Lecture 2:数据合规与安全

一、安全多方计算

姚氏百万富翁问题：For definiteness, suppose Alice has i millions and Bob has j millions, where $1 \leq i, j \leq 10$. We need a protocol for them to decide whether $i < j$, such that this is also the only thing they know in the end(aside from their own values).Let M be the set of all N.

通俗理解的方法：第一个百万富翁，选择十个盒子，按照顺序排列，分别代表 1 到 10，并用自己的财产数字与盒子的数字进行比较，如果小于该数字，则在盒子里面放一个苹果，若大于等于则放一个香蕉。请第二个百万富翁来，让他选择自己财产数额对应的箱子(在第一个百万富翁不参与的情况下)，随后，第二个百万富翁把剩余所有箱子销毁。最后，两个百万富翁一起打开最后剩下的那个箱子，则可以得出谁更富有。

(1)Bob picks a random N-bit integer, and computes privately the value of $E_a(x)$;call the result k.(B无私钥，有公钥。选择一个大数x，加密$E(x) = k$)

(2)Bob sends Alice the number $k - j + 1$.(B计算$k - j + 1$，并告诉A)

(3)Alice computes privately the values of $y_u = D_a(k - j + u)$ for $u = 1, 2, ..., 10$.(A有公钥有私钥，计算$k - j + 1, k - j + 2, ..., k - j + j, ..., k - j + 10$)，并分别解密
$y_1, y_2, ..., y_j, ..., y_{\{10\}}, y_j = D(k - j + j) = D(k) = x$

(4)Alice generates a random prime of $\frac{N}{2}$ bits, and computes the values $z_u = y_u (\mathrm{mod}\, p)$ for all u; if all z_u differ by at least 2 in the mod p sense, stop; otherwise generares another random prime and repeat the process until all z_u differ by at least 2; let p, z_u denote this final set of numbers;(A选择质数p，对y取模，得到$z_1, z_2, z_3, z_4, ..., z_j, ..., z_{\{10\}}, z_j = x(\mathrm{mod}\, p)$)

(5)Alice sends the prime p and the following 10 numbers to B: $z_1, z_2, ..., z_i$ followed by $z_i + 1, z_{i+1} + 1, ..., z_{10} + 1$; the above numbers should be interpreted in the mod p sense.(A把z_i后的都+1，并把p和 10 个数都发送给B)

(6)Bob looks at the j-th number(not counting p) sent from Alice, and decides that $i \geq j$ if it is equal to $x \bmod p$, and $i < j$ otherwise.(B拿出第j个，有可能是z_j，也有可能是$z_j + 1$，计算$x \bmod p$。如果与拿出的相等，说明$j \leq i$，否则$j > i$)

(7)Bob tells Alice what the conclusion is.

第一步，经过特定的操作，让A构造出n把锁，B有且仅有第j把锁的钥匙，但是A不知道j是多少；

第二步，A给B n把锁锁着的标志位，其中前i个标志位置 0，后$n - i$个置 1；

第三步，B检查第j把锁锁着的标志位是否为 0.如果为 0 则$i \geq j$，否则$i < j$。

二、联邦学习

联邦学习：在分布式设备或系统上训练模型，参与方仅传输模型参数，在不共享数据的基础上联合建模。原始数据不出域、通信量低、计算负载均衡，但是上传参数存在隐私泄露风险。

对于联邦学习，有一个比较著名的比喻是“小羊吃草”，小羊和草分别被比作模型和数据。对于传统的机器学习，需要将各个草场的草移动到小羊所在的中心区域。

三、数据脱敏

Data Msking 是根据指定的脱敏规则，针对敏感信息进行数据变形或遮蔽，降低数据的敏感级别，扩大数据可共享和被使用的范围，达到保护隐私数据安全的目的。

分类：

- 动态数据脱敏： 适用于不脱离生产环境，对敏感数据的查询和调用结果进行实时脱敏。在访问敏感数据的同时实时进行脱敏处理，可以为不同角色、不同权限、不同数据类型执行不同的脱敏方案，从而确保返回的数据可用而安全。
- 静态数据脱敏：适用于脱离生产环境，脱敏后发生至测试、开发、数据分析等场景。是数据的“搬移并仿真替换”。将数据脱敏处理后，下发给下游环节取用和读写。脱敏后数据与生产环境相隔离，满足业务需求的同时保障生产数据库的安全。

