第11章

推理控制

> 隐私是一个短暂的概念。当人们不再相信上帝能看到一切时，它就开始了，当政府意识到有一个空缺需要填补时，它就停止了。
>
> 罗杰·李约瑟

> “匿名数据”是那些圣杯之一，就像“健康的冰淇淋”或“选择性破解的加密货币”
>
> – CORY DOCTOROW

11.1 简介

正如大烟草公司花了几十年时间否认吸烟导致肺癌，大石油公司花了几十年时间否认气候变化，大数据也花了几十年时间假装敏感的个人数据可以很容易地“匿名化”，因此可以用作工业原材料不侵犯数据主体的隐私权。

匿名化是一个鼓舞人心的术语，意思是从数据中剥离识别信息，以便在不泄露有关可识别数据主体的信息的情况下进行有用的统计研究。它的局限性已经在四波研究中被探索过，每一次都是对当今技术的回应。第一波浪潮发生在 20 世纪 70 年代末和 80 年代初的美国人口普查中，其中包含的统计数据本身很敏感，但出于正当理由（例如向各州分配资金）需要总计；以及从大学分数到员工工资再到银行交易的其他结构化数据库的背景下。统计学家开始研究信息如何泄露，并制定推理控制措施。

第二次浪潮出现在 1990 年代，当时医疗记录已计算机化。卫生服务管理人员和医学研究人员都将此视为宝库，并希望删除患者的姓名和地址足以使数据成为非个人数据。由于数据的丰富性，这被证明是不够的，这导致了包括美国在内的几个国家的争吵。

美国、英国、德国和冰岛。此后，由于未充分匿名化的数据被泄露甚至被出售，出现了多起丑闻。

第三次浪潮发生在 2000 年代中期，当时人们意识到他们可以使用搜索引擎来识别大型消费者偏好数据集中的人，例如电影评级和搜索引擎日志。理论的进步出现在 2006 年，当时辛西娅·德沃克 (Cynthia Dwork) 及其同事开发了差异隐私理论，该理论量化了通过限制查询和添加噪音可以防止推理的程度，使我们能够在需要的地方添加噪音。

这现在被用于美国人口普查，他们的经验告诉我们很多关于它的实际限制。

第四次浪潮出现在 2010 年代后期，社交媒体、无处不在的基因组学以及通过电话应用程序收集并广泛出售给营销人员的个人位置历史的大型数据库。越来越多大规模出售个人信息的公司假装这些信息不是个人信息，因为名称以某种方式被标记化了。越来越多的新闻文章表明这种说法通常是多么虚假。

例如，2019 年 12 月，《纽约时报》报道称分析了 1200 万美国人在几个月内的手机位置历史记录，毫不费力地找到了名人、暴徒、警察、特勤局人员甚至色情行业的顾客 [1885]。

我们面临着使用匿名化和相关隐私技术可以完成的事情与从医学研究人员到营销人员再到政治家的利益相关者愿意相信的事情之间存在巨大差距。这种差距一直是很多讨论的主题，就像烟草和碳排放一样，也是政治争论的主题。随着我们对重新识别风险的了解变得越来越详细和确定，政府和行业的希望也越来越不切实际。政府反复征求建议书，数据用户呼吁承包商，创造无法创造的服务；十分之十，隐私服务合同是由更无知或不道德的运营商赢得的。

必须说，并非所有政府都只是无知。例如，英国和爱尔兰多年来允许公司假装数据是匿名的，而这些数据显然不是匿名的，这激怒了其他欧盟成员国多年，这是导致欧盟通过《通用数据保护条例》的因素之一 (GDPR)，正如我稍后将在第 26.6.1 节中讨论的那样。自它生效以来，一厢情愿的回旋余地变得越来越小 尽管即使是欧洲机构有时也对去识别化可以实现的目标抱有乐观的看法。

11.2 推理控制的早期历史

推理控制可以追溯到 1920 年代，当时经济数据的编制方式掩盖了个别公司的贡献，但它首先是在人口普查数据的背景下进行系统研究的。人口普查收集有关个人的许多敏感信息，包括姓名、地址、家庭关系、种族、就业、教育程度和收入，然后使地理和政府单位（如州、

县、区和区。这些信息用于确定选区,设定政府对公共服务的资助水平,并作为各种其他政策决定的输入。人口普查数据是一个很好的简单案例,因为数据采用标准格式,并且通常事先知道允许的查询。

有两种广泛的方法,这取决于数据是否在发布之前一劳永逸地进行了清理,或者隐私机制是否一次运行一个查询并确定它是否被允许。在数学上,两种类型的处理是相同的。对于受给定隐私约束的特定类型的数据,只允许一定数量的查询;问题是你是提前确定这些,还是动态响应用户需求。

第一种类型的一个例子来自直到 1960 年代的美国人口普查数据。
千分之一的记录在磁带上可用 减去姓名、确切地址和其他敏感数据。数据中还添加了噪音,以防止拥有一些额外知识(例如公司城镇雇主支付的薪水)的人追踪个人。除了样本记录外,还给出了各种属性的局部平均值。但是具有极端价值的记录 例如非常高的收入 被压制了。如果没有这种压制,生活在小村庄的富裕家庭可能会增加村庄的平均收入,足以扣除他们自己的家庭收入。

在第二种处理中,可识别的数据存储在数据库中,隐私保护来自限制可能进行的查询。例如,一个简单的规则可能是你不回答任何问题,除非结果是使用三个或更多数据主体的数据计算的 所谓的三规则。早期的尝试并不是很成功,因为人们不断地根据推理提出新的攻击。典型的攻击会构造一些关于包含目标个体的样本的查询,然后回过头来推断出一些机密事实。例如,你可能会问“告诉我收入在 50,000 美元到 55,000 美元之间的双人家庭的数量”,“告诉我收入在 50,000 美元到 55,000 美元之间的 40-45 岁男性为户主的家庭的比例”,“告诉我比例以年收入在 50,000 美元到 55,000 美元之间的男性为户主,其子女已经长大并离家出走的家庭数量”,依此类推,直到您找到目标个人。我们连续添加上下文以击败查询控件的查询称为跟踪器。

在许多情况下都会出现相关问题。例如,一名新西兰记者通过仔细检查军事和外交人员名单中随时间推移的发布模式,推断出该国信号情报服务 GCSB 中许多人员的身份 [849]。结合低级来源得出高级结论在国家安全背景下被称为聚合攻击。

11.2.1 推理控制的基本理论

推理控制的基本理论是由 Dorothy Denning 等人在 70 年代末和 80 年代初开发的,主要是为了应对美国人口普查的问题 [538]。这一波研究在 1989 年的一份调查报告中进行了总结,作者是

