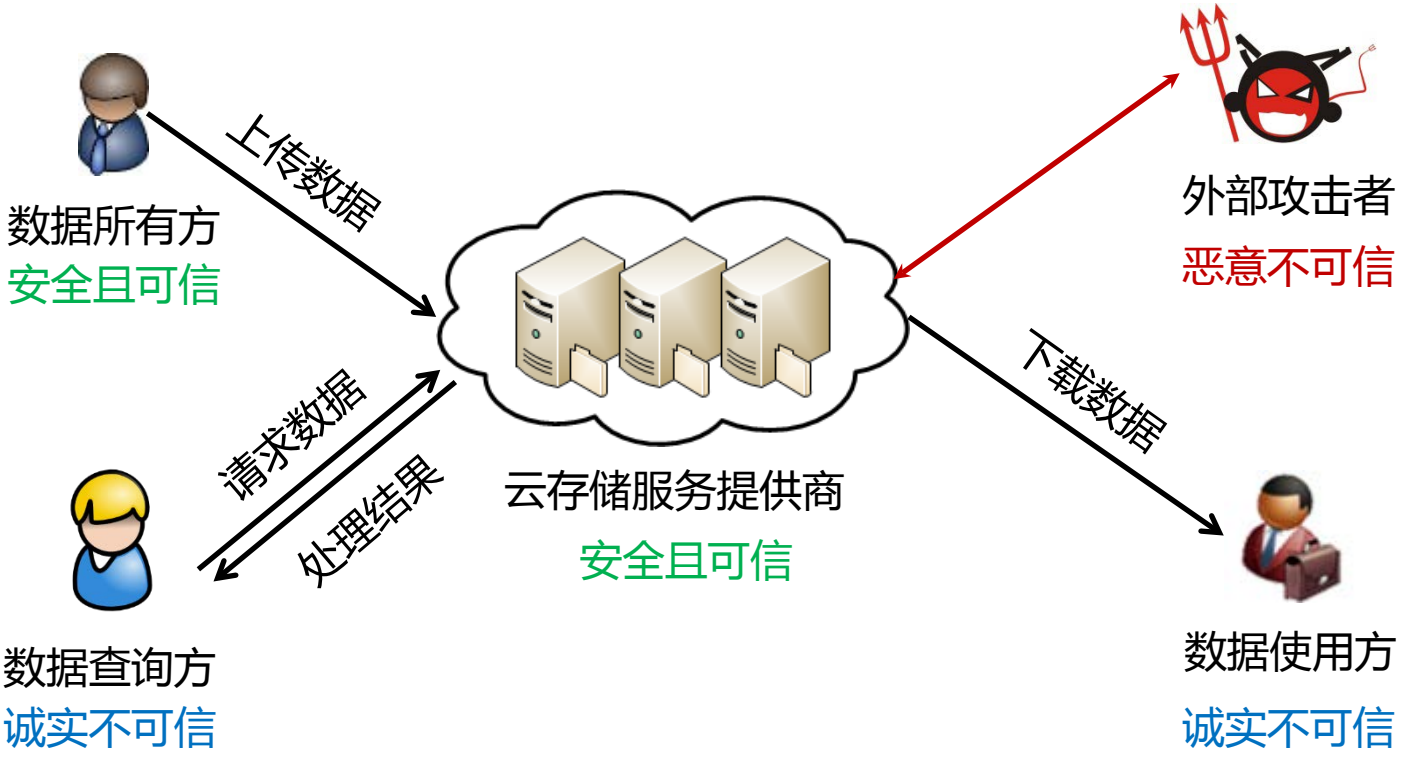
数据库安全

Contents

目录

隐私保护数据发布体系架构

特定情境下云存储服务安全可信，然而数据查询仍会泄露用户个人隐私



传统方法：数据脱敏发布

姓名	工作	性别	年龄	患病
Alice	作家	女	30	感冒
Bob	工程师	男	35	肝炎
Cathy	作家	女	30	HIV
Mike	律师	男	38	HIV
Emily	舞蹈家	女	30	感冒
Fred	工程师	男	38	肝炎
Gladys	舞蹈家	女	30	HIV
Henry	律师	男	39	感冒
Irene	舞蹈家	女	32	感冒

