北京郵電大學
Beijing University of Posts and Telecommunications

大数据共享及其安全隐私

石瑞生
网络空间安全学院

提纲

- 隐私的概念
- 数据匿名化
- 匿名化技术与反匿名化技术的军备竞赛
- 差分隐私技术

大数据共享与隐私保护

- 数据已成为数字经济时代核心的生产要素。
  - 从大数据整体态势上看，数据的规模将变得更大，出现了数据资源化的趋势。
  - 多源数据通过关联分析和深度开采，数据的价值才愈发凸显。因此，数据共享机制在大数据服务的发展中扮演着日益重要的角色。
- 阻碍数据大规模共享的一个重要问题就是隐私保护问题。
  - 为了保护隐私，已经提出了很多保护隐私的计算方法，包括同态加密、安全多方计算、函数加密等等。加密能够解决很多安全隐私的问题，但仅仅依靠加密技术是不够的。
  - 在社交网络和其它的公共网站上，可以公开免费获得大量关于个人的数据，任何一个想要做坏事的人都可以从任意数量的在线资源通过交叉引用来建立关于他们的目标的轮廓(profile)。
  - 因此，对于需要公开的数据或者准备共享的数据，在数据发布之前需要进行匿名化处理以保护用户的隐私。

隐私的概念

提纲

- 隐私的定义
- 如何度量隐私
- 隐私保护面临的威胁

隐私的定义

- 简单地说，隐私就是个人、机构等实体不愿意被外部世界知晓的信息。
  - 在具体应用中，隐私即为数据所有者不愿意被披露的敏感信息，包括敏感数据以及数据所表征的特性。
  - 通常我们所说的隐私都指敏感数据，如个人的薪资、病人的患病记录、公司的财务信息等。
- 什么是敏感数据?
  - 敏感数据就是，不是所有人都能够获得的公开数据。
- 隐私的定义中包含主体（外界）和客体（敏感数据）两个概念，不同文化、不同个体对这两个概念的界定会有很大差异。

外界如何定义？例如，你的工资信息，对于公司的人力资源部门和你的家人，不是隐私；对于之外的人，就是隐私。不同的文化和个体，对这个范围的界定也会不同。个人身份信息，对国家是公开的，对企业是保密的。

主观上对敏感数据的认定不同：当针对不同的数据以及数据所有者时，隐私的定义也会存在差别的。例如保守的病人会视疾病信息为隐私，而开放的病人却不视之为隐私。

隐私的分类

- 从隐私所有者的角度而言，隐私可以分为两类
  - 个人隐私（Individual privacy）：任何可以确认特定个人或与可确认的个人相关、但个人不愿被暴露的信息，都叫做个人隐私，如身份证号、就诊记录等。
  - 共同隐私（Corporate privacy）：共同隐私不仅包含个人的隐私，还包含所有个人共同表现出但不愿被暴露的信息。如公司员工的平均薪资、薪资分布等信息，再如两个人之间的关系信息。

隐私的度量与量化表示

- 数据隐私的保护效果是通过攻击者披露隐私的多寡来侧面反映的。隐私度量可以统一用“披露风险”（Disclosure Risk）来描述。
  - 披露风险表示攻击者根据所发布的数据和其它背景知识（Background Knowledge），可能披露隐私的概率。通常，关于隐私数据的背景知识越多，披露风险越大。
  - 若s表示敏感数据，事件S_k表示“攻击者在背景知识K的帮助下揭露敏感数据s”，则披露风险r(S,K)表示为： $r(S,K)=Pr(S_k)$
- 对数据集而言，若数据所有者最终发布数据集D的所有敏感数据的披露风险都小于阈值α，$\alpha\in[0,1]$，则称该数据集的披露风险为α。
  - 特别地，不做任何处理所发布数据集的披露风险为1；当所发布数据集的披露风险为0时，这样发布的数据被称为实现了完美隐私（Perfect Privacy）。

完美隐私实现了对隐私最大程度的保护，但由于对攻击者背景知识的假设是不确定的，因此实现对隐私的完美保护也只在具体假设、特定场景下成立，真正的完美保护并不存在。

威胁分析：谁有可能侵犯大众的隐私呢？

- 随着互联网的兴起，网络隐私成为一个大家日益关注的问题
- 谁在侵犯大众的隐私？
  - 政府：公共安全；国际政治
    - 国家级敌手（State-level adversary）
    - 2013年7月，斯诺登事件，美国的“棱镜计划”
      - 数据与元数据
  - 企业：经济利益
    - 苹果，谷歌，BAT，等互联网公司
  - 黑客及一些犯罪组织：黑产

网络空间的黑色产业链
　　系统漏洞，被明码标价
　　入侵工具，像武器一样容易购买
　　水军，花钱就能够发起内容攻击
云计算技术，使得普通人的攻击能力日益增强

用户隐私泄漏事件

- 90年代，美国马萨诸塞州政府雇员医疗数据的用户隐私泄漏事件
- 美国在线（AOL）数据发布（2006年）
- 网飞公司(Netflix)数据隐私泄漏事件（2006年）
- 社交网络上隐私泄漏事件

社交网络上隐私泄漏事件

社交网络给我们带来了很多方便：能够帮助我们更方便与朋友保持联系，更便捷地获取高质量的信息（经过朋友筛选的、可能是自己感兴趣的信息）。

但是，社交网络也是采集隐私信息的理想场所。

社交网络上看似平常的社交互动行为，其实已经涉及个人隐私保护问题。

例如，照片可能会包含时间信息、地理位置信息等敏感信息；【上传照片的可交换图像文件(exchangeable image file, EXIF)信息中可能包含拍照的时间、GPS坐标等信息】

尽管大多数应用已经对这些敏感信息进行了过滤，但基于照片本身进行分析来确定照片拍摄的地理位置信息，很多时候也并不困难。

2012年7月15日18点28分，京东商城的小家电总监庄佳——微博名@vikizhuang发布了3张西红柿熟透的照片，并称：“惊喜地发现结出小番茄啦，只有3只，舍不得摘~~”；11分钟后，即18点39分，京东商城首席执行官刘强东在新浪微博（@刘强东）也发布了一张西红柿熟透的照片，配文中称：“阳台上的西红柿终于熟了”。

细心的网友对比后发现，两人发布的照片是同一场景下的同一棵西红柿，由此一段“西红柿爱情”被疯狂传播。

案例二：网友40分钟找出演员王珞丹的住址

清华大学学生罗霄宇仅仅根据电视剧《杜拉拉升职记》的女主角王珞丹在微博上发布的从她家向外拍的两张照片，利用谷歌地球(Google Earth)和简单的地理常识，在短短40分钟之内就推断出王珞丹的家庭住址。

美国在线（AOL）数据发布（2006年）

- 2006年8月4日，美国在线（AOL）公司的研究部门在互联网上发布了超过65万用户在过去三个月的搜索关键字，以供公众对搜索技术进行研究。
  - 该公司对发布的数据进行了匿名化处理，但仅仅是把用户的账号用一个随机号码代替，并没有对用户所提交的搜索关键字进行任何处理。
  - 随后，《纽约时报》成功将部分数据去匿名化，并在经过当事人同意后，公开了其中一位搜索用户的真实身份。
- 这起隐私泄漏事件引起了人们的广泛关注，并导致美国在线公司首席技术官辞职。随后，美国在线公司因为此事件在北加州地方法院被起诉。

网飞公司(Netflix)数据隐私泄漏事件

- 2006年，网飞公司投资100万美元举办了一个为期三年的推荐系统算法竞赛，并发布了一些用户的影评数据供参赛者测试。出于隐私保护，网飞公司在发布数据前将所有用户的个人信息移除，仅保留了每个用户对各个电影的评分以及评分的时间戳。
- 然而，来自德州大学奥斯汀分校的两位研究人员利用网飞用户影评数据与公开的互联网电影数据库（IMDB)用户影评数据之间的相关性，将网飞公司的一部分匿名用户与公开的IMDB用户进行了一一对应，由此获得了IMDB用户在网飞公司网站上的全部电影浏览信息（包括涉及敏感题材的电影）。
- 为此，2009年，网飞公司遭到了4位用户的起诉，也不得不取消了原定于2010年举行的第二届算法竞赛。

Arvind Narayanan and Vitaly Shmatikov, **De-anonymizing Social Networks**, 30th IEEE Symposium on Security and Privacy, 2009