亚当和沃特曼 [17]。许多现代隐私系统的开发者往往没有意识到这项工作,并重复了 1960 年代的许多错误。以下是基本思想的概述。

特征公式是选择一组查询记录的表达式（在某些数据库查询语言中）。一个例子可能是“计算机实验室所有教授级别的女性雇员”。通过所有属性（或其否定)的逻辑与获得的最小查询集被称为基本集或单元格。如果集合大小太小,则与查询集对应的统计信息可能是敏感统计信息。推理控制的目标是防止泄露敏感统计数据。

如果我们让 D 是公开的统计数据集合,P 是敏感且必须保护的集合,那么我们需要 D ✓ P0来保护隐私,其中 P0是 P 的补集。如果 D = P0 ,则表示保护准确地说。

,

不精确的保护通常会在数据库可以回答的查询范围方面带来一些成本,因此可能会降低其实用性。

11.2.1.1 查询集大小控制

最简单的保护机制是指定最小查询集大小,这样如果计算答案的记录数小于某个阈值 t,则不会回答任何问题。但这还不够。假设 t = 6;那么一个明显的跟踪器攻击就是查询六个病人的记录,然后查询这些记录加上目标的记录。并且您还必须防止攻击者查询除一条记录以外的所有记录:如果有 N 条记录并且查询集大小阈值为 t,则在 t 和 N t 之间的记录必须是查询的主题才能被允许。这也适用于子集。例如,当我写这本书的第一版时,我们实验室只有一位正教授是女性。所以我们只需要两个查询就可以知道她的薪水：“教授的平均薪水？”和“男教授的平均工资？”。因此,如果 K L 和 |L|,则必须避免连续查询记录集 K 和 L。 |K| <吨。

11.2.1.2 跟踪器

这是一个单独的跟踪器的例子,一个自定义公式,允许我们间接计算禁止查询的答案。还有通用跟踪器 可以揭示任何敏感统计数据的公式集。

多萝西·丹宁 (Dorothy Denning)、彼得·丹宁 (Peter Denning) 和梅耶·施瓦茨 (Mayer Schwartz) 在 20 世纪 70 年代末发现了一个有点令人沮丧的发现,即一般的追踪器通常很容易找到。如果最小查询集大小 n 小于统计总数 N 的四分之一,并且对允许的查询类型没有进一步限制,那么我们可以找到提供通用跟踪器的公式 [541]。所以跟踪器攻击很容易,除非我们限制查询集大小或以其他方式控制允许的查询。这样的查询审计被证明是一个 NP 完全问题。

11.2.1.3 小区抑制

下一个问题是如何处理抑制敏感统计数据的副作用。例如,英国关于 2010 年人口普查的规则要求“任何统计单位在确定自己的身份后,都不可能使用该知识,通过推导来确定国家统计输出中的其他统计单位”[1416]。举一个简单具体的例子,假设一所大学要发布各种课程组合的平均分,以便人们可以检查跨课程的评分是否公平。假设现在图 11.1 中的表格包含学习两门科学科目的学生人数,一门是主修科目,一门是辅修科目。

主要的:	生物物理化学地质学			
次要的:				
生物学	-	16	17	11
物理	7	-	32	18
化学	33	41	-	2
地质学	9	13		-

图 11.1:包含细胞抑制前数据的表格

英国人口普查规则暗示最小查询集大小为 3,这在这里也很有意义:如果我们将其设置为 2,那么学习“地质与化学”的两个学生中的任何一个都可以计算出另一个的分数。所以我们不能发布“地质与化学”的平均值。但是,如果化学的平均分已知,则可以根据“生物学-化学”和“物理-化学”的平均分重建。所以我们必须在化学行中至少取消一个标记,出于类似的原因,我们需要在地质栏中取消一个标记。但是,如果我们压制“地质学与生物学”和“物理学与化学”,那么我们最好也压制“物理学与生物学”,以防止这些值被依次推导出来。我们的表现在看起来像图 11.2,其中“D”表示“为了披露目的而隐瞒的值”。

主要的:	生物物理化学地质学			
次要的:				
生物学	-	丁	17	18
物理	7	-	32	
化学	33	13	-	丁
地质学	9			-

图 11.2:细胞抑制后的表格

由于 Tore Dalenius,这个过程被称为互补细胞抑制。
如果数据库模式中有更多属性 例如,如果数字也按种族和性别细分,以表明遵守反歧视法 那么可能会丢失更多信息。在数据库方案包含 m 元组的情况下,消隐单个单元格通常意味着抑制2m 1 个其他单元格,这些单元格排列在一个超立方体中,其中一个顶点具有敏感统计信息。

因此,即使是精确的保护也会迅速使数据库无法使用。有时可以避免互补细胞抑制,例如当收入很高(或很少

疾病)在全国范围内制表,不包括在地方数据中。但是当我们发布微观统计数据时,它通常是必要的,如上面的考试成绩表。它可能仍然不够,除非我们可以在总数中添加噪音 因为机密数据的可能价值进一步受到我们披露的信息的限制,并且可能还有一些辅助信息,例如没有总数是负数的事实.

11.2.1.4 其他统计披露控制机制

另一种方法是 k-匿名,由 Pierangela Samarati 和 Latanya Sweeney 提出,这意味着其数据用于计算数据发布的每个人都无法与其他 k 人区分开来 [1795]。它的局限性在于它是隐私机制的操作定义,而不是隐私属性的数学定义;如果 k 个个体都具有相同的敏感属性,那也没有多大帮助。在数据库开放在线查询的情况下,我们可以使用隐含查询控制:只有当通过将 m 属性设置为 true 或 false 给出的2m个隐含查询集中的每一个至少有 k 条记录时,我们才允许对 m 属性值进行查询.另一种方法是限制查询的类型。最大顺序控制限制了任何查询可以拥有的属性数量。然而,为了有效,限制可能必须很严格。识别一个人只需要 33 位信息,而且大多数数据集的人口要少得多。一种更彻底的方法(在可行的情况下)是拒绝将样本总体划分为太多集合的查询。

我们在前一章中看到了如何在分区安全中使用格来定义分区之间允许的信息流的部分顺序与代码字的组合。它们还可以以稍微不同的方式用于在某些数据库中系统化查询控制。例如,如果我们有三个属性 A、B 和 C(比如居住地区、出生年份和医疗状况),我们可能会发现虽然对这些属性中的任何一个的查询都是非敏感的,但对 A 的查询也是如此和 B 以及 B 和 C,A 和 C 的组合可能是敏感的。因此,对所有这三个问题的调查也是不允许的。因此,格子自然地分为禁止查询的“上半部分”和允许查询的“下半部分”,如图 11.3 所示。

11.2.1.5 更复杂的查询控件

简单查询控制有许多替代方案。在 20 世纪后期,美国人口普查使用了“n 个受访者,k% 支配规则”:它不会发布一个统计数据,其中 k% 或更多是由 n 个或更少的值贡献的。