典型数据脱敏方法：删除标识信息

姓名	工作	性别	年龄	患病
***	作家	女	30	感冒
***	工程师	男	35	肝炎
***	作家	女	30	HIV
***	律师	男	38	HIV
***	舞蹈家	女	30	感冒
***	工程师	男	38	肝炎
***	舞蹈家	女	30	HIV
***	律师	男	39	感冒
***	舞蹈家	女	32	感冒

典型数据脱敏方法：盲化个人信息

姓名	工作	性别	年龄	患病
***	作家	女	[30-33)	感冒
***	工程师	男	[33-36)	肝炎
***	作家	女	[30-33)	HIV
***	律师	男	[36-39)	HIV
***	舞蹈家	女	[30-33)	感冒
***	工程师	男	[36-39)	肝炎
***	舞蹈家	女	[36-39)	HIV
***	律师	男	[39, +∞)	感冒
***	舞蹈家	女	[30-33)	感冒

数据发布典型攻击方式

- 攻击者已知用户姓名、性别、工作情况等个人信息，推测用户患病情况
- 个人信息可通过其他来源泄露的数据信息、社会工程学攻击等方式获得

姓名	性别	工作	年龄
Alice	女	作家	[30-33)
Bob	男	工程师	[33-36)
Cathy	女	作家	[30-33)
Mike	男	律师	[36-39)
Emily	女	舞蹈家	[30-33)
Fred	男	工程师	[36-39)
Irene	女	舞蹈家	[30-33)

攻击者已知数据

?

性别	工作	年龄	患病
男	工程师	[33-36)	肝炎
男	工程师	[36-39)	肝炎
男	律师	[36-39)	HIV
女	作家	[30-33)	感冒
女	作家	[30-33)	HIV
女	舞蹈家	[30-33)	HIV
女	舞蹈家	[30-33)	HIV

发布数据

数据发布典型攻击方式

- 问题：攻击者已知用户的个人信息，可从发布数据表中推测敏感信息
- 原因：现实世界中的数据存在关联性

性别	工作	年龄	患病
男	工程师	[33-36)	肝炎
男	工程师	[36-39)	肝炎
男	律师	[36-39)	HIV
女	作家	[30-33)	感冒
女	作家	[30-33)	HIV
女	舞蹈家	[30-33)	HIV
女	舞蹈家	[30-33)	HIV