数据脱敏技术浅谈

脱敏方法：为了适应不同数据脱敏的应用场景，在保持数据原始特征及业务一致性的基础上，提供了多种数据脱敏方法

- 无效化：方案在处理待脱敏的数据时，通过对字段数据值进行 截断、加密、隐藏 等方式让敏感数据脱敏，使其不再具有利用价值。一般采用特殊字符代替真值
- 随机值：随机值替换，字母变为随机字母，数字变为随机数字，文字随机替换文字的方式来改变敏感数据，这种方案的优点在于可以在一定程度上保留原有数据的格式，往往这种方法用户不易察觉的。
- 数据替换：数据替换与无效化方式比较相似，不同的是这里不以特殊字符进行遮挡，而是用一个设定的虚拟值替换真值。
- 对称加密：对称加密是一种特殊的可逆脱敏方法，通过加密密钥和算法对敏感数据进行加密，密文格式与原始数据在逻辑规则上一致，通过密钥解密可以恢复原始数据，要注意的就是密钥的安全性。
- 平均值：平均值方案经常用在统计场景，针对数值型数据，我们先计算它们的均值，然后使脱敏后的值在均值附近随机分布，从而保持数据的总和不变。
- 偏移和取整：偏移和取整通过随机移位改变数字数据，偏移取整在保持了数据的安全性的同时保证了范围的大致真实性，比之前几种方案更接近真实数据，在大数据分析场景中意义比较大。

数据脱敏的整体架构

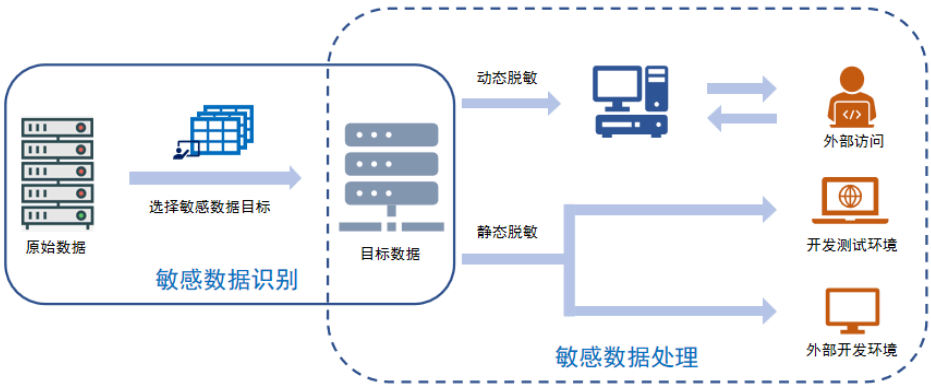

图 1　整体架构

脱敏标准：为了适应不同数据脱敏的应用场景，基于多种数据脱敏方法，我们提供多个数据脱敏算法以满足不同数据脱敏标准

- 遮蔽脱敏：对数据的全部或者一部分用符号替换；
- 一致性脱敏：原始数据的关联关系在进行数据脱敏后也能保持；

- 保持数据格式脱敏：保留数据的主要格式；
- 保持数据特征脱敏：保留数据的主要特征；
- 泛化脱敏：保留原始数据局部特征的前提下使用其他方式替代原始数据的方式；
- 可逆性脱敏：脱敏后数据可以使用对应表，对数据进行恢复操作，从脱敏数据可以获取原始数据。

四、差分隐私

什么是差分隐私？“隐私”的数学模型：第一次用可证明的数学模型定义了隐私和隐私保护（Dwork06）

怎么实现差分隐私实现：通过对真实数据添加随机噪声进行扰动，实现用户隐私的量化保护

安全性：随机噪声对真实数据的扰动是差分隐私安全性的来源。

可用性：（独立）随机噪声叠加后的相互抵消使得扰动数据的统计结果具有较高准确度。

差分隐私： For every pair of inputs that differ in one row. Adversary should not be able to distinguish between any D_1 and D_2 based on any O.

