安全	•PA14数据导入导出安全 •PA13数据处理环境安全 •PA12数据正当使用 •PA11 数据分析安全 •PA10数据脱敏	•PA16数据发布安全 •PA15数据共享安全 •PA17 数据接口安全	•PA19存储媒体销毁处置 •PA18数据销毁处置

通用安全过程域

•PA20数据安全策略规划 •PA26终销数据安全	•PA21组织和人员管理 •PA27监控与审计	•PA22合规管理 •PA28鉴别与访问控制	•PA23数据资产管理 •PA29需求分析	•PA24数据供应链安全 •PA30安全事件应急	•PA25元数据管理

•数据安全治理技术支撑体系

数据安全治理技术支撑体系

- **安全状况摸底**
  - 数据资产梳理与定位
  - 静态梳理（数据定位）
  - 动态梳理(数据使用)
  - 权限梳理(授权状况)
  - 数据分级分类
  - 合规 漏洞 配置 权限
  - 数据库安全风险扫描
- **数据安全管控**
  - 业务访问安全
  - 运维访问安全
  - 测试开发安全
  - 数据外发安全
  - 数据存储安全
- **数据治理稽核**
  - 行为审计与分析
  - 权限变化监控
  - 异常行为分析
  - 建立安全基线

数据共享的主要问题

利用大数据打造企业核心竞争力，提升政府治理能力和公众服务水平，已成为社会的共识。数据是智慧社会的支撑，及时开放、共享数据，充分挖掘数据资源"宝矿"，才能发挥最大效益。

国际上"开放政府联盟"宣布成立，八国集团首脑签署了《开放数据宪章》。在此期间，美国、英国、法国、加拿大和澳大利亚等西方发达国家和巴西、阿根廷等新兴国家纷纷推出数据开放的国家战略。

2015年国务院印发的《促进大数据发展行动纲要》中首次提出"大力推动政府部门数据共享"以来，各级政府部门以及医疗、交通等重点行业积极响应。尽管如此，大数据共享的发展仍然不容乐观。

数据共享的痛点

用户担心个人隐私暴露，不愿意共享数据

隐私原因

行为、视频、照片 → 客户端A

云服务器

行为、视频、照片 → 客户端B

政府、企业等实体担心数据共享引发敏感信息泄露，违反法律、法规

政策原因

General Data Protection Regulation

《欧盟通用数据保护条例》

传统企业数据

敏感信息不可交换

政府数据 ✖ 互联网企业数据

孤立的数据难以形成合力，阻碍数据资源的充分利用

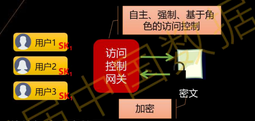
基于属性密码的安全数据共享

1、定义

➢属性加密方案（Attribute Based Encryption，ABE）是一种新型公钥加密技术，能够在加密的同时实现对数据的细粒度访问控制。

➢不再需要部署集中式的访问控制网关，是未来实现密态数据共享的重要工具之一。

传统解决方案



传统解决方案的不足：

- 一对一
- 访问控制网关容易成为重点攻击目标
- 密钥管理复杂
- 很难实现细粒度的访问控制
- 访问控制策略不够灵活

VS

属性加密方案

属性密码的优势：

- 一对多
- 在加密的同时实现访问控制，不再需要访问控制网关
- 密钥管理简单
- 实现细粒度的访问控制，可精确到文件
- 灵活访问控制策略，支持与、或、优先级