其他技术包括抑制具有极值的数据。人口普查可能会将高净值个人纳入国家统计数据,但不会纳入地方数据,而一些医学数据库也会对不太常见的疾病进行同样的处理。

例如,那个时期的英国处方统计系统抑制了当地统计数据中的艾滋病药物销售 [1249];即使在 20 世纪 90 年代初的艾滋病危机期间,有些县也只有一名患者接受了这种治疗。

一些系统试图绕过静态查询控制强加的限制

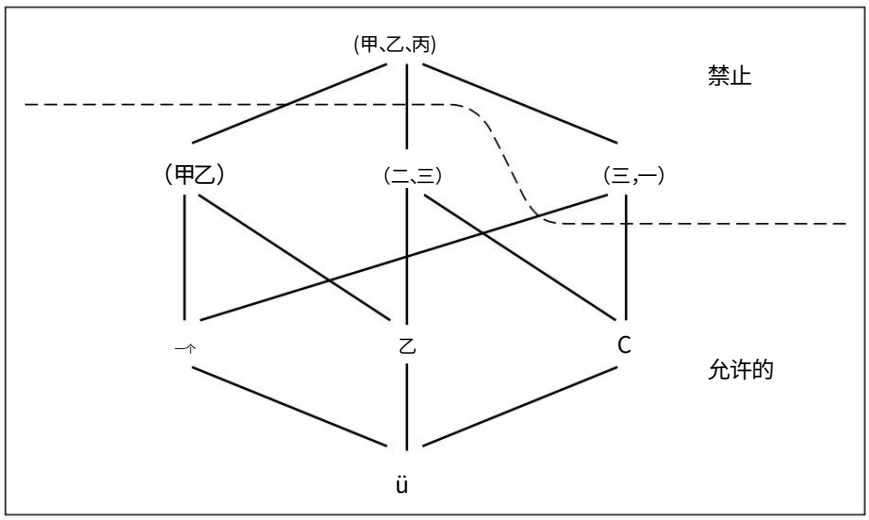

图 11.3：- 具有三个属性的数据库的表格

通过跟踪谁访问了什么。称为查询重叠控制,这涉及拒绝来自用户的任何查询,结合用户已经知道的内容,会泄露敏感统计信息。这听起来像是个好主意,但在实践中它有两个通常无法克服的缺点。

首先,所涉及的处理的复杂性会随着时间的推移而增加,而且通常呈指数级增长。其次,很难确定您的用户没有串通,或者一个用户以两个不同的名字注册。即使您的用户今天都是诚实和独特的人,明天他们中的一个人总是有可能被接管。

11.2.1.6 随机化

现在应该很清楚,如果各种查询控制是统计数据库中使用的唯一保护机制,它们通常会造成不可接受的统计性能损失。因此,查询控制通常与各种随机化结合使用,旨在从攻击者的角度降低信噪比,同时尽可能少地损害合法用户的信噪比。

直到 2006 年,所有使用的方法都相当临时。他们从扰动开始,或者向数据中添加零均值和已知方差的噪声;但这往往会在样本集较小时准确地损害合法用户的结果,而当样本集大到足以使用简单查询控件时,它们会保持原样。后来的变体是控制表格调整,您可以在其中识别敏感单元格并将它们的值替换为不同的值,然后调整表格中的其他值以恢复加法关系 [490]。然后是随机样本查询,我们使所有查询集大小相同,从可用的相关统计数据中随机选择它们。因此,所有发布的数据都是从小样本中计算出来的,而不是从整个数据库中计算出来的,我们可以使用一个伪随机数

生成器键入输入查询以使结果可重复。随机样本查询是一种自然保护机制,其中所调查的相关性足够强,小样本就足够了。最后,还有交换,这是 Tore Dalenius 的另一项创新;许多人口普查局交换了一定比例的记录,这样一个有两个十几岁的孩子且收入在第二个四分位数的家庭可能会被交换为下一个县镇上的一个类似家庭。

自 2006 年以来,我们有一个可靠的理论来说明我们可以从添加随机性中获得多少保护:差异隐私。这现在用于 2020 年美国人口普查,我们将在本章后面更详细地讨论它。

11.2.2 经典统计安全的限制

与任何保护技术一样,统计安全性只能在特定环境中针对特定威胁模型进行评估。是否足够取决于应用程序的细节。

一个例子是 1990 年代中期由一家名为 Source Informatics 的公司开发的系统,用于分析药物处方的趋势,该系统在英国关于匿名数据隐私的关键诉讼中有所体现 1。该系统的目标是通过按地区跟踪不同药品的销售情况,告诉制药公司他们的销售人员的效率如何。隐私目标是不泄露有关可识别患者的任何信息或个别医生的处方习惯 2。因此,从药店收集处方(减去患者姓名),然后进一步的去标识化阶段也删除了医生的身份。

周:1 4		2个	3个	
医生甲 17 26 19 22				
乙医生 25 31 9 29				
C 医生 32 30 39 27				
博士 D 16 19 18 13				

图 11.4:去标识化的药物处方数据样本

该系统的第一个版本只是用“医生 A”、“医生 B”等替换了四五个诊所中的医生姓名,如图 11.4 所示。

在评估它时,我们意识到警觉的药物代表可以根据处方模式识别医生:“嗯,B 医生一定是苏珊·琼斯,因为她在 1 月的第三周去滑雪,并在这里查看处方中的下降 。而 C 医生可能就是为她掩护的 Mervyn Smith。”解决办法是用每个医生开出的每种特定药物的百分比代替处方的绝对数量,随机放弃一些医生,并通过将数字向后或向前移动几周来随机扰乱时间 [1249]。

1完全披露:我是评估员,代表英国医学协会行事。
2医生一直被药品销售代表追捕,他们经常说他们会使用某种产品或其他产品只是为了让他们脱离手术。奇怪的是,如此重要的隐私案件的隐私目标是医生继续说善意谎言的能力。

这是经典统计安全技术可以提供稳健解决方案的系统类型的一个很好的例子。应用程序定义明确,数据库不太丰富,允许的查询相当简单,并且随着时间的推移保持稳定。即便如此,英国卫生部还是起诉了数据库运营商,声称该数据库可能会损害隐私。该部门的动机是维持对向行业提供此类数据的垄断。

他们输了,这建立了一个先例,即(至少在英国)推断安全控制可以,如果它们是稳健的,就隐私法而言,统计数据可以免于被视为“个人信息”[1804]。

不过,总的来说,这并不容易。首先,隐私机制并不组合:拥有两个独立的应用程序很容易,每个应用程序都通过相同数据的扰动版本提供相同的结果,但是可以访问这两个应用程序的攻击者可以轻松识别个人。这实际上发生在 Source Informatics 案例中;到 2015 年,出现了另一个使用不同机制的竞争系统,人们意识到可以访问这两个系统的制药公司偶尔可以推断出一些医生的处方行为。如果我们今天要重新实现这样一个系统,我们将通过使用差异隐私来防止这种情况发生,我将在本章稍后部分对此进行描述。