数据匿名化技术

人们是如何对数据进行匿名化处理的呢？

- 发布-遗忘模型（Release-and-Forget Model）
  - 顾名思义，该模型包含两部分内容。
    - 1）发布（Release）：数据管理员对数据进行匿名化处理后发布数据，包括公开发布数据，秘密地向第三方发布数据，或者在自己的组织内部发布数据；
    - 2）然后她会忘记，这意味着她不会试图在发布后追踪记录的情况。
    - 而在数据发布之前，数据发布者并没有轻率地将要发布的数据对象置于危险之中，而是对敏感数据进行了处理。
- 我们学习“发布-遗忘”模型的原因有两个。
  - 首先，这种模型被广泛使用。
  - 其次，这种技术往往是有缺陷的。许多再识别技术的最新进展特别针对“发布-遗忘”匿名化。

如何对数据进行匿名化处理？

- 为了保护本表中的人员隐私，医院数据库管理员将在发布此数据之前采取以下步骤

1）识别身份信息。
首先，管理员将挑选出她认为可以用来识别个人的任何字段。通常，她不仅要挑选像姓名和社会安全号码这样的显示标识符，而且还要考虑字段的组合，在组合的情况下可能将表中的记录与患者的身份关联起来。

2）抑制

3）泛化

4）聚合

表 6-1　原始(非匿名)数据

姓名	种族	出生日期	性别	邮政编码	疾病
Sean	黑人	9/20/1965	男	02141	呼吸短促
Daniel	黑人	2/14/1965	男	02141	胸痛
Kate	黑人	10/23/1965	女	02138	眼疼
Marion	黑人	8/24/1965	女	02138	喘息
Helen	黑人	11/7/1964	女	02138	关节疼痛
Reese	黑人	12/1/1964	女	02138	胸痛
Forest	白人	10/23/1964	男	02138	呼吸短促
Hilary	白人	3/15/1965	女	02139	高血压
Philip	白人	8/13/1964	男	02139	关节疼痛
Jamie	白人	5/5/1964	男	02139	发烧
Sean	白人	2/13/1967	男	02138	呕吐
Adrien	白人	3/21/1967	男	02138	背疼

抑制（Suppression）：不发布

表 6-1　原始(非匿名)数据

姓名	种族	出生日期	性别	邮政编码	疾病
Sean	黑人	9/20/1965	男	02141	呼吸短促
Daniel	黑人	2/14/1965	男	02141	胸痛
Kate	黑人	10/23/1965	女	02138	眼疼
Marion	黑人	8/24/1965	女	02138	喘息
Helen	黑人	11/7/1964	女	02138	关节疼痛
Reese	黑人	12/1/1964	女	02138	胸痛
Forest	白人	10/23/1964	男	02138	呼吸短促
Hilary	白人	3/15/1965	女	02139	高血压
Philip	白人	8/13/1964	男	02139	关节疼痛
Jamie	白人	5/5/1964	男	02139	发烧
Sean	白人	2/13/1967	男	02138	呕吐
Adrien	白人	3/21/1967	男	02138	背疼

表 6-2　抑制四个标识符字段

种族	疾病
黑人	呼吸短促
黑人	胸痛
黑人	眼疼
黑人	喘息
黑人	关节疼痛
黑人	胸痛
白人	呼吸短促
白人	高血压
白人	关节疼痛
白人	发烧
白人	呕吐
白人	背疼

一个根本的矛盾：严重的抑制使数据对研究几乎毫无用处。虽然研究人员可以使用剩余的数据来跟踪疾病的发生率，但由于年龄，性别和住所信息已经被删除，研究人员将无法得出许多其它有趣和有用的结论。

泛化（Generalization）

- 泛化（Generalization）：为了更好地实现效用和隐私之间的平衡，匿名者可能会泛化而不是抑制标识符。
  - 这意味着她将改变而不是删除标识符值以增强隐私，同时保持数据的实用性。
  - 例如，匿名者可以选择不发布姓名字段，将出生日期归纳为出生年份，并通过只保留前三个数字来概括邮政编码。

表 6-3　泛化

种族	出生日期	性别	邮政编码	疾病
黑人	1965	男	021 *	呼吸短促
黑人	1965	男	021 *	胸痛
黑人	1965	女	021 *	眼疼
黑人	1965	女	021 *	喘息

- 聚合（Aggregation）：通常，分析师只需要汇总统计数据，而不是原始数据。数十年来，统计人员一直在研究如何发布汇总统计数据，同时保护数据主体免于再识别。
- 然而，一系列经过匿名化处理的脱敏数据导致的隐私泄漏事件，使人们对匿名化的效用产生了严重的怀疑。

匿名化技术与反匿名化技术的军备竞赛

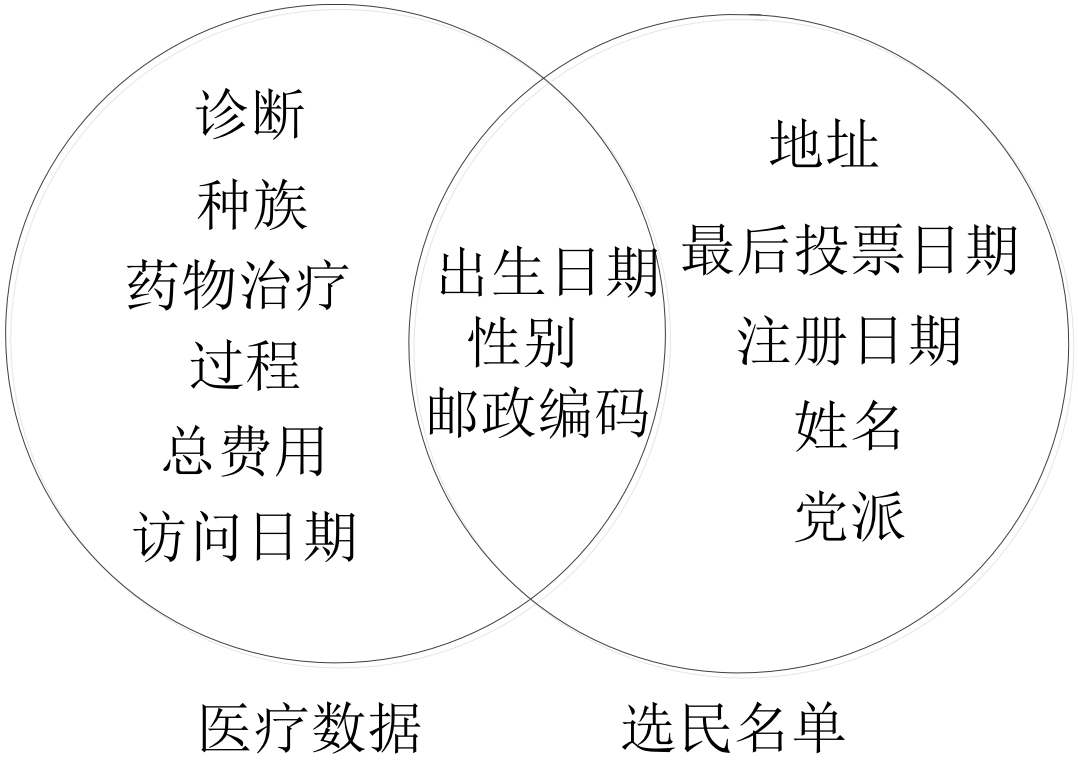
重大用户隐私泄漏事件

- 20世纪最著名的用户隐私泄漏事件发生在美国马萨诸塞州。
  - 90年代中叶，该州团体保险委员会(Group Insurance Commission)决定发布州政府雇员的“经过匿名化处理的”医疗数据，以助公共医学研究。在数据发布之前，委员会对潜在的隐私问题已有所认识，因此删除了数据中所有的敏感信息，例如姓名、住址和社会安全号码(social security number)。
  - 然而1997年，麻省理工学院博士生拉坦娅·斯威尼 (Latanya Sweeney)（现任哈佛大学教授）成功破解了这份匿名数据，并找到了时任马萨诸塞州州长威廉·威尔德(William Weld) 的医疗记录，还将该记录直接寄给了州长本人。