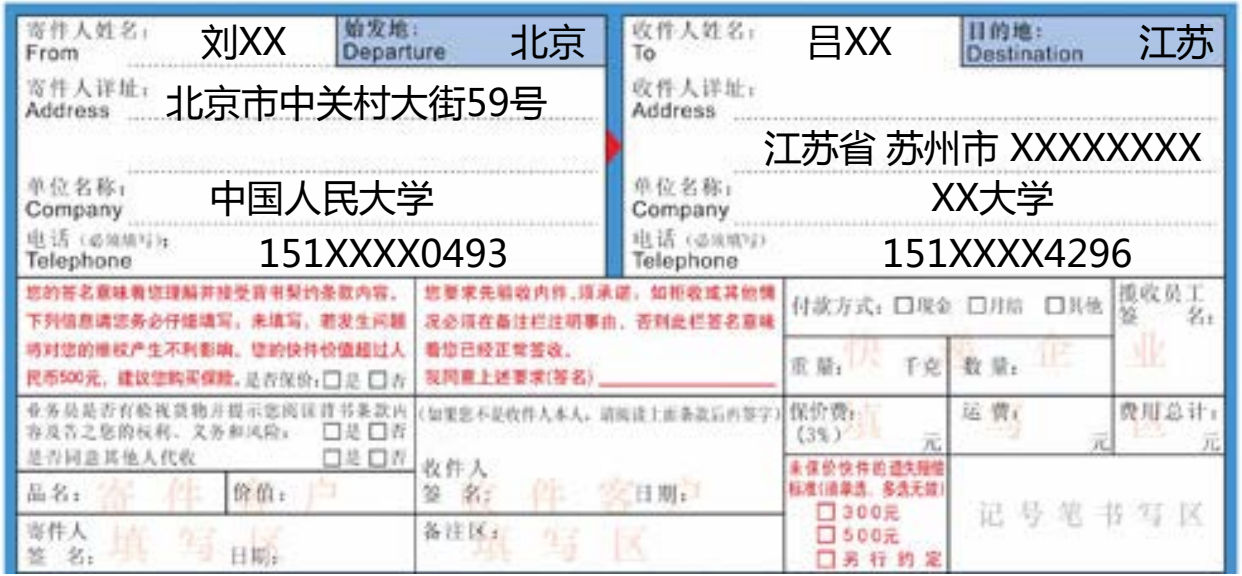
发布数据

- {男, 律师, [36 -39)} → Mike
- {男, 工程师, [33-36)}→ Bob
- {男, 工程师, [36-39)}→ Fred
- {女, 作家, [30 -33)} → ???
- {男, 舞蹈家, [30-33)}→ ???

数据库安全

隐私保护数据发布

数据发布典型攻击方式



{卖家：北京，中国人民大学，刘XX，15110060493}

→ {买家：江苏，**南京大学**，吕XX，15131184296} → {图书，30.00元}

原因：容易获得卖家信息，邮寄地址可从速递单号查询到，电话号码通过百度得到

数据库安全

Contents

目录

4.1 K-匿名（K-Anonymity）

- 目标：保证数据表发布时，个人信息的组合结果至少有K个重复项
- 方法：删除数据项；盲化数据项

性别	工作	年龄	患病
男	工程师	[33-36)	肝炎
男	工程师	[36-39)	肝炎
男	律师	[36-39)	HIV
女	作家	[30-33)	感冒
女	作家	[30-33)	HIV
女	舞蹈家	[30-33)	HIV
女	舞蹈家	[30-33)	HIV

发布数据

性别	工作	年龄	患病
男	专业工作	[33-39)	肝炎
男	专业工作	[33-39)	肝炎
男	专业工作	[33-39)	HIV
女	艺术家	[30-33)	感冒
女	艺术家	[30-33)	HIV
女	艺术家	[30-33)	HIV
女	艺术家	[30-33)	HIV

K-匿名发布数据（K=3）

4.2 L-多样（L-Diversity）

- 要解决 问题：经过K-匿名处理后的数据，仍可能批量泄露用户敏感信息
- 原因：L-多样性没有对等价类中敏感信息的取值做任何限制

姓名	性别	工作	年龄
Alice	**女**	**作家**	**[30-33)**
Bob	男	工程师	[33-36)
Cathy	**女**	**作家**	**[30-33)**
Doug	男	律师	[36-39)
Emily	**女**	**舞蹈家**	**[30-33)**
Fred	男	工程师	[36-39)
Irene	**女**	**舞蹈家**	**[30-33)**

性别	工作	年龄	患病
男	专业工作	[33-39)	肝炎
男	专业工作	[33-39)	肝炎
男	专业工作	[33-39)	感冒
女	**艺术家**	**[30-33)**	**HIV**
女	**艺术家**	**[30-33)**	**HIV**
女	**艺术家**	**[30-33)**	**HIV**
女	**艺术家**	**[30-33)**	**HIV**