11.2.2.1 主动攻击

Source Informatics 系统每周都会添加一批新的记录,但有时用户可以将单个可识别记录插入数据库。在这种情况下,主动攻击会特别强大。 1990 年代后期的一个突出案例是冰岛的一个医学研究数据库。一家瑞士制药公司资助了当地一家初创公司,向雷克雅未克政府提供了一笔交易:如果您允许我们开采它用于研究,我们将为您建立一个现代医疗卡系统。政府签署了协议,但冰岛的医生大多反对这笔交易,认为这是对患者隐私和专业自主权的威胁。

根据他们提议的设计,每次生成医疗记录时,都会将其发送给冰岛隐私专员,后者的系统会删除患者的姓名和地址,用其社会安全号码的加密版本代替,并将其传递给研究数据库。隐私专员控制着加密密钥。然而,系统中任何想要查找(比如)总理医疗记录的人只需输入一些记录或其他记录 比如阿司匹林的处方 然后在一两秒钟后看着它在研究系统上弹出。冰岛政府无论如何都继续推进,耐心地选择退出。许多医生建议患者选择退出,11% 的人这样做了。最终,冰岛最高法院裁定,欧洲隐私法要求数据库可以选择加入而不是选择退出,这让该项目付出了代价。

冰岛对研究人员特别有吸引力,因为人口非常同质,是一千年前少数定居者的后裔,而且有很好的家谱记录。这也让冰岛语数据库中的隐私问题更加尖锐。通过将医疗记录与公开的家谱联系起来,可以通过诸如叔叔、阿姨、叔祖父、曾祖母等的数量等因素来识别患者 实际上是通过形状

他们的家谱。关于该设计在理论上是否甚至可以满足法律隐私要求的争论很多[66],欧洲隐私官员对欧洲隐私法体系可能产生的后果表示严重关切[515]。这将我们带到了丰富的上下文数据这一更广泛的问题上,它推动了推理控制的第二波工作。

11.2.3 丰富医疗数据中的推理控制

1990 年代后半期见证了“互联网繁荣”。万维网是新生事物,随着企业(和政府)试图弄清楚如何将其业务转移到网上,大量资金涌入科技领域。医疗保健 IT 人员在安全和隐私方面面临许多问题;记录已经从纸质转移到计算机,但现在所有的计算机都开始相互交谈 [63]。您可以使用电子邮件将测试结果从医院发送到医生的手术室,还是使用网络表格?您将如何加密它,谁来管理密钥?你能否通过删除姓名和地址来使完整的医疗记录足够安全以用于研究,而不是仅仅删除个人处方等情节数据?研究人员以前曾通过坐在医院图书馆阅读纸质记录来进行流行病学研究,如果你能在办公桌前进行这项工作“显然”会更好。然而,流行病学家通常希望能够将患者一生中的发作联系起来,这样他们就可以看到治疗和生活方式选择的长期影响。

匿名化要困难得多。

许多国家的卫生 IT 人员同时面临这个问题。新西兰建立了一个数据库,其中包含加密的患者姓名以及一条规则,即对于少于 6 条记录的查询不得回答,但意识到这还不够,并限制了少数经过特殊许可的医学统计学家的访问 [1422]。柏林墙的倒塌给德国带来了一个尖锐的问题,因为前东德的癌症登记处拥有一流的数据,这些数据对研究非常有用,但有患者姓名和丰富的背景数据,而这些现在属于西德严格的隐私法。注册表必须快速安装保护机制,这涉及去识别和严格的使用控制 [266]。在瑞士,一些研究系统也在隐私监管机构的坚持下被替换 [1681]。英国医学协会在 1995-6 年反对集中研究数据库的提议,并在著名精神病学家 Dame Fiona Caldicott 的领导下成立了一个委员会,以提出前进的方向。

1997 年,Latanya Sweeney 开始关注医疗记录的丰富背景改变了统计安全游戏这一事实,她在她的博士论文中尝试建立一个可以正确匿名医疗记录的系统,并发现它有多难是。她表明,即使是医疗保健金融管理局的“公共用途”文件,通常也可以通过将它们与商业数据库进行交叉关联来重新识别 [1849]。她展示了 69% 的美国居民可以通过出生日期和邮政编码来识别,并讨论了擦除包含各种上下文数据(包括自由格式文本)的医疗记录的极端困难 [1849]。当时,Medicare 系统将受益人加密的记录(患者姓名和社会安全号码加密)视为个人数据,因此只能由受益人使用

值得信赖的研究人员。还有公共访问记录,将标识符剥离到仅以一般术语识别患者的水平,例如“居住在佛蒙特州的 70-74 岁白人女性”。尽管如此,研究人员发现,通过将公共访问记录与商业数据库进行交叉关联,仍然可以识别出许多患者。斯威尼通过查明马萨诸塞州州长威廉·韦尔德的记录,引起了公众的注意。这使医学研究数据的匿名性被提上了美国的政治议程。

正如我在第 10.4 节中所述,克林顿政府于 2000 年根据 HIPAA 发布了隐私规则,为数据的公开共享定义了“安全港”标准,然后在 2002 年布什政府采用了更为宽松的规则。 2017 年,Sweeney 及其同事检查了一项 2006 年针对加利福尼亚州 50 个家庭的公共卫生研究,该研究在研究文献中被引用了数百次,并表明他们可以通过姓名和地址识别 28% 的参与者 [1850] .即使在将参与者的出生年份编辑为 10 年范围之后,他们仍然可以通过姓名和地址精确定位 3% 和 18% 因为住房类型等辅助信息。

英国遵循了类似的轨迹。 Dame Fiona Caldicott 的报告确定了医疗服务中的 60 多个非法信息流 [367]。一些研究数据集被非常粗心地去识别化;其他人(包括有关艾滋病毒/艾滋病患者的数据)后来被故意重新识别,因此在匿名承诺下收集数据的人和艾滋病慈善机构被欺骗了。议会随后通过了一项法律,赋予部长们监管医疗数据二次使用的权力,但大方向是受信任的研究人员;一个委员会审查数据访问申请。在某些情况下获得了患者同意,但不适用于涉及医院事件统计数据库的研究,该数据库包含从 1998 年至今英格兰和威尔士超过 10 亿次医院治疗的记录。 HES 数据可供研究人员使用,但患者的姓名和地址已被删除并替换为加密标识符。 (每个许可数据的研究组织的加密密钥都不同。)