$$\log\left(\frac{\Pr[A(D_1)=O]}{\Pr[A(D_2)=O]}\right)<\varepsilon(\varepsilon>0)$$

.Simulate the presence or absence of a single record.Guarantee holds no matter what the other records are. Privacy Parameters ε

$$\Pr[A(D_1)=O]\leq e^{\varepsilon}\Pr[A(D_2)=O]$$

Controls the degree to which D_1 and D_2 can be distinguished. Smaller the ε more the privacy(and better the utility)

Laplace 机制

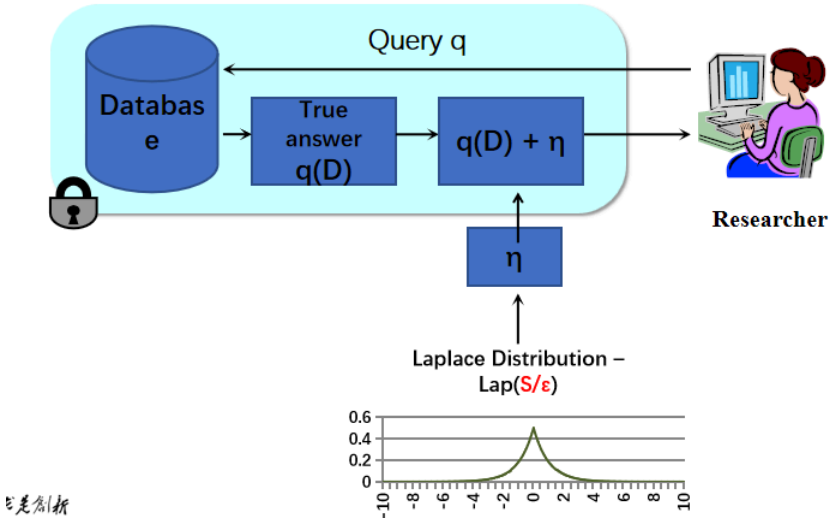

图 2　Laplace 机制

$$f(x|\mu,b)=\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

Mean μ, variance $2b^2$, Lap(b) when $\mu=0$

Sensitivity: Consider a query $q:I\rightarrow R.S(q)$ is the smallest number s.t. for any beighboring tables D,D',

$$q(D) - q(D') \leq S(q)$$

Theorem: If **sensitivity** of the query is **S**, then the algorithm $A(D) = q(D) + \text{Lap}\left(\frac{S(q)}{\varepsilon}\right)$ guarantees ε-differential privacy

Consider neighboring databases D and D'. Consider some output O,

$$\frac{\Pr[A(D) = O]}{\Pr[A(D') = O]} = \frac{\Pr[q(D) + \eta = O]}{\Pr[q(D') + \eta = O]} = \left(\frac{e^{-\frac{|O - q(D)|}{\lambda}}}{e^{-\frac{|O - q(D')|}{\lambda}}} \right) \leq e^{|q(D) - q(D')\frac{1}{\lambda}} \leq e^{\frac{S(q)}{\lambda}} = e^{\varepsilon}$$