Alice、Cathy、Emily、Irene均患HIV

4.2 L-多样（L-Diversity）

- 目标：：保证数据表发布时，敏感信息的组合结果至少有L种情况
- 方法：进一步盲化用户个人信息，使敏感信息结果更加丰富

性别	工作	年龄	患病
男	专业工作	[33-39)	肝炎
男	专业工作	[33-39)	肝炎
男	专业工作	[33-39)	感冒
女	艺术家	[30-33)	HIV
女	艺术家	[30-33)	HIV
女	艺术家	[30-33)	HIV
女	艺术家	[30-33)	HIV

K-匿名发布数据（K=3）

性别	工作	年龄	患病
男	专业工作	[33-39)	肝炎
男	专业工作	[33-39)	肝炎
**	在职	[30-39)	感冒
女	艺术家	[30-39)	HIV
女	艺术家	[30-39)	HIV
女	艺术家	[30-39)	HIV
女	艺术家	[30-39)	HIV

L-多样发布数据（L=2）

4.3 T-接近（T-Closeness）

- 要解决问题：经过L-多样处理后的数据，仍可能批量泄露用户敏感信息的范围
- 原因：现实世界中数据分布存在概率性规律，敏感属性可能具有语义相似性

姓名	性别	工作	年龄
Alice	**女**	**作家**	**[30-33)**
Bob	男	工程师	[33-36)
Cathy	**女**	**作家**	**[30-33)**
Doug	男	律师	[36-39)
Emily	**女**	**舞蹈家**	**[30-33)**
Fred	男	工程师	[36-39)
Irene	**女**	**舞蹈家**	**[30-33)**

性别	工作	年龄	患病
男	专业工作	[33-39)	肝炎
男	专业工作	[33-39)	肺炎
男	专业工作	[33-39)	感冒
女	**艺术家**	**[30-33)**	**胃溃疡**
女	**艺术家**	**[30-33)**	**胃痉挛**
女	**艺术家**	**[30-33)**	**胃胀气**
女	**艺术家**	**[30-33)**	**胃癌**

Alice、Cathy、Emily、Irene均患胃病

4.3 T-接近（T-Closeness）

- 目标：保证数据发布时，敏感信息的组合概率分布与完整表格中敏感信息的概率分布差值小于门限值T。
- 方法：对表中敏感信息出现概率进行评估，按照门限值T对记录进行重排。

性别	工作	年龄	患病
男	专业工作	[33-39)	肝炎
男	专业工作	[33-39)	肺炎
男	专业工作	[33-39)	感冒
女	艺术家	[30-33)	胃溃疡
女	艺术家	[30-33)	胃痉挛
女	艺术家	[30-33)	胃胀气
女	艺术家	[30-33)	胃癌

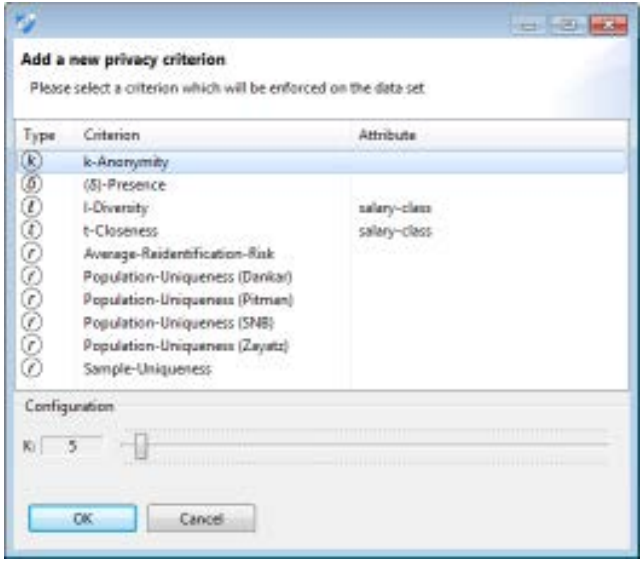
L-多样发布数据（L=3）

性别	工作	年龄	患病
**	在职	[30-39)	肝炎
**	在职	[30-39)	胃溃疡
**	在职	[30-39)	胃痉挛
**	在职	[30-39)	肺炎
**	在职	[30-39)	感冒
**	在职	[30-39)	胃胀气
**	在职	[30-39)	胃癌