但加密患者姓名是不够的。假设我想查找前首相托尼·布莱尔的记录。快速网络搜索显示,他于 2003 年 10 月 19 日和 2004 年 10 月 1 日因心律不齐在伦敦哈默史密斯医院接受治疗。这足以找出他的加密 ID 并查看他所做的一切。这样的泄漏对任何人来说都是侵入性的;对于名人来说,这可能具有新闻价值。更重要的是,在许多系统中都有明文邮政编码和出生日期;同样,这种组合足以识别大约 98% 的英国居民3。即使将出生日期替换为出生年份,如果记录很详细,或者如果不同个人的记录可以关联,我仍然可能会泄露患者隐私。例如,查询“显示所有 36 岁的女性及其女儿分别为 14 岁和 16 岁的记录,其中母亲和恰好一个女儿患有牛皮癣”可以从数百万中找到一个人。

查询集大小控制可能会阻止这种跟踪器,但研究人员确实希望进行具有很多条件的复杂查询,以找到具有

3UK 邮政编码比美国邮政编码具有更高的分辨率,每个邮政编码通常有 30 座建筑物。邮政编码加出生日期不唯一的 1% 左右的人大多是同卵双胞胎,或者住在大学宿舍或军营的年轻人。

几百甚至几十个病人。无论是有意还是无意,此类查询都可以以识别个人的方式进行编写。

2006 年,英国隐私团体组织了一场运动,提醒人们注意风险,并邀请他们行使选择退出二次数据使用的权利。
2007 年,议会的健康特别委员会对电子病历进行了调查,听取了来自广泛观点的证据 4并提出了许多建议,包括应允许患者阻止在研究中使用他们的数据 [925]。隐私问题并不是患者可能合理要求不使用其数据的唯一原因;例如,一位虔诚的天主教妇女可能会要求不得将她的数据用于开发堕胎或节育药丸。政府拒绝了这一点。

David Cameron s government, elected in 2010, weakened privacy protection, just as George Bush had done ten years earlier.正如我在第 10.4.4.3 节中详细讨论的那样,在谈论废除繁文缛节并使英国成为世界上医学研究的最佳地点时,他推出了“care.data”,这是一个可以添加测试结果的中央研究数据库,现有 HES 数据库中的处方和 GP 数据。 2013 年 11 月,HES 数据可通过 BT 在线销售 [948],2014 年 2 月,HES 数据库的副本已出售给全球 1,200 个组织,不仅包括学术研究人员,还包括商业公司,从制药公司到咨询公司 [774]。美国一家大型咨询公司已将所有 23GB 的数据上传到谷歌云,“因为它对 Excel 来说太大了”,并向客户提供这些数据,尽管法律要求数据必须留在英国。这些数据曾用于非健康目的,特别是被精算师用来优化保险费。很快通过了一项法律,规定只有在“对医疗保健有益”时才能共享和分析健康和社会数据,而绝不能用于其他目的。聘请了另一家咨询公司来制作另一份报告,并告知选择退出的人重新选择退出。一个学术案例研究讲述了这个故事,分析了医疗保健法和数据保护法之间的紧张关系,并指出“这场辩论的中心是保护和维护患者数据匿名的能力,没有简单的答案”[1548].

11.2.4 第三波:偏好和搜索

下一波浪潮在 2006 年爆发,那时大量交易已转移到网上,推荐系统因 eBay 和亚马逊而出现,搜索引擎使大海捞针变得容易。那一年发生的两起事件让公众知道了这一点。

首先,AOL 公布了 657,000 人在三个月内进行的 2000 万次搜索查询的所谓匿名记录。搜索者的姓名和 IP 地址被替换为数字,但这无济于事。调查记者查看了搜索结果并迅速确定了一些搜索者的身份,他们对隐私泄露事件感到震惊 [167]。数据是“出于研究目的”发布的:泄漏导致向联邦贸易委员会提出投诉,随后公司的首席技术官辞职,公司解雇了这名员工

4利益声明:我是委员会的特别顾问。

发布了数据和他们的主管。搜索历史,或者相当于您的点击流,是高度敏感的,因为它反映了您的想法和意图。

其次,Netflix 为更好的推荐算法提供了 100 万美元的奖金,并公布了 50 万订户的收视率,但去掉了他们的名字。
当时,它只有 600 万美国客户,并向他们运送实体 DVD,因此这只是其客户中的一小部分。 Arvind Narayanan 和 Vitaly Shmatikov 表明,通过将匿名记录与互联网电影数据库中公开表达的偏好进行比较,可以重新识别许多订户 [1384]。这部分是由于“长尾”效应 :一旦你忽略每个人都看的 100 部左右的电影,人们的观看偏好就会非常独特。由于美国法律保护电影租赁隐私,此次攻击对 Netflix 来说是一次严重的尴尬。

欧洲和加拿大的隐私监管机构的回应是推广隐私增强技术 (PET) 他们希望如果安全研究人员更加努力地工作,我们可以想出更有效的匿名化丰富数据的方法 [649]。 Microsoft 的研究人员相信他们的话,并发展了差异隐私理论,我在 11.3 中对此进行了解释。这并没有让隐私监管机构摆脱束缚,因为它阐明了匿名化的局限性。

然而,多年来,政策制定者将其作为一种解决方案进行讨论,却不了解它更详细地解释了为什么我们无法解决研究人员对详细数据的需求与数据主体的隐私权之间的紧张关系。

11.2.5 第四次浪潮:位置和社交

2010 年代,智能手机和社交网络改变了世界。
本书 2008 年第二版的第 23 章描述了早期的社交网络场景,当时 Facebook 刚刚从 Myspace 接管。我注意到罗伯特·普特曼 (Robert Putman) 的书《独自打保龄球》记录了随着 1960 年代电视的到来 [1563],通过教堂、俱乐部和社团等自愿协会的社会参与度下降,以及互联网早期的 Usenet 新闻组和邮件列表已设法将其中的一些恢复原状。社交网络的最佳点是将其推广给所有人。

无论您的兴趣如何深奥,您都可以与分享这些兴趣的人建立联系,无论他们身在何处。我们预测社交网络会直接带来各种隐私问题,因为社交背景很难隐藏。 (除了我之外,还有谁和密码学家、数字权利活动家以及对 200 年前的舞曲感兴趣的人一起出去玩过吗?)坚持会增加更多的危险,因为当青少年对性和毒品的吹嘘又回来困扰时他们后来在工作面试中。我们错过了两件事,一个是大量数据已经迁移到云端,另一个是可以从有关人员的上下文数据中推断出的大量敏感个人信息。到 2011 年,谷歌将其核心竞争力描述为“众包数据的统计数据挖掘”;随着数据集变得越来越大,并且基本统计技术随着机器学习而得到增强,我们可以学习的数量也在增加。

“更多数据”的一个示例是位置历史记录。到 2012 年,Yves-Alexandre de Montjoye 和他的同事已经表明,四个手机位置通常足以识别一个人,即使你只拿到他们的手机信号塔