差分隐私的组合性

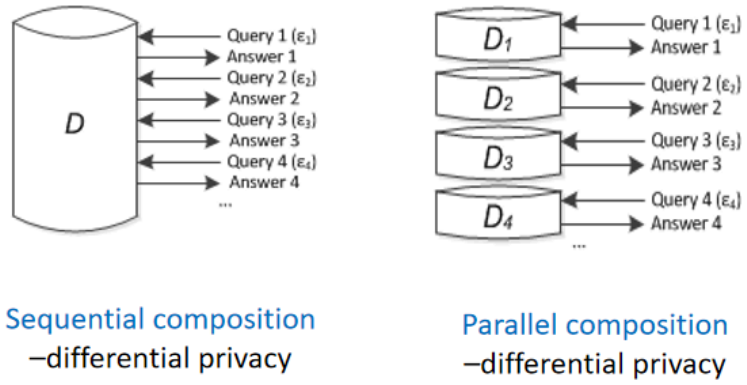

图 3　组合性

为什么是差分隐私？

- 功能需求：加密方法解决数据计算过程中的隐私泄露，差分隐私解决计算结果的隐私泄露。
- 性能需求：在海量数据收集的场景中，加密方法计算开销过大。
- 风控需求：企业需要一种“一劳永逸”的用户数据脱敏方法，即在数据收集后，存储、处理、分析阶段不需要额外的用户隐私保护措施，从而有效控制风险责任。

图 4　差分隐私

五、全同态加密

全同态加密：在密文域的操作等效于明文域的对应操作，支持加法和乘法数据操作的同时不泄露任何数据信息.

优势：实现了在密文域上的任意运算，避免运算过程中需要先解密而导致的用户敏感信息泄露，实现了数据的可用不可见

劣势：计算成本高

六、零知识证明

零知识证明：证明者（Prover）能够在不向验证者（Verifier）提供任何有用的信息的情况下，使验证者相信某个论断是正确的

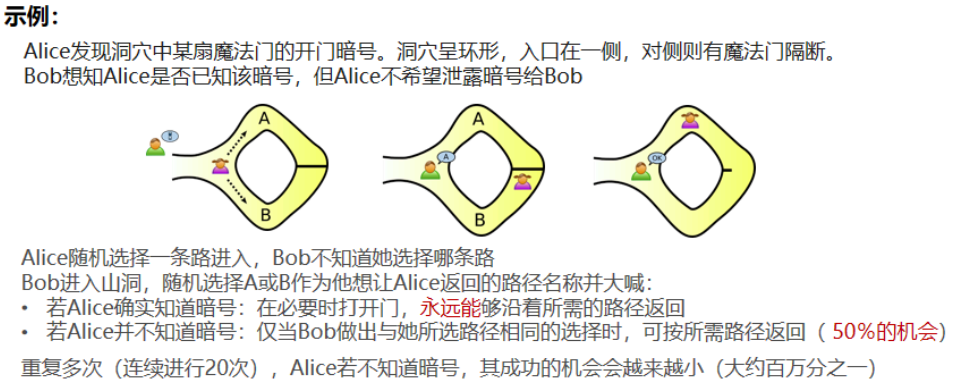

图 5　示例

零知识证明特征：

- 完整性：诚实的验证者可以相信诚实证明者拥有正确论断
- 可靠性：不诚实证明者无法说服诚实验证者其拥有正确论断
- 零知识性：验证者不知道除论断的正确性以外的任何信息

零知识证明技术：

- 交互式零知识证明：通过证明者和验证者之间的多轮交互完成论断的证明
- 简洁非交互式零知识证明（zk-SNARKs）：
  - 简洁：零知识证明可以快速验证
  - 非交互式：证明者和验证者之间只需要交互一次
- 非交互式可扩容透明零知识证明（zk-STARKs）：
  - 可扩容：验证时间与电路规模呈亚线性增长
  - 透明：依赖于可公开验证的随机数来生成用于证明和验证的公共参数

优势：

- 严格的隐私保护，通过密码学可证明安全理论实现对用户信息的隐私保护
- 广泛的应用场景，包括匿名支付、身份认证、区块链扩容等，未来市场潜力大

劣势：

- 证明生产成本高：证明者的时间和空间负责度较高，无法在低算力设备上执行
- 需要不可证伪假设：大多数零知识证明技术需要不可证伪假设，例如知识假设
- 存在量子计算威胁：部分零知识证明使用双线性曲线群，量子计算机可以打破其安全模型

实用化挑战

- 性能优化：优化执行引擎，改进算法和电路，实现线性证明者执行时间
- 硬件加速：设计新型基于硬件（GPU、ASIC）加速的安全零知识证明方案
- 零知识证明编译器：开发高兼容性零知识证明编译器，包括高级开发框架、布尔电路、R1CS 编译器

七、数据合规