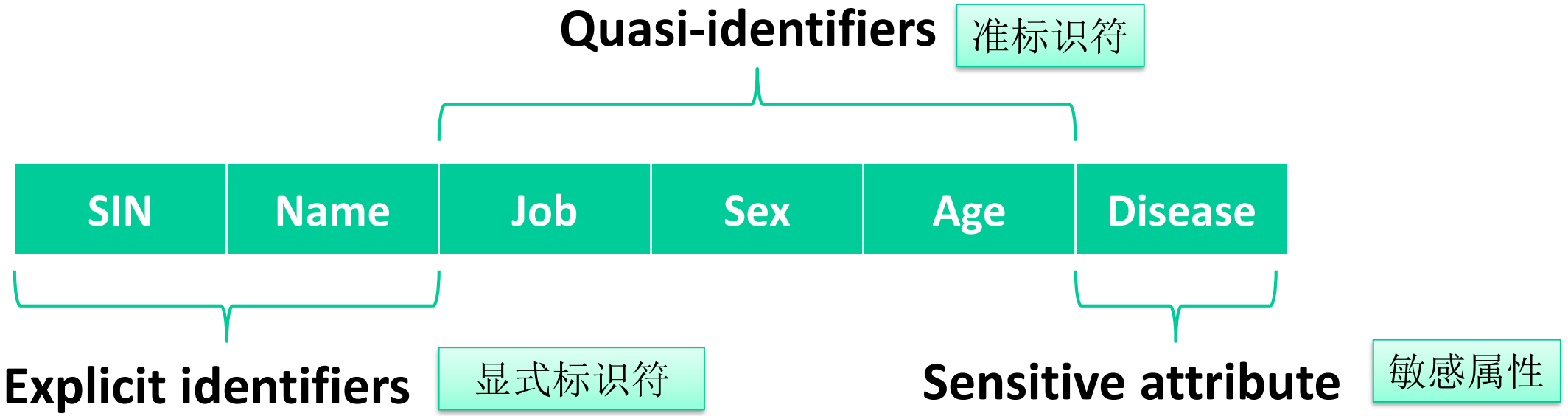
攻击方法

- 攻击方法
  - 斯威尼的攻击手法非常简单。
  - 首先，她下载了团体保险委员会的“匿名”数据。这份数据虽然删除了很多个人信息，但仍保留了三个关键字段：患者的出生日期、性别和邮编。
  - 然后，她花20美元购买了一份公开的马萨诸塞州剑桥市的投票人名单（州长亦在其中），名单中包含投票人的姓名、住址、邮编、出生日期、性别以及其他信息。
  - 最后，斯威尼将投票人名单与匿名医疗数据进行匹配，发现医疗数据中仅有6人与州长的生日相同，而其中只有3人是男性，当中又仅有一人与州长的邮编相同。
  - 由此，斯威尼准确地定位了州长的医疗记录。

斯威尼进一步研究发现，87%的美国人拥有唯一的出生日期、性别和邮编三元组信息；同时发布三元组信息，事实上几乎等同于直接公布姓名。



防御措施（K-匿名隐私保护模型）



按照斯威尼的思路，可以数据记录的属性分为三类：
1）显式标识符（Explicit Identifiers）：能唯一标识单一个体的属性，如身份证号码、姓名等。
2）准标识符（Quasi-Identifiers，QID）：联合起来能唯一标识一个人的多个属性，如邮编、生日、性别等联合起来则可能是准标识符。
3）敏感属性（Sensitive Attribute）：包含隐私数据的属性，如疾病、薪资等。
简单的匿名化处理，只是把显式标识符过滤掉。但是，一个人的记录还是可以通过链接到他的准标识符而被识别出来。

如果表中的一个记录有某个值qid，那么至少有k - 1个其他记录也有该值qid。

K匿名隐私保护模型

- K-匿名方法通常也是采用泛化和抑制技术对原始数据进行匿名化处理以便得到满足k-匿名规则的匿名数据，从而使得攻击者不能根据发布的匿名数据准确的识别出目标个体的记录。
- 泛化（Generalization）通常是将QID的属性用更概括、抽象的值替代具体描述值。
  - 泛化的核心思想就是一个值被一个不确切的，但是忠于原值的值代替。
    - 数据集中的数据和对象通常包含原始概念层的细节信息，数据泛化（概化）是将数据集中与任务相关的数据由较低的概念层次抽象到较高的概念层次的过程。
- 抑制（Suppression）是指针对标识符做不发布处理。
  - 因为标识符和某些属性有很强的查询能力，所以针对这些属性做抑制处理是比较恰当的选择。有时抑制方法可以降低或减小泛化的代价。

K匿名隐私保护模型

身份证号	姓名	工作	性别	年龄	疾病
12345679	Steven	工程师	男性	35	肝炎
12345678	Frank	律师	男性	38	艾滋病
12345601	Andy	律师	男性	30	肝炎
12345603	Alice	作家	女性	30	艾滋病
12345610	Lily	舞蹈家	女性	30	艾滋病
12345670	Ellen	作家	女性	30	艾滋病
12345607	Gloria	舞蹈家	女性	30	艾滋病

（a）原始数据集

工作	性别	年龄	疾病
工程师	男性	35	肝炎
律师	男性	38	艾滋病
律师	男性	30	肝炎
作家	女性	30	艾滋病
舞蹈家	女性	30	艾滋病
作家	女性	30	艾滋病
舞蹈家	女性	30	艾滋病

（b）去除显式标识符

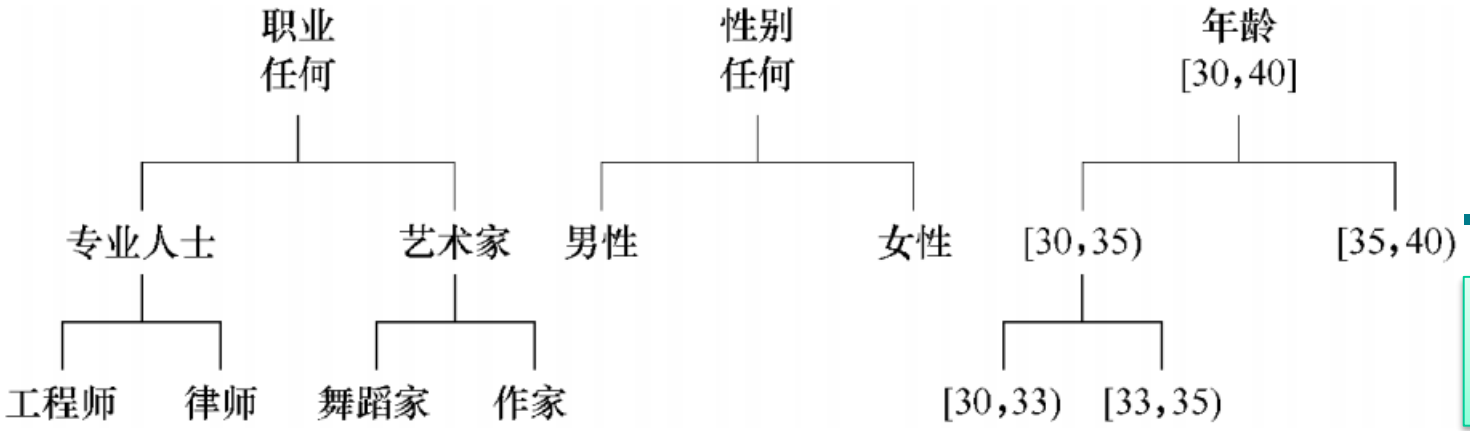
工作	性别	年龄	疾病
工程师	男性	35	肝炎
律师	男性	38	艾滋病
律师	男性	30	肝炎
作家	女性	30	流感
舞蹈家	女性	30	艾滋病
作家	女性	30	艾滋病
舞蹈家	女性	30	艾滋病

（c）记录链接式攻击

例如，攻击者知道他的老板在住院，因此，他就知道他的老板的医疗记录将会出现在医院泄露出的患者数据库中。并且，对于这个攻击者来说，得到他老板的邮编、出生日期和性别也不是难事，而这些就可以作为相关攻击时的准标识符。

攻击者知道Frank是一个38岁的男律师，并且这个数据集中包含Frank的数据，就能够推断出Frank的疾病是艾滋病。

这种攻击方式被称为记录链接式攻击(attack of record link)。



经过匿名化处理的数据集可以抵御前述的记录链接式攻击 (attack of record link)。

工作	性别	年龄	疾病
工程师	男性	35	肝炎
律师	男性	38	艾滋病
律师	男性	30	肝炎
作家	女性	30	艾滋病
舞蹈家	女性	30	艾滋病
作家	女性	30	艾滋病
舞蹈家	女性	30	艾滋病

（c）记录链接式攻击

表 6-6　经过 *K* 匿名处理后的数据集

工作	性别	年龄	疾病
专业人士	男性	[30,40)	肝类
专业人士	男性	[30,40)	艾滋病
专业人士	男性	[30,40)	肝炎
艺术家	女性	30	艾滋病
艺术家	女性	30	艾滋病
艺术家	女性	30	艾滋病
艺术家	女性	30	艾滋病

L多样性隐私保护模型

- 但是，通过观察我们很容易发现一个新的问题：如果攻击者知道Lily是一名30岁的女性舞蹈演员，并且该数据集中包含Lily的数据，从公开的数据集中能够推断出Lily的疾病是艾滋病。

这种攻击方式被称为属性链接（Attribute linkage）类攻击。

当遭受属性链接类攻击时，攻击者也许不能精确地识别目标受害者的记录，但可能从被公布的数据T中基于与受害者所属的团体相联系的一系列敏感值集合推理出他的敏感值。

表 6-6　经过 K 匿名处理后的数据集

工作	性别	年龄	疾病
专业人士	男性	[30,40)	肝类
专业人士	男性	[30,40)	艾滋病
专业人士	男性	[30,40)	肝炎
艺术家	女性	30	艾滋病
艺术家	女性	30	艾滋病
艺术家	女性	30	艾滋病
艺术家	女性	30	艾滋病

L多样性隐私保护模型

Machanavajjhala 等人出了多样性原则，并称之为“ℓ-多样性”来阻止属性链接(attribute linkage)攻击。 ℓ-多样性要求每个 qid 组至少包含有ℓ个有“较好代表性”的敏感值。“较好的代表性”的最简单理解是确保每个 qid 组中的敏感属性都有ℓ个不同的值。

T相近隐私保护模型

- 满足K匿名和 L-多样性的要求，是不是就没有问题了呢?