位置 [1333]。如今,更高分辨率的数据被广泛使用,因为许多智能手机应用程序要求访问您的位置 这不仅涉及 GPS（室外平均精度可能为 8 米）,还涉及哪些 wifi 热点在范围内（这可以告诉您在建筑物中的位置）。大多数人毫不犹豫地点击同意,现在有一个完整的公司生态系统买卖位置跟踪数据 现在精确到几米而不是几百米。这些数据不仅卖给了营销公司,还卖给了私家侦探,包括赏金猎人,他们用它来追踪逃保的人[489]。

2019 年 12 月,《纽约时报》在几个月内掌握了 1200 万美国人的位置踪迹,并以图形方式展示了人们现在可以有多紧密地被追踪。您的日常轨迹显示了您的家、您何时离开、您如何前往工作、您在途中停下来喝咖啡的地方、您的办公室在哪里、您去哪里吃午饭 一切。记者在他们的数据库中发现了一位曾在教堂为特朗普总统唱歌的名人;数百名在五角大楼和中央情报局工作的人,以及总统的特勤局保镖,他们都可以跟着他们回家;和访问性行业的人。他们找到了一个在微软工作过的人,然后访问了亚马逊,然后在下个月开始在亚马逊工作。他们观察了一场骚乱,发现他们可以跟踪暴乱者和警察回家 [1885 年]。在公开市场上购买这些数据的容易程度与执法部门必须克服的障碍才能通过认股权证获得这些数据之间形成了鲜明的对比。位置数据公司都声称他们的数据是匿名的;然而,即使他们实际上可能不会使用电话簿或选民名册从您的街道地址中查找您的名字,一些人也会根据您浏览器中的一个或多个 cookie 出售与广告标识符相关联的位置数据。有了低分辨率的位置数据,当你去拉斯维加斯的黑帽大会时,在线赌博公司就可以在你面前投放广告。有了高分辨率数据,外国情报机构可以找到在五角大楼工作的人,也可以访问同性恋俱乐部或妓院。它还可以跟着他们回家。

“更好的推理”的一个例子来自社交网络数据的行为分析。此处的头条新闻始于 Michal Kosinski 及其同事编写了一款 Facebook 应用程序,该应用程序提供免费的心理测量测试,并说服数万人使用它。他们发现他们可以从 Facebook 的四个点赞中判断一个人是异性恋还是同性恋;给定 60 个赞,他们可以评估用户的“五大”人格特征:你是乐于体验还是谨慎、认真或随和、外向或内向、随和或超然、神经质或自信 [1086]。他们还可以判断您是白人还是黑人、保守派还是自由派、基督徒还是穆斯林、是否吸烟、是否饮酒、是否处于恋爱关系中以及是否吸毒 准确度各不相同。这导致他的一些同事以工业规模收集 Facebook 数据用于营销和政治竞选,导致了剑桥分析丑闻,我将在第 3 部分中讨论。后来的研究表明,拥有行为数据只能给出版商额外 4% 广告收入与他们通过上下文广告获得的收入相比,可以想象这种做法可能会被禁止 [1239]。

然而,行业观察家指出,平台赚得比这更多,因为它们获得了最大份额的广告收入 因此可以预期它们会抵制任何

这样的隐私法 [1181]。

在许多情况下,您可以获得位置数据和社交数据,并且可以大规模获取它们。例如,澳大利亚维多利亚州政府公开了一个交通票使用数据库,涵盖 2015-8 年间 1500 万张票的十亿次旅程。尽管卡 ID 已被匿名化,但居民通常只需要一两次旅行就可以通过触摸和触摸 次来识别自己的卡;研究人员发现他们可以识别他们的同行者 [502]。接下来,他们确定了使用澳大利亚联邦议会通行证的人,这些人经常乘坐火车前往他们的选区;假设可以从议员的推文中得到证实。该数据集使研究人员能够分析旅行时间的敏感性。他们发现,即使将旅行时间缩短到一天,去掉小时和分钟,四个地点也能识别出超过三分之一的旅行者。

我们现在有许多社交渠道和位置数据。位置历史泄露的数据如此之多,因为它揭示了我们与谁一起生活、一起工作和聚会。社交网络包含我们的联系方式、偏好和自拍,甚至可以使这些测量更加准确。社会分析可以深入到堆栈的最低层。例如,事实证明,匹配两个社交图谱相当容易,即使它们不是彼此的精确副本;所以给定一个国家的匿名手机通话数据记录,你可以通过将它们与(比如)社交网络的朋友图进行比较来重新识别它们 [1719]。手机数据已经泄露了很多关于我们性格的信息:外向的人打电话更多,随和的人接到的电话更多,电话之间的时间差异预示着尽责性 [1334]。

更多的数据和更好的推论相结合,也在医学研究中引发了新的争议。谷歌的人工智能子公司 DeepMind 于 2016 年宣布与伦敦一家医院合作开发一款诊断肾损伤的应用程序。第二年,事实证明,医院在未经同意的情况下,不仅向 DeepMind 提供了肾损伤患者的记录,还向 DeepMind 提供了所有患者的所有 160 万条完全可识别的记录 [1542]。隐私监管机构谴责了医院,因为这种访问权限只应授予参与直接患者护理而不是产品研究的公司;但是它并没有试图强制 DeepMind 删除数据。该公司使用来自美国的 VA 数据来开发诊断应用程序。它确实成立了一个声称将控制该技术的道德委员会,并承诺不会将医院数据提供给其母公司谷歌,但在 2017 年,道德委员会的一位知名成员辞职,声称这是在装点门面,并在 2018 年宣布谷歌正在吸收 DeepMind 的健康业务 [909]。在这起缓慢的火车事故之后,有消息称谷歌已经因获取 5000 万美国患者的记录而受到抨击 [121]。

那么是否可以正确地进行匿名化?答案是肯定的;在某些情况下,它是。虽然无法创建可用于回答任何问题的匿名数据集,但有时当我们着手回答一组特定的研究问题时,我们可以提供可靠的隐私措施。

这将我们带到了差异隐私理论。

11.3 差分隐私

2006 年,Cynthia Dwork、Frank McSherry、Kobbi Nissim 和 Adam Smith 发表了一篇开创性论文,展示了如何系统地分析增加噪音以防止泄露数据库中敏感统计数据的隐私系统 [595]。他们的理论,差分隐私,使安全工程师能够限制泄露的可能性,即使存在具有无限计算能力和丰富边信息的对手,因此可以被视为等同于一次一密和密码学中的无条件安全认证码。尽管它最初是一篇关于理论密码学的论文,但它已被视为统计数据库安全性和一般匿名化的黄金标准。起点是 Kobbi Nissim 和 Irit Dinur 较早的一篇论文,他们在 2003 年表明,如果对数据库的查询每个都返回私有信息位的线性函数的近似值,那么只要误差足够小,num 重建数据库所需的查询数量不会增长太快;毕竟,这种重建攻击是基于线性代数的,因此攻击者无需进行精心设计的跟踪器攻击,而是可以进行大量随机查询,然后进行代数计算并找出所有内容 [562]。因此,如果查询数量超过有限数量,则防御者必须添加噪音,问题是有多少。