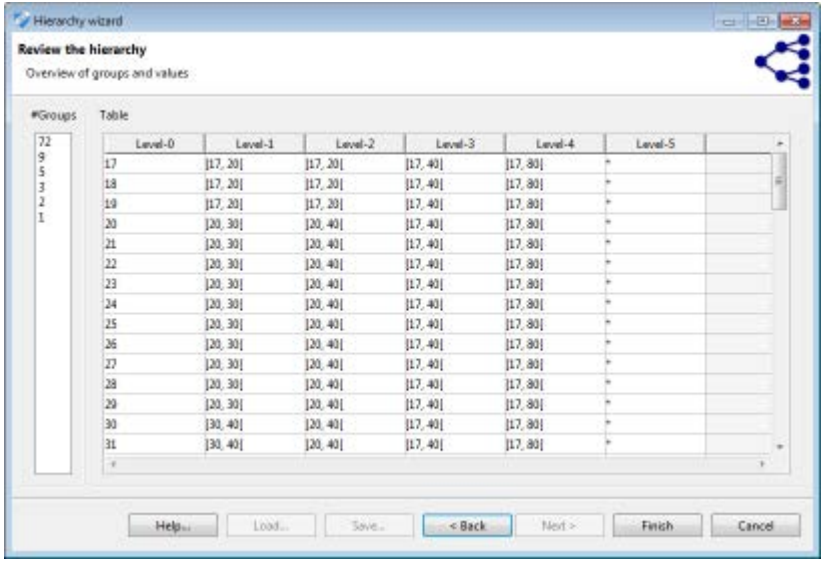
T接近发布数据（$T=\frac{3}{7}$，同时满足L=3）

4.4 实例：数据匿名化工具ARX

- 简介：开源个人隐私数据匿名化工具（Data Anonymization Tool）
- 链接： http://arx.deidentifier.org/
- 特性：支持K-匿名、L-多样、T-临近、δ-披露等多种隐私数据发布方案