表 6-7　概率分布问题

工作	性别	年龄	收入/万美元
*	男性	*	2.05
*	男性	*	5.19
*	男性	*	7.05
作家	女性	30	9.59
舞蹈家	女性	30	1.99
作家	女性	30	8.55
舞蹈家	女性	30	2.00

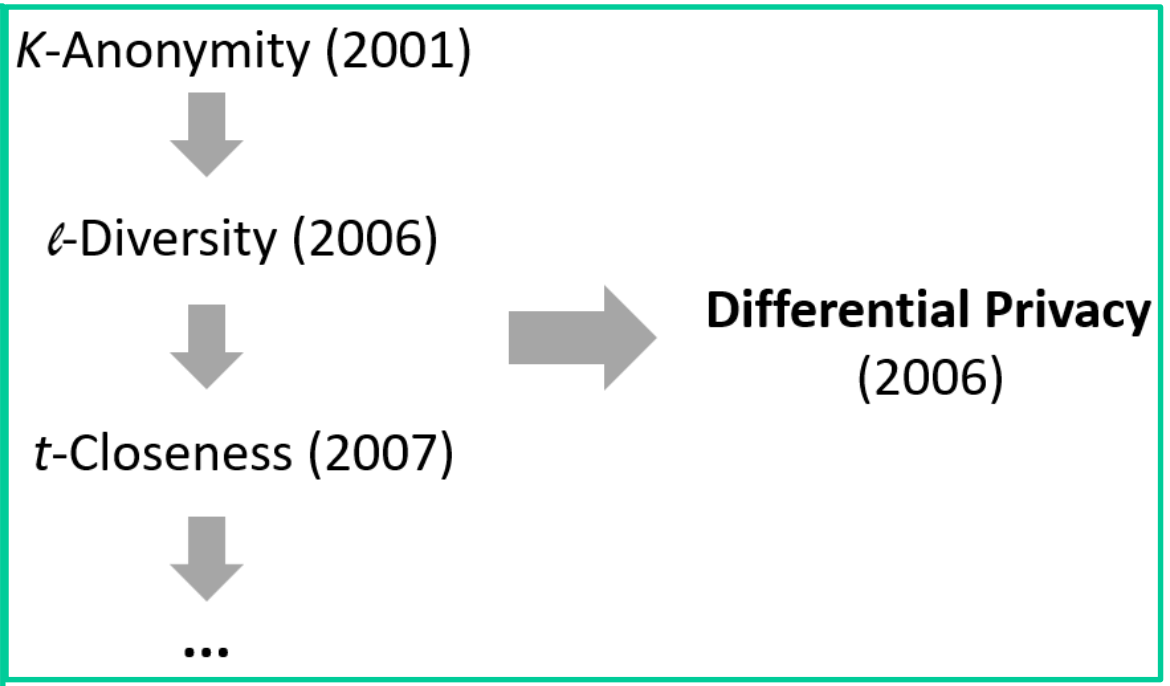
考虑一种情况:Ellen是一名30岁的女性舞蹈家,她的收入是多少?能否从表中推断出来?

攻击者如果知道Ellen是一名30岁的女性舞蹈演员，并且数据集中包含Ellen的数据，就能够推断出Ellen收入是1.99万或者2万；由于这个两个数值很接近，攻击者已经获得了想要的答案。

基于该问题，研究者提出了t-Closeness模型。这个模型需QID上任一群组中的敏感值的分布接近于整体表中的属性分布。

攻防技术演进

- 从k-匿名开始的一系列工作（t-closeness,l-diversity,等），陷入了一个“新隐私保护模型不断被提出但又不断被攻破”的循环中。
- 从根本上说，这一系列工作的缺陷在于为简化隐私保护理论上的推导，他们对攻击者的背景知识和攻击模型都给出了相当多的假定。但是那些假定在现实中并不完全成立，因此人们总能找到各种各样的方法来进行攻击。
- 直到差分隐私被提出后，这一问题才得到较好的解决。
  - 来自微软研究院的德沃柯(Dwork)等人于2006年提出了差分隐私模型



差分隐私

提纲

- 差分隐私模型简介
- 工作原理
- 应用与挑战

差分隐私模型简介

来自微软研究院的德沃柯(Dwork)等人于2006年提出了差分隐私模型。差分隐私是一种通用且具有坚实的数学理论支持的隐私保护框架，可以在攻击者掌握任意背景知识的情况下对发布数据提供隐私保护。

差分隐私具有两个最重要的优点：

(1)差分隐私严格定义了攻击者的背景知识：除了某一条记录，攻击者知晓原数据中的所有信息——这样的攻击者几乎是最强大的，而差分隐私在这种情况下依然能有效保护隐私信息。

(2)差分隐私拥有严谨的统计学模型，极大地方便了数学工具的使用以及定量分析和证明。

正是由于差分隐私的诸多优势，使其一出现便迅速取代了之前的隐私模型，成为隐私研究的核心，并引起理论计算机科学、数据库与数据挖掘、机器学习等多个领域的关注。

基本思想

当用户（也可能是潜藏的攻击者）向数据提供者提交一个查询请求时，如果数据提供者直接发布准确的查询结果，则可能导致隐私泄漏，因为用户可能会通过查询结果来反推出隐私信息。