差异隐私的关键见解是,为了避免无意中泄露,任何人对查询结果的贡献都不应该造成太大的差异,所以你根据数据的敏感度来校准噪声的标准差.如果对于所有数据库 X 和X0在一行中不同,则隐私机制称为 -不可区分,从 X 获得任何答案的概率在从X0获得答案的 (1 +) 的倍数范围内;换句话说,您限制了比率的对数。因此,您可以使用具有拉普拉斯分布的噪声来获得与噪声总和的不可区分性,并且事物组合在一起,因此这一切在数学上变得易于处理。 的值设置了准确性和隐私之间的权衡 ,必须由策略设置。

小值提供强隐私;但设置 = 1000 基本上是发布您的原始数据。

现在有越来越多的研究文献探索如何将这种机制扩展到静态数据库、动态数据库、数据流、机制设计和机器学习。但是,在实际应用中能否实现在学习有关群体的有用信息的同时对个人一无所获的承诺?

11.3.1 将差异隐私应用于普查

差异隐私现在正在 2020 年美国人口普查中得到全面检验。
人口普查不得发布任何可识别任何个人或机构数据的信息;根据法律,收集的数据必须保密 72 年,并且在此之前仅用于统计目的。首先,人口普查局根据现代分析工具审查了 2010 年人口普查的安全性 [752]。 2010 年,汇总的人口普查编辑文件 (CEF) 从美国居民那里收集数据,然后进行编辑以去除重复项并填写缺失项

来自纳税申报表等数据的条目,每个居民都有 44 位机密数据（总共 1.7Gb）。问题是微数据摘要包含的数据比这多得多;把所有东西都写出来,你会得到几十亿个联立方程,理论上可以解决机密数据。

在实践中呢?人口普查人员根据 Kobbi Nissim 和 Irit Dinur 的工作实施了想法,发现他们在大约 38% 的时间里正确地获得了所有变量,覆盖了略低于 20% 的人口。在四台服务器上花了一个月的时间,所以这并不完全是微不足道的。然而,教训是传统的统计数据库安全方法并不真正有效。他们确实提供了一些隐私,因为 2010 年的人口普查将非常可识别的家庭与其他街区交换了,所以并不是每个人都受到了损害。如果他们把所有的家庭都换了,那还可以,但是用户不会忍受的;他们给出了一个区块的确切人口数量这一事实是一个真正的漏洞。零碎地处理数据库重建是困难的;这就是差异隐私的价值。

最大的政策问题是你在哪里设置 。这也是一个实证问题。
2018 年,人口普查人员做了一个端到端的测试,报告了四个表格。到 2020 年,整个系统会将 CEF 处理成一个微数据详细信息文件 (MDF),从中可以得出表格。可预见的问题包括数字不会加起来;所以独立的美洲原住民部落的成员人数加起来不会等于美洲原住民的总数,这必须向公众解释。差异隐私方法将保护每个人,而旧系统只保护被交换的人,而且必须一次性完成。每条记录都可能根据总体隐私预算进行修改,因此 CEF 和 MDF 之间没有确切的映射。

新的自上而下算法生成一个没有地理标识符的国家直方图,然后开始自上而下构建一个地理直方图,这样各州的数字就与全国的数字相加（这是国会重新划分选区所需要的）。然后通过州、县、地区、街区组和街区递归地进行构建,之后生成微观数据。这可以并行完成并实现稀疏性发现（例如,很少有超过 100 岁的人属于 5 个或更多种族）。事实证明,自上而下的方法比逐块应用噪声更准确,因为该县数据的误差小于块,而国家数据基本上没有误差。有几个边缘案例需要特殊处理:监狱不会变成大学宿舍,但如果有五个宿舍,你可能会报告四六个。人户联结也很难;您可以计算一个街区的男性人数或家庭人数,但以单身男性为户主的家庭中的孩子人数更为敏感。但是很多以前被压抑的东西,现在已经不需要了;您不再需要列举所有可能使用的辅助信息来源;最后将公布错误统计数据。

现在轮廓设计已经完成,您可以使用一个模拟器来探索 的可能值。您可以将其插入更好统计数据的边际社会效益与身份盗用的边际社会成本之间权衡的经济分析中[928];结果表明 的值在 4 到 6 之间。

11.4 注意差距?

在政治方面,多年来,无论是医学研究还是市场研究,在研究中使用轻微去识别化的数据都涉及隐私倡导者和数据用户之间的零星游击战,监管机构通常站在数据用户一边,除非在丑闻。正如我将在第 26.6.1 节中描述的那样,监管者既不知所措又矛盾重重,并且大多不具备与大型互联网服务公司或政府部门较量的政治支持。不管怎样,这些“大数据”利益通常都善于抓住监管者。例如,2008 年,英国首相戈登·布朗要求英国信息专员和英国最大的医学研究慈善机构负责人制定研究数据使用指南;他们无视隐私权,对成本和收益采取工具性观点,并将数据的二次使用称为“数据共享”。正如您所料,隐私律师和安全学者都不满意结果 [96]。

2009 年,美国著名法学教授保罗·欧姆 (Paul Ohm) 撰写了一篇极具影响力的论文“隐私的违背承诺”[1465]。他指出,“科学家们已经证明,他们通常可以非常轻松地‘重新识别’或‘去匿名化’隐藏在匿名数据中的个人”,并承认“我们犯了一个错误,在根本性的误解下努力,这让我们的隐私比以前少得多”我们假设。这个错误几乎遍及所有信息隐私法律、法规和辩论,但监管机构和法律学者却很少关注它。”在过去的 30 年里,计算机科学家们知道匿名化并没有真正起作用,但法律和政策人员却对此置若罔闻。终于,一位杰出的律师在法律期刊上使用律师通俗易懂的语言阐明了事实,讲述了 AOL 和 Netflix 的故事。除其他外,他还嘲笑谷歌声称 IP 地址不是个人信息的说法(它认为其搜索日志因此不属于数据保护范围),谴责数据是否属于个人的二元思维,并呼吁采取更现实的做法关于隐私和数据保护的辩论。这可能会改变一切吗?