所支持的隐私数据发布方案



数据盲化定义页面

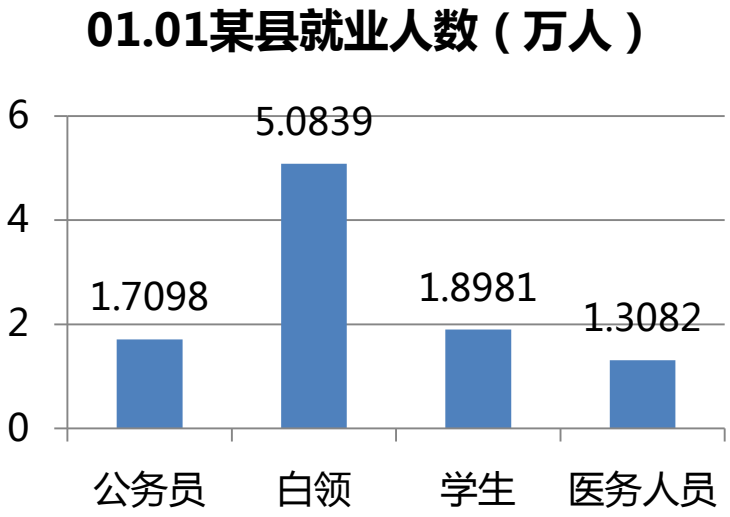
数据库安全

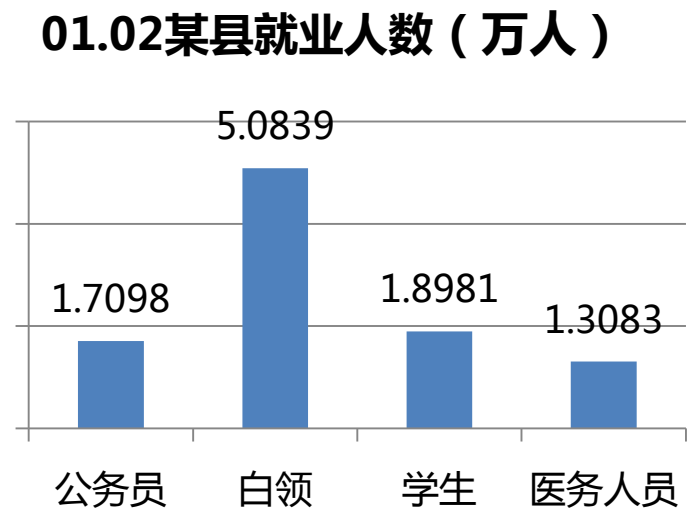
Contents

目录

5 隐私保护统计数据发布

- 目标：各云存储服务商会周期性发布数据统计信息，更好地服务大众
- 问题：统计数据发布也会泄露个体用户隐私信息





攻击结果：01月02日某县新增就业是医务人员

统计数据发布典型攻击方式

- 在统计数据库中，允许用户查询聚集类型的信息（例如合计、平均值等），但是不允许查询单个记录信息。例如，查询“程序员的平均工资是多少？”是合法的，但是查询“程序员张勇的工资是多少？”就不允许。
- 问题：攻击者可以从合法的查询中推导出不合法的信息。
- 原因：存在着隐蔽的信息通道

实例1：存在以下两个合法的查询：

- 本公司共有多少女高级程序员？
- 本公司女高级程序员的工资总额是多少？

如果第1个查询的结果是“1”，那么第2个查询的结果显然就是这个程序员的工资数。为此，可以规定任何查询至少要涉及N个以上的记录(N足够大)。

统计型数据发布典型攻击方式

实例2：某个用户A想知道另一用户B的工资数额，他可以通过下列两个合法查询获取：

- 用户A和其他N个程序员的工资总额是多少？
- 用户B和其他N个程序员的工资总额是多少？

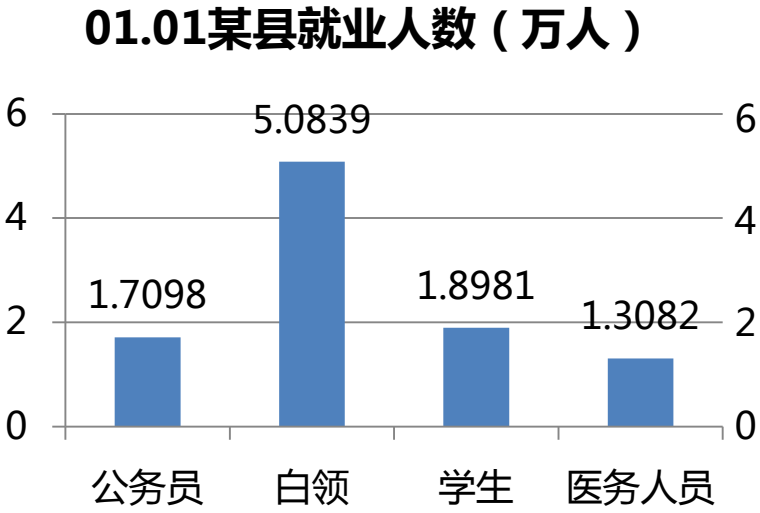
假设第一个查询的结果是X，第二个查询的结果是Y，由于用户A知道自己的工资是Z，那么他可以计算出用户B的工资=Y-(X-Z)。