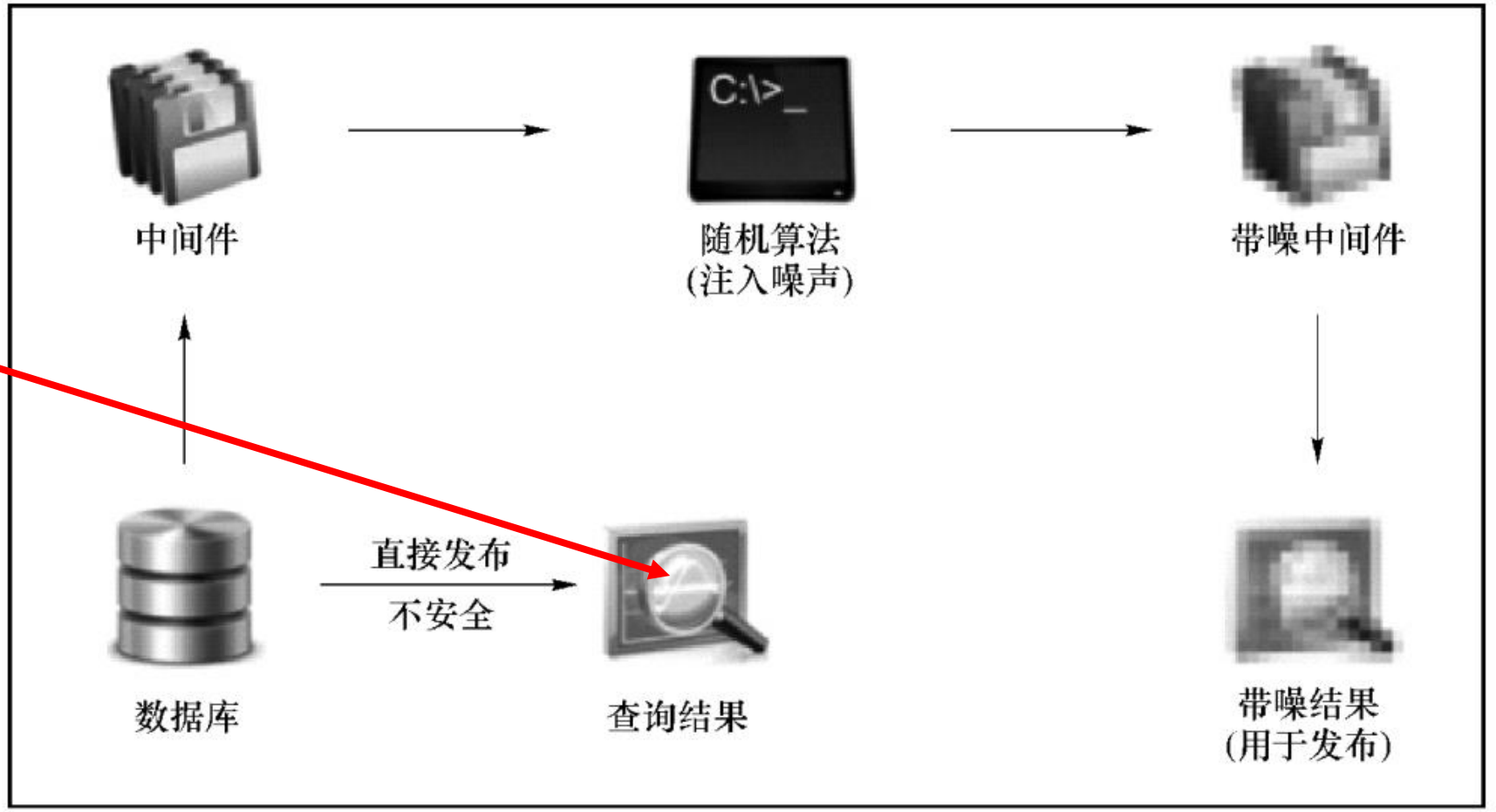
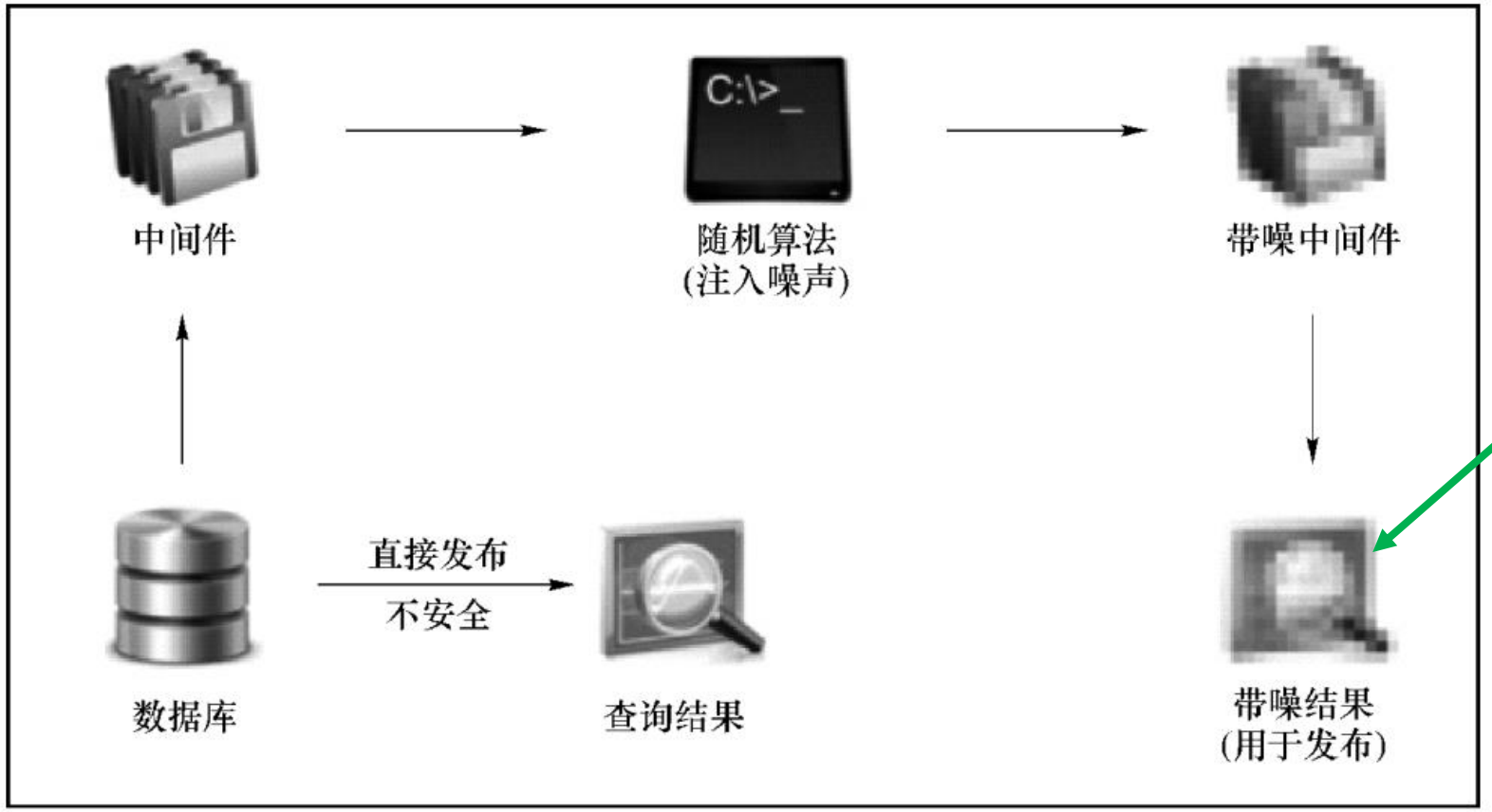


图 6-3　差分隐私技术的基本原理

基本思想



为了避免这一问题，差分隐私系统要求从数据库中提炼出一个中间件，用特别设计的随机算法对中间件注入适量的噪音，得到一个带噪中间件；再由带噪中间件推导出一个带噪的查询结果，并返回给用户。

图 6-3　差分隐私技术的基本原理

这样，即使攻击者能够从带噪的结果反推得到带噪中间件，他也不可能准确推断出无噪中间件，更不可能对原数据库进行推理，从而达到了保护隐私的目的。

定义

- 定义：对于任意一对相邻数据库（定义为差别最多有一个记录的两个数据库）D1和D2， 任意一个可能的带噪中间件S， 一个提供ε-差分隐私保护的算法A必须满足：
  - $Pr[A(D1)=S] \leq \exp(\varepsilon) \cdot Pr[A(D2)=S]$ 。
- 简单来说，定义一的要求是，即便攻击者已经知道了原数据中的绝大部分记录（即D1和D2的相同部分）以及带噪中间件（即S），他依然无法准确判断原数据到底是D1还是D2。
- 换言之，即便攻击者已经知道了原数据中的绝大部分元组，他依然无法对剩余的元组做出准确的推断。这是因为对于输入D1和D2，算法A输出S的概率是相近的。

统计学模型

- 示例
  - 对于任意一个可能的带噪中间件S，Pr[A(D1)=S] 和Pr[A(D2)=S]的比率总是被约束在[exp(-ε),exp(ε)]之间。
- 差分隐私的参数ε描述了上述两个概率分布的相似性。
  - ε越小，概率的相似性越高，也就越难区分D1和D2，从而达到更高程度的隐私保护。
  - 值得一提的是， 在实际应用中，选取多大的参数ε仍然是一个未决的问题。

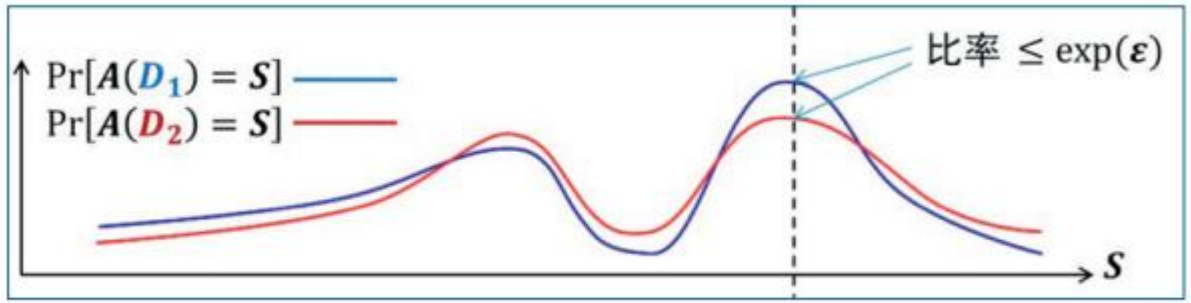


图2 差分隐私的统计学模型

核心算法及技术难点

- 差分隐私的核心在于其随机算法的设计：设计者首先需要证明算法输出的带噪中间件满足定义，然后在满足上述标准的情况下尽量少地加入噪音。
- 德沃柯最先提出了差分隐私的通用随机算法：拉普拉斯机制(Laplace mechanism)