2、应用场景——密态数据的访问控制

➢同时提供加密和细粒度访问控制功能

➢访问控制策略由传统集中式设置变为分布式设置，适用于分布式场景

➢适配结构化数据（如数据库页表）、非结构化数据（如文档、图片等）

3、国内外技术进展

➢从2005年至今，属性密码备受学术界和工业界关注，一直是热门研究方向。

➢国内外研究侧重纯理论、特定场景解决方案研究，研究成果之间不成体系。

➢目前属性加密实用化研究较为分散，互不兼容，距离实用化还有一定距离！

基于属性密码的安全数据共享

围绕属性加密方案的实用化技术，中科院信工所团队在效率优化技术、核心功能增强、安全防护机制、跨平台软件实现、应用落地解决方案等领域形成一系列创新研究成果：

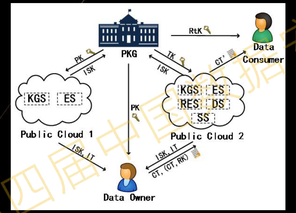
➢在IEEE TDSC、IEEE TIFS、IEEE TPDS、IEEE TM、IEEE TSC、ESORICS等期刊和会议上发表学术论文20余篇。

➢形成跨平台软件函数库，可兼容主流操作系统和开发环境。

基于属性密码的安全数据共享

• 支持快速权限代理的加密数据共享方案

主要问题：属性加密方案的核心算法和权限代理的模指数、双线性对运算次数与属性个数呈正线性相关。例如，一次加密操作可达403次模指数运算



技术创新：

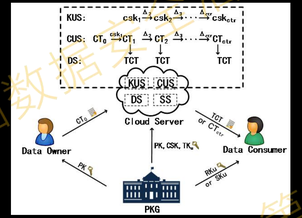
✓提出了基于外包计算的支持快速权限代理的属性加密方案，成功将所有核心算法和权限代理算法的在线计算量从O(n)到0

✓解决了基于属性加密的密态数据共享技术的效率瓶颈

✓该成果发表在CCF网络与信息安全领期刊IEEE TDSC

• 支持用户即时撤销机制的属性加密方案

主要问题：实际部署的密码系统离不开用户撤销和用户密钥撤销机制**既存的属性密码撤销机制无法兼容效率和可靠性**



技术创新：

✓提出了基于分步解密思想的可撤销属性加密方案，可立即对用户完成撤销操作，同时抵抗撤销密钥泄漏攻击

✓解决了现有属性加密撤销机制无法兼容效率和可靠性的问题，解决实际部署的后顾之忧

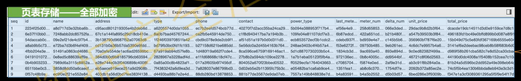
✓该成果发表在期刊IEEE TSC

基于属性密码的安全数据共享

文件加密和访问控制

✓实现不同医院医疗数据的互联互通共享，便利百姓就医， 延伸医疗大数据产业链条
✓支持端到端加密，确保数据中心无法获得明文数据，打通共享通道

数据库字段级加密和访问控制



✓数据库表的所有字段均使用ABE加密。
✓不同用户使用各自密钥解密同一张页表，得到不同的字段。

可搜索加密技术

企业敏感数据外包存储的安全性与可用性同样重要

- 2000 Song等提出SSE机制
- 2013 Cash等提出支持布尔查询的OXT机制
- 2013 Jarecki等提出支持查询授权的MSSE机制
- 2014 Cash等提出分块存储索引
- 2016 Sun等结合ABE实现支持查询授权的MSSE机制
- 2020 Du等提出支持关键词粒度动态查询授权是DMSSE机制

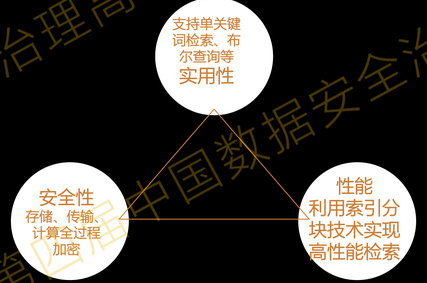
存在的问题：

1、检索与检索结果的访问控制粒度不统一，检索授权粒度过粗或过细，缺乏实用性

2、索引数据膨胀率高，且依赖于内存存储，难以满足大规模密态数据高效查询检索需求

可搜索加密技术

- 融合属性加密与可搜索加密的分块数据存储组织技术，实现数据加密存储、密态数据搜索、查询快速响应



- **安全性**：存储、传输、计算全过程加密，支持适配商密、国密等算法
- **可用性**：支持单关键词检索、布尔检索；提供表、列级细粒度访问控制策略
- **高性能**：采用索引分块存储方法，亿级记录规模的密态数据，单关键词检索响应时间小于1秒；索引大小仅为数据的45%