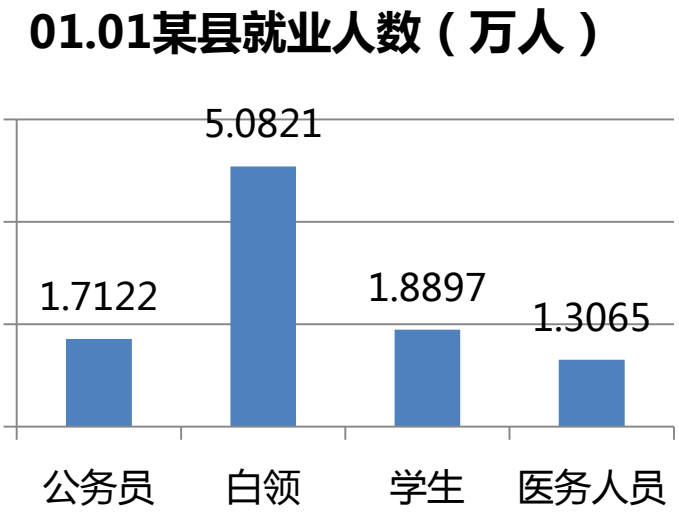
这个例子的关键之处在于两个查询之间有很多重复的数据项（即其他N个程序员的工资）。因此可以再规定任意两个查询的相交数据项不能超过M个。可以证明，在上述两条规定下，如果想获知用户B的工资额，用户A至少需要进行1+(N-2)/M次查询。

当然可以继续规定任一用户的查询次数不能超过1+(N-2)/M，但是如果两个用户合作查询就可以使这一规定仍然失效。

核心技术：差分隐私（Differential Privacy）

- 目标：令攻击者从发布数据中无法判断某个体数据是否在数据集内
- 方法：统计数据发布结果中增加小量噪声项



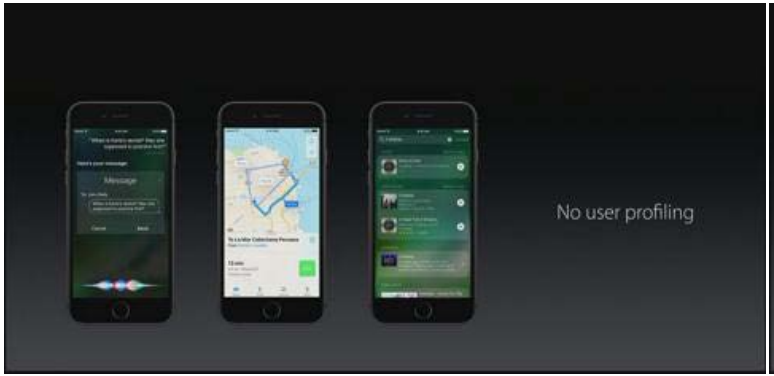


核心技术：差分隐私（Differential Privacy）

- 差分隐私保护的思想就是要保证任一个体在数据集中或者不在数据集中时，对最终发布的查询结果几乎没有影响。具体来说，设有两个几乎完全相同的数据集（两者的区别仅在于一个记录不同），分别对这两个数据集进行查询访问，同一个查询在两个数据集上产生同一结果的概率比值接近于1.
- 差分隐私保护研究方向：
  - 隐私保护统计数据发布：非交互式框架，隐私保护+满足数据分析，技术：直方图、划分、采样-过滤等
  - 面向数据挖掘和学习的差分隐私保护技术：解决高层隐私需求带来的问题，技术：top-k频繁模式挖掘、分类及回归分析
  - 基于差分隐私的查询处理：解决如何以较小的隐私预算与较低误差来响应查询，交互式框架+线性与批量查询
  - 基于差分隐私的应用系统等：在各类环境中通用

实例：苹果公司在大数据分析中引入差分隐私

2016年WWDC大会上，苹果公司发布称采用差分隐私技术保护用户隐私



数据库安全

Contents

目录

数据库安全

总结

- 研究方向与前沿成果
  - 隐私保护分级与风险评估
  - 兼容实用性和动态感知的隐私保护策略
  - 隔断信息链接的方法
  - 在数据库即服务（DaaS）框架下的外包数据库加密，支持数据库所有操作，不改变云存储架构和现有用户习惯
  - 与区块链技术融合的分布式数据库安全与隐私保护



谢　谢！