德沃柯最先提出了差分隐私的通用随机算法：拉普拉斯机制（Laplace Mechanism）。

拉普拉斯机制的核心思想是通过向中间件加入拉普拉斯噪声来满足定义中的约束条件。具体来说，对于一个数据查询 F，拉普拉斯机制首先生成真实结果 $F(D)$ 作为中间件，然后通过发布带噪结果 $F(D)+\eta$ 来回答查询，其中噪声 η 服从拉普拉斯分布。

德沃柯等人证明了当 $\lambda \geqslant \Delta F/\varepsilon$ 时，拉普拉斯机制就能满足 ε-差分隐私机制，这样会使带噪结果中的噪声量过大。

此时需要运用各种技术手段来在保证隐私的同时进行降噪(如将带噪结果进行处理,或将F替换为一个结果与F相近但敏感度低的查询),以提高带噪结果的可用性。

核心算法及技术难点

- 麦克雪莉(McSherry)和图沃(Tulwar)所提出的指数机制(exponential mechanism)，也是差分隐私的经典通用算法。
  - 该机制与拉普拉斯机制最大的不同在于，后者适用于当数据查询的返回值为实数值的场合，而前者则适用于数据查询的范围值域为离散值域的场合。
- 现有的许多差分隐私算法在很大程度上都可以认为是拉普拉斯机制与指数机制的组合与应用。

应用与挑战

- 差分隐私极强的隐私保护能力和严谨的数学定义使其自诞生起就得到了广泛应用。例如，苹果公司就宣称采用了差分隐私技术保护用户的数据。
- 差分隐私中，我们假设攻击者拥有非常强大的背景信息，即知晓除了一条记录外的所有原数据，并在这样的假定攻击下保护数据隐私。
- 这样的假设其实是一把双刃剑：好的一面是，差分隐私提供了绝对的安全——即使如此强大的攻击者真的存在，差分隐私算法依然能够保护隐私；而坏的一面是，如此高强度的保护必然带来大量的噪音，影响带噪结果的可用性。

如果如此强大的攻击者在现实中并不存在，那么差分隐私就加入了过量的噪音，造成浪费。在我们对差分隐私的实际运用中，确实观察到了对于某些查询，过量的噪音导致结果完全不可用的现象。
所以在实际应用场景上，差分隐私是否过于严苛，是经常讨论的话题，也出现了一些改进差分隐私、 合理弱化假设的尝试。

小结

- 隐私的概念
- 数据匿名化
  - 发布-遗忘模型（Release-and-Forget Model）
  - 脱敏：1）识别身份信息；2）抑制；3）泛化；4）聚合
- 匿名化技术与反匿名化技术的军备竞赛
  - K匿名；L多样化；T相近
- 差分隐私技术
  - 基本原理
  - 核心算法及技术难点
  - 应用与挑战

阅读材料 – 数据匿名化

1) **L. Sweeney**. Guaranteeing anonymity when sharing medical data, the Datafly system. Proceedings, Journal of the American Medical Informatics Association. Washington, DC: Hanley & Belfus, Inc., 1997
2) Sweeney, Latanya. "Achieving k-anonymity privacy protection using generalization and suppression." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 05 (2002): 571-588.
3) Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. "l-diversity: Privacy beyond k-anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, no. 1 (2007): 3-es.
4) **Li, Ninghui**, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." In *2007 IEEE 23rd International Conference on Data Engineering*, pp. 106-115. IEEE, 2007.
5) **Arvind Narayanan** and Vitaly Shmatikov, **De-anonymizing Social Networks**, 30th IEEE Symposium on Security and Privacy, 2009
6) D. Kifer and A. Machanavajjhala. No free lunch in data privacy. SIGMOD, 2011.
7) X. He, A. Machanavajjhala and B. Ding. Blowfish privacy: tuning privacy-utility trade-offs using policies. SIGMOD, 2014.

阅读材料 – 差分隐私

1) Cynthia Dwork, Frank McSherry, Kobbi Nissim, and **Adam Smith**. "Calibrating noise to sensitivity in private data analysis." In *Theory of cryptography conference*, pp. 265-284. Springer, Berlin, Heidelberg, 2006.
2) F.McSherry and K.Talwar. Mechasim Design via Differential Privacy. Proceedings of the 48th Annual Symposium of Foundations of Computer Science, 2007.
3) Dwork, Cynthia. "Differential privacy: A survey of results." In *International conference on theory and applications of models of computation*, pp. 1-19. Springer, Berlin, Heidelberg, 2008.
4) Privacy integrated queries: an extensible platform for privacy-preserving data analysis by Frank D. McSherry. In Proceedings of the 35th SIGMOD International Conference on Management of Data (SIGMOD), 2009.
5) Dwork, Cynthia, and Aaron Roth. "**The algorithmic foundations of differential privacy**." *Foundations and Trends in Theoretical Computer Science* 9, no. 3-4 (2014): 211-407.

LDP

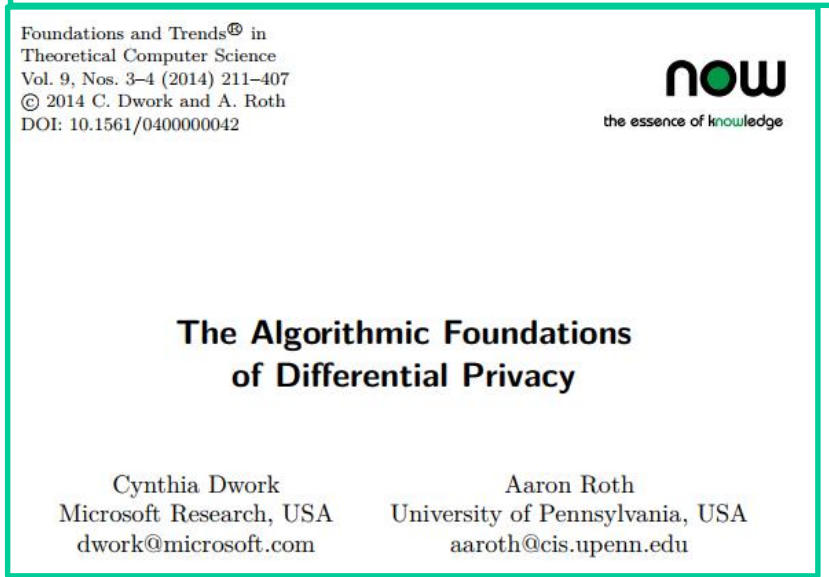
1) Raskhodnikova, Sofya, **Adam Smith**, Homin K. Lee, Kobbi Nissim, and Shiva Prasad Kasiviswanathan. "What can we learn privately." In *Proceedings of the 54th Annual Symposium on Foundations of Computer Science*, pp. 531-540. 2008.
2) Erlingsson, Úlfar, Vasyl Pihur, and Aleksandra Korolova. "Rappor: Randomized aggregatable privacy-preserving ordinal response." In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054-1067. 2014.
3) Bassily, Raef, and **Adam Smith**. "Local, private, efficient protocols for succinct histograms." In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 127-135. 2015.
4) 2017_Learning with Privacy at Scale_apple differential privacy system

In 2006, Cynthia Dwork, Frank McSherry, Kobbi Nissim and Adam D. Smith published an article formalizing the amount of noise that needed to be added and proposing a generalized mechanism for doing so.[1] Their work was a co-recipient of the 2016 TCC Test-of-Time Award[5] and the 2017 Gödel Prize.[6]

- Cynthia Dwork (born June 27, 1958)

2006年，Cynthia Dwork、Frank McSherry、Kobbi Nissim和Adam D. Smith发表了一篇文章，将需要添加的噪音量形式化，并提出了一个通用的机制。

In 2006, Cynthia Dwork, Frank McSherry, Kobbi Nissim and Adam D. Smith published an article formalizing the amount of noise that needed to be added and proposing a generalized mechanism for doing so.


Foundations and Trends® in Theoretical Computer Science
Vol. 9, Nos. 3–4 (2014) 211–407
© 2014 C. Dwork and A. Roth
DOI: 10.1561/0400000042

now
the essence of knowledge

The Algorithmic Foundations of Differential Privacy

Cynthia Dwork
Microsoft Research, USA
dwork@microsoft.com

Aaron Roth
University of Pennsylvania, USA
aaroth@cis.upenn.edu

北京郵電大學
Beijing University of Posts and Telecommunications

感谢聆听！

阅读材料 – 差分隐私

1) M. C. Tschantz, S. Sen and A. Datta, "SoK: Differential Privacy as a Causal Property," *2020 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2020, pp. 354-371.