共享学习技术

- 共享学习技术是一种在多方参与且互不信任场景下，聚合多方数据并保护数据隐私和安全，实现数据共享的分布式学习范式
- 共享学习技术在工业界有着广泛的应用前景，有望成为下一代人工智能协同算法和协作网络的基础

安全合规

满足安全及合规要求
数据不泄露
保证数据和模型安全

联合建模

打破数据孤岛
联合多方建模
聚合数据价值
连接不同场景

提升赋能

合作推动行业
建立激励机制
互利共赢
赋能行业AI

共享学习技术的基本概念

共享学习技术的国内外研究现状

共享学习技术

功能	隐私学习	联邦学习	竞合学习	可信机器学习	分裂学习	共享学习
降维						√
差分隐私	√	√			√	√
可信计算环境	√		√	√	√	√
安全多方计算	√	√	√		√	√
同态加密	√	√	√			√
数据/模型受控	√ / X			X / √		√ / √

有效性

非独立同分布的数据
连接状态不稳定的网络
计算能力不一致的节点

安全性

重构攻击
模型反演攻击
成员推理攻击

共享学习技术面临的问题和挑战

共享学习作为未来人工智能发展的底层技术，它依靠安全可信的数据保护措施下连接数据孤岛的模式，将不断推动全球人工智能技术的创新与飞跃。随着共享学习在更大范围和更多行业场景的渗透及应用，它在更高层面上对各类人群、组织、行业和社会都将产生巨大影响。

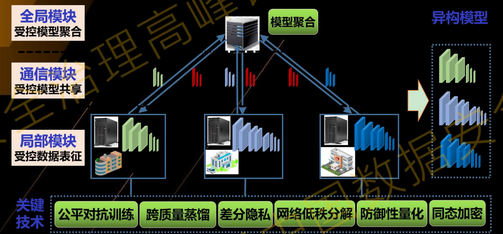
共享学习的社会价值
- 加速人工智能技术创新发展
- 保障隐私信息及数据安全
- 促进全社会智能化水平提升

共享学习的商业价值
- 带动跨领域的企业级数据合作
- 催生基于联合建模的新业态和模式
- 降低技术提升成本和促进创新技术发展

共享学习的应用场景
- 金融风控
- 视觉安防
- 自动驾驶
- 智慧医疗
- 智慧零售

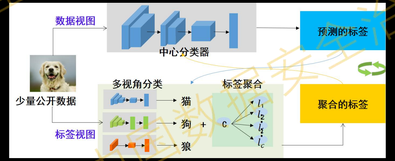
共享学习技术的重大价值

共享学习技术



全过程受控的局部-全局协调算法

关键技术：基于耦合视图的数据学习
问题难点：跨组织多源数据共享，异构鸿沟难填
技术思路：聚合多视角弱标签，耦合视图联合训练



相关工作发表在AAAI 2020上
Coupled-view Deep Classifier Learning from Multiple Noisy Annotators

关键技术：基于知识蒸馏的模型共享
问题难点：模型保障高效知识传递的同时面临隐私泄露
技术思路：教师-学生训练，不直接访问数据下模型传承

相关工作发表在期刊IEEE TIP和会议AAAI 2020上
Efficient Low-Resolution Face Recognition via Bridge Distillation
Look One and More: Distilling Hybrid Order Relational Knowledge for Cross-Resolution Image Recognition

共享学习若干技术进展

赋能广泛的数据隐私与算法安全服务

总结与展望

- 大数据时代，数据安全性与数据可用性同等重要。属性加密、共享学习等技术适用于很多数据共享应用场景，具有很好的应用前景；
- 相关技术在理论研究方面取得了丰硕的成果，相对而言还没有广泛的应用；
- 希望产学研用各方能够加深合作， 共同提升数据安全治理的技术支撑能力。

THANKS