差分隐私应用 –

- 匿名化方法是较为常用的隐私保护的方法。在对外发布的数据库中，机构主体通常会将姓名等敏感信息做匿名化处理。但是，在大数据时代，由于可以获取到外部数据库，通过比对和关联分析，可以推理出敏感信息，而造成隐私信息泄露。
- 因此匿名化方法往往无法提供良好的个人敏感信息保护。

匿名化方法：

敏感数据库（匿名保护）

姓名	出生地	年龄	喜好	存款
xxx	北京市	19	舞蹈	2w

非敏感数据库（无匿名保护）

姓名	性别	出生地	星座	年龄	喜好
小明	男	山东省	水瓶座	25	篮球
小红	女	北京市	天蝎座	19	舞蹈

在可以获取到外部数据库的情况下，匿名化方法往往无法提供良好的个人敏感信息保护

差分隐私应用 – 机器学习

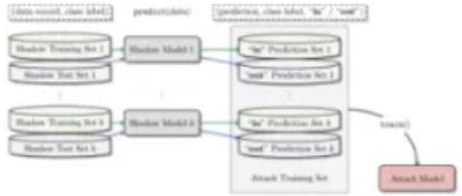
在机器学习领域，如果不发布数据，而只发布训练模型，个人隐私仍然不能得到有效的保障，例如，模型的逆向攻击或者成员推理攻击，都可以通过对模型的解析推理窃取原始数据。
因此，为了保护数据，我们需要寻找有数学保证的隐私保护方法。

隐私保护的挑战：模型隐私

在机器学习领域，如果不发布训练数据，而**只发布训练模型**，个人隐私会得到有效的**保证吗**?

Model Inversion Attacks [Matt Fredrikson et al. CCS 2015]

Membership Inference Attacks [Reza Shokri et al. S&P 2017]



需求：**有数学保证的隐私保护方法**

机器学习模型也会遭受多种攻击，导致敏感信息泄露

隐私保护的挑战: 差分隐私

- **更加严格的、更加数学化**的隐私保护方法：**差分隐私**

差分隐私从数学上给出了严格的证明，可以有效屏蔽包括成员推理攻击、属性推理攻击等攻击手段，因此，受到了Google、苹果、微软等科技公司的关注并被广泛应用。

差分隐私可以有效屏蔽诸如多种攻击手段:

- **Membership Inference Attack** [Michael Backes et al. (SIGSAC 2016)]
- **Attribute Inference Attack**[Nicholas Carlini et al. (USENIX 2019)]
- **Memorization Attack** [Bargav Jayaraman and David Evans (USENIX 2019)]

三种添加随机噪声的方式

✓ **有三种添加随机噪声的方式以保证模型的差分隐私性**

□ 输出扰动

$$\theta^* = \text{argmin}_\theta f(\theta; x, y), \quad \theta^{priv} = \theta^* + z$$

噪声

□ 目标函数扰动

$$f^{priv}(\theta; x, y; z) = f(\theta; x, y) + f(z),$$
$$\theta^{priv} = \text{argmin}_\theta f^{priv}(\theta; x, y; z)$$

噪声

□ 梯度扰动

$$\theta_{t+1}^{priv} = \theta_t^{priv} - \alpha_t \left(\nabla_\theta f\left(\theta_t^{priv}; S\right) + z_t \right)$$
$$\theta^{priv} = \theta_T^{priv}$$

噪声

（1）输出扰动：

输出扰动方法的基本思想是：在通过传统方法训练的机器学习模型参数上加入噪声，得到满足差分隐私定义的机器学习模型。

输出扰动方法的优点是：操作简单，原理清晰。缺点是：对输出模型的扰动可能会影响模型的性能，甚至会导致输出模型无法应用到新数据集。

（2）目标函数扰动：

目标函数扰动方法的基本思想是：在机器学习模型所需要优化的目标函数中加入噪声，得到满足差分隐私定义的机器学习模型 θpriv 。

目标函数扰动的缺点是：对目标函数的扰动可能会使模型不能收敛到最优，而影响模型的性能。

（3）梯度扰动：

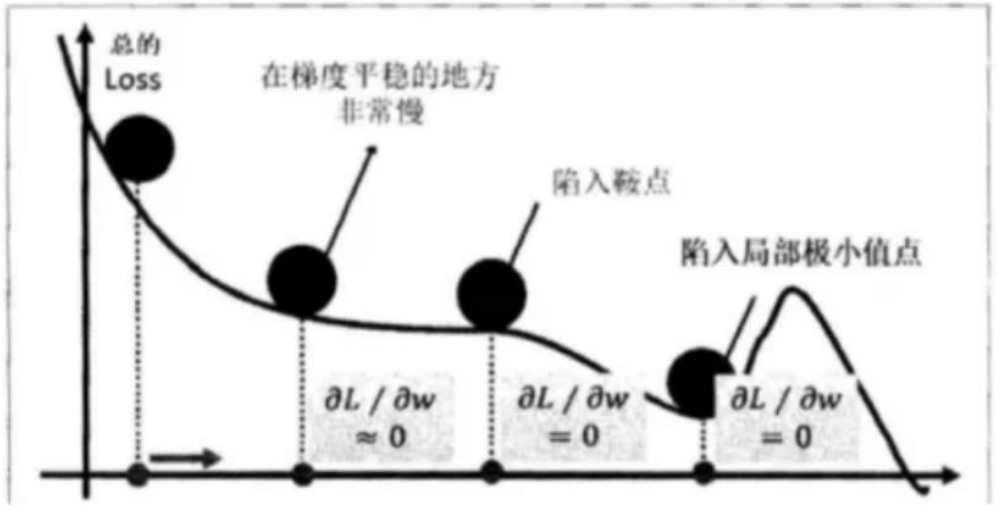
梯度扰动的基本思想是：对每一回合中目标函数的梯度加入噪声，得到满足差分隐私定义的机器学习模型。

梯度扰动通过在梯度加入噪声，因此还可以得到较优的训练模型，是最得到广泛应用的一种扰动方式。

随机噪声与模型性能

随机噪声vs干净数据

- ✓ 随机噪声是否一定导致性能下降？ 不一定！！！
  - □ 跳出鞍点/局部最小点，加速收敛



  - □ 增加模型鲁棒性（对抗性）

在模型中加入随机噪声并不一定导致性能下降。

1）如图所示，通过梯度扰动的方法，在目标函数的梯度加入噪声，将可能跳出鞍点/局部最小点，并加速收敛。

2）此外，对于满足差分隐私定义的机器学习模型中，由于已经加入噪声，将减少由数据样本扰动所引起的模型性能下降，即增加模型鲁棒性。

差分隐私：数据异质性差分隐私算法

在传统的差分隐私算法中，将所有训练数据等同视之，利用任意数据对模型进行训练时均添加同样的随机噪声。

但是在实际训练时，不同的数据点对模型的贡献是不同的。因此，对于一些贡献小的点，若不添加噪声，仍然可以得到满足差分隐私定义的机器学习模型。

基于此，我们对于传统的差分隐私进行改进，得到了数据异质性差分隐私算法。

经典差分隐私

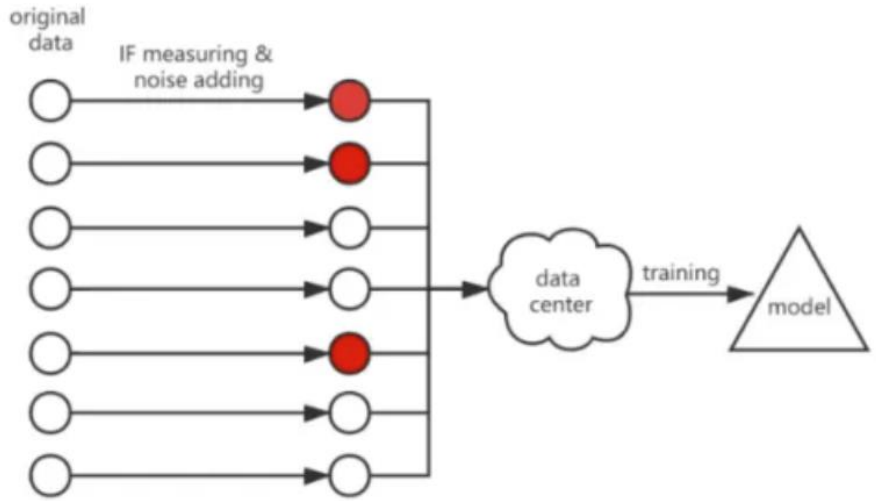
$$\theta_{t+1}^{priv} = \theta_t^{priv} - \alpha_t\left(\nabla_\theta f\left(\theta_t^{priv}; x, y\right) + \boldsymbol{z_t}\right)$$

$$\theta^{priv} = \theta_T^{priv}$$

- 传统算法将所有训练**数据等同视之**，利用任意数据对模型进行训练时**均添加同样的随机噪声**

数据异质性差分隐私算法

- 若某条**数据对模型输出的影响很小**，攻击者**本就无法分辨**该条数据是否参与训练，那么在利用该条数据训练模型时，就不必对其添加噪声
- 基于这种想法，提出**数据异质性差分隐私算法。**



在改进的数据异质性差分隐私算法中，若某条数据对模型输出的影响很小，攻击者本就无法分辨该条数据是否训练，那么在利用该条数据训练模型时，就不必对其添加噪声。

差分隐私：数据异质性差分隐私算法

差分隐私：数据异质性

- 相较于传统算法，该算法在梯度下降前先**对数据点对模型的影响**进行判断。
- **如果影响很小**，以至于攻击者无法从中得到有用的信息，那么则**不添加噪声**，以此减少模型训练中噪声添加的总量，进而提升模型精度

Yiling Kang, **Yong Liu**. Differential Privacy Based on Data Heterogeneity. 2022

Algorithm 2 PIDBDP

Input: dataset D, learning rate α, local iteration rounds T_{local}, global update rounds R

Get the pre-trained model $\widetilde{\theta}$ and calculate $H_{\widetilde{\theta}}^{-1}$.

Initialize $\theta_0^{(g)}, \theta_0^{(l)} \leftarrow \widetilde{\theta}, D_{priv} \leftarrow \{\}$.

for $r = 0$ **to** $R - 1$ **do**

  for $t = 0$ **to** $T_{local} - 1$ **do**

    Choose data instance z_t randomly.

    Calculate contribution built by z_t: $c_t^{(o)} = c_{z_t} + E$.

    $c_t = 2c_t^{(o)} + b^{(c)}$, where $b^{(c)} \sim \mathcal{N}(0, \sigma_{(c)}^2 I_m)$.

    if $e^{-\epsilon_1} \leq \| \frac{c_t + \theta_r^{(g)}}{\theta_r^{(g)}} \|_2 \leq e^{\epsilon_1}$ **then**

      $\theta_{t+1}^{(l)} \leftarrow \theta_t^{(l)} - \alpha \nabla_\theta \ell(z_t, \theta_t^{(l)})$.

    else

      Sample $b_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$, and train the model by $p(z_t) = z_t + b_t$, i.e.

      $\theta_{t+1}^{(l)} \leftarrow \theta_t^{(l)} - \alpha \left(\nabla_\theta \ell(p(z_t), \theta_t^{(l)}) - d_t + s_t \right)$.

    end if

  end for

  $\theta_{r+1}^{(g)} = \theta_{T_{local}}^{(l)}$.

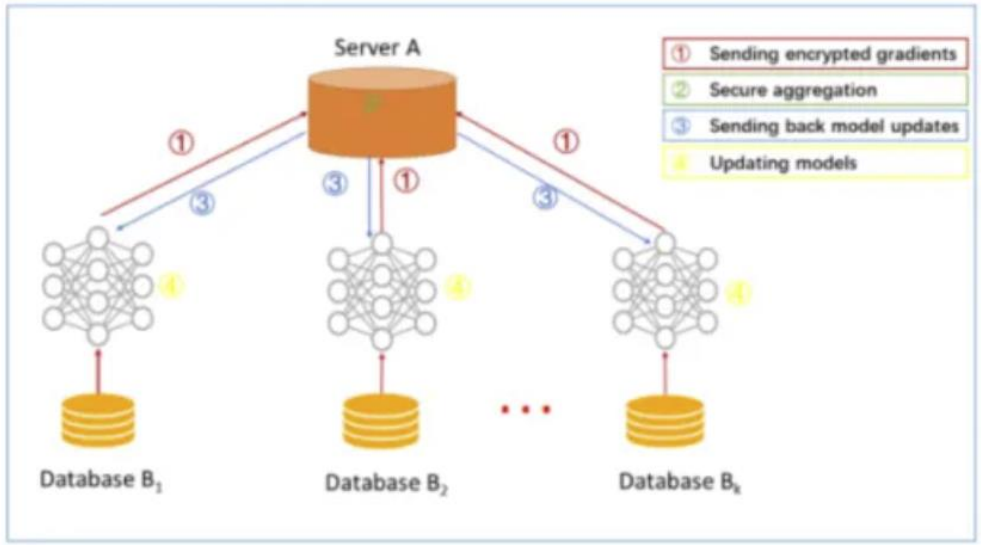
end for

Add all (last version) perturbed data instances to D_{priv}, if there is no perturbation, add the original one.

return $\theta_{priv} = \theta_R^{(g)}, D_{priv}$.

差分隐私应用 – 联邦学习

首先，差分隐私可以应用于联邦学习中。在实际的联邦学习模型中，模型参数可能多达百万甚至上亿，此时若使用同态加密等其他隐私保护算法，其实际的计算量和传输负载将难以承受，而应用差分隐私则可以解决这个问题。



- 数据不动模型动

相对分布式学习，联邦学习的难点：

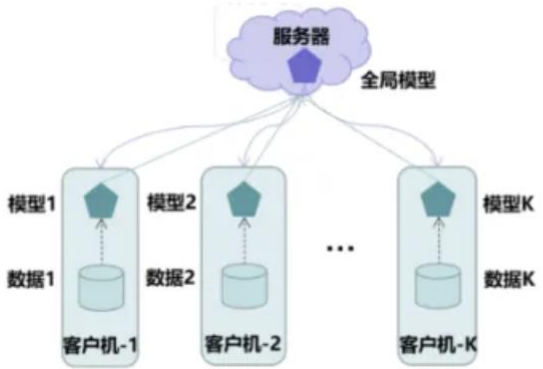
- 数据集非独立同分布
- 不平衡的数据量
- 慢速且不稳定的通信连接

联邦学习:非独立同分布

联邦学习：非独立同分布

基于隐私考虑，联邦学习使用模型交互代替数据交互，而由于用户的使用习惯不同等原因，联邦学习中各客户机的本地数据之间是**非独立同分布**（Non-IID）的：

- **数据分布不同**
- **数据量不相等**

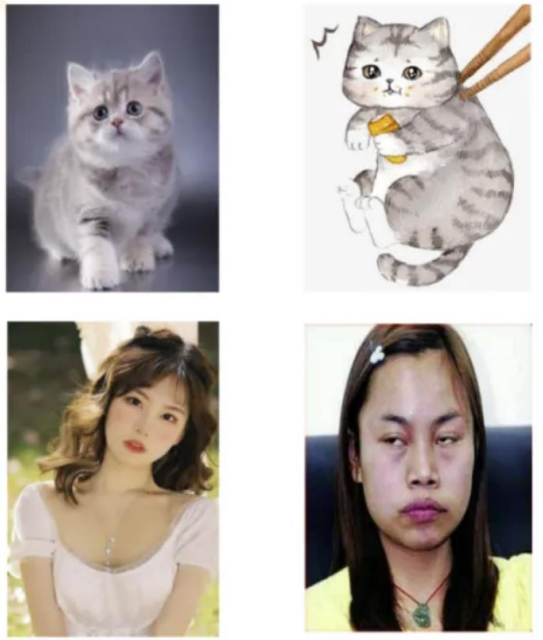


联邦学习的全局模型是本地模型的加权平均，所以Non-IID问题严重影响了全局模型的效果。

不同的本地数据分布 + 不平衡的本地数据量 → 本地模型之间差异较大 → 影响全局模型的效果

基于隐私考虑，联邦学习使用模型交互代替数据交互，而由于用户的使用习惯不同等原因，联邦学习中各客户机的本地数据之间是非独立同分布（Non-IID）的。
数据的非独立同分布可以分为：数据分布不同和数据量不相等。
而联邦学习的全局模型是本地模型的加权平均，所以Non-IID问题严重影响了全局模型的效果。

非独立同分布



在概率论与统计学中，独立同分布（Independent and identically distributed，IID）是指一组随机变量中每个变量的概率分布都相同，且这些随机变量互相独立。例如，如图中所示，左上图是一只猫的真实照片，但右上图是一只猫的卡通图片，这两只猫的图片是非独立同分布的。