2012 年,皇家学会的一份报告呼吁科学家在可能的情况下公开发布他们的数据,但承认重新识别风险的现实:“然而,计算机科学的大量工作现在已经证明,个人记录的安全性无法通过主动寻求身份的匿名程序来保证数据库中的信息 [1627]。

同年,英国信息专员还制定了匿名化操作规范 [80];由于 ICO 是隐私监管机构,这样的代码可以保护公司免受责任,并且它是大力游说的目标。最终的代码要求数据用户仅以一般术语描述他们的机制,并将举证责任转移给任何反对者 [81]。这比 ICO 在信息自由案件中适用的负担要轻,在信息自由案件中,可以拒绝对公共数据的请求,前提是数据主体的“朋友、前同事或熟人”可能知道相关背景。这涉及到一个与战术匿名相关的概念

> 隐私集,或者我可能不想知道关于我的一些事实的一组人。对于大多数人来说,这是您的家人、朋友和同事可能有 100 到 200 人。对于名人,可以是每个人;问题可以

当某人突然成名时出现。我们大多数人在大城市中都可以匿名,但名人却不能。

另一个有用但完全不同的概念是匿名集,它是您可能会混淆的人的集合。我们都熟悉侦探电影或小说,其中波洛稳步将可能犯下谋杀罪的人数从十几人减少到一人。差异隐私等战略机制侧重于保持匿名集足够大,而许多战术机制评估有权访问某些应用程序的人与您的隐私集重叠的风险。

但是您始终必须仔细考虑威胁模型。虽然当担心的是尴尬时,担心您的隐私集可能就足够了,但当它是骗子时,您需要担心匿名集。正如我们在第 3 章中提到的,网络钓鱼攻击通常涉及有关受害者的信息泄露,这使攻击者能够冒充受害者使用某些服务,或冒充受害者的服务。简而言之,在网络钓鱼方面,任何可以将您的身份与某些相关上下文联系起来的人都可以攻击您。

11.4.1 战术匿名及其问题

ICO 还建立了英国匿名网络 (UKAN),该网络由学术界和国家统计局协调。 2016 年,UKAN 出版了一本关于公司应如何做出匿名决定的指南,并由 ICO [626] 正式签署。它的作者将机密性视为风险而非责任;不仅要根据识别数据主体的技术可能性,还要根据决定是否可以尝试这样做的制度和社会背景来做出决定。威胁模型应该基于似是而非的入侵场景。他们谈论的是治理过程而不是旁路;他们认为差异隐私是“极端的”;他们将匿名化视为一个过程,并建议不要使用像“匿名化”这样的“成功术语”;他们将“去识别化”定义为“无法直接从数据中重新识别”。管理重新识别风险的措施应与风险及其可能的影响成比例;由于最终来自其他数据集的三角测量,匿名化措施的寿命可能有限。因此,此类机制必须被视为战术匿名,而不是美国人口普查中精心设计的战略匿名。 UKAN 作者似乎没有认真考虑差异隐私。

尽管存在缺陷,但如果您要在英国依赖匿名化(无论是战术上还是战略上),则需要注意 UKAN 框架,因为它是监管机构决定是否对您采取执法行动的准绳。随着匿名化逐渐变得不那么有效,它可能会为数据用户和监管机构提供减震器和责任保护。

它本来可以为在英国开展欧盟业务的公司提供一些保护,但随着英国离开欧盟,这将不再适用。然而,它确实包含合理数量的实用建议,用于评估在数据和环境都得到合理理解的应用程序中战术匿名化的风险。因此,现在有几家公司的产品和服务旨在帮助数据用户遵守它。

在此框架下公开运营的公司的一个例子是移动网络运营商沃达丰,该公司销售“位置洞察”产品。该公司将其客户的手机位置汇总到包含隐含的出发地、目的地和交通方式的旅程中。起点-终点矩阵连同沿主要公路和铁路的流量出售给地方政府和运输公司。隐私机制包括:首先,允许所有订户选择退出;其次,加密电话IMSI,为每个设备提供不同的假名,并使用缓慢变化的密钥;手机信号塔很容易重新识别。确实可以说这里的风险很低;也许当地议会或公共汽车公司的分析师可以识别您,特别是如果您住在一个小村庄(就像我一样;距离最近的村庄 200 米有四栋房子)。

所以匿名集可能太小了。然后你必须看看隐私集的大小。但是假设你在一家成为维权人士攻击目标的公司工作。如果他们在委员会招募人员,他们可能会以住在孤立房屋中的公司员工为目标,以恐吓他们或他们的家人5。

已经变得明显的实际问题首先与规模有关,其次与自我监管的内在冲突有关。规模不仅体现在可能与外部匹配以识别人员的数据源数量上,还体现在组织内部数据仓库的规模和复杂性不断增长上。 Hadoop6是一个决定性因素:公司现在可以存储所有内容,因此很难跟踪存储的内容。由于没有数据库模式,但数据只是堆积起来,你不知道链接风险,特别是如果你的公司有一个多租户集群,其中有来自不同子公司的各种学生。此类数据仓库现在用于欺诈预防、客户分析和有针对性的营销。公司想要负责任,但您如何向您的开发和测试团队提供实时数据?您如何与学者和初创公司合作?如何销售数据产品?

匿名化技术在这个规模上都非常初级,因为你不知道发生了什么,它超出了差异隐私或任何其他你可以清楚地分析的范围。您可以对摄取的数据进行标记以去除明显的名称,然后控制访问并对时间序列和位置流使用特殊技巧,但噪声添加对轨迹不起作用,并且有很多创造性的方法可以重新识别位置数据(例如名人进出出租车的照片)。在人们获得部分授权和部分访问权限的情况下,事情会变得更加困难。

未来的问题可能来自人工智能和机器学习;这是现在的时尚,追随 2010 年代中期导致公司建立大型数据仓库的“大数据”时尚。你现在正在训练通常无法解释它们所做的事情的系统,使用你并不真正理解的数据。我们已经知道很多事情都可能出错。保险系统提高了少数民族社区的保费,违反了反歧视法。机器学习系统在训练数据的同时吸收了现有的社会偏见;随着机器翻译系统读取千兆字节的在线文本,它们变得越来越

52003 年,我被选为我们大学管理机构的成员,在大学提议建造一座用于医学研究的动物新大楼后,我们成为了动物权利活动家的目标。一些同事让激进分子出现在他们的家中对他们大喊大叫,一些激进分子后来在牛津进行了类似的活动后被判犯有恐怖主义罪。几乎任何人都可能突然成为目标。

6 最初由 Yahoo 开发的开源软件,用于以 PB 级规模存储数据
服务器集群并使用 NoSQL 访问它。

更擅长翻译,但他们也变得种族主义、性别歧视和恐同（我们将在 25.3 节中更详细地讨论这个问题。另一个问题是,如果神经网络接受个人数据训练,那么它通常能够识别这些人如果它再次遇到它们 所以你不能只是训练它然后发布它,希望它的知识在某种程度上是匿名的,因为我们可能希望从大量数据中得出平均值。同样,你只是不明白ML 系统在做什么,所以你对匿名所做的任何声明都应该受到怀疑。仅仅说“我们不出售你的数据,我们只针对广告”是不够的:如果你让伊朗秘密警察针对广告在说波斯语的同性恋者那里,他们可以简单地弹出提供免费披萨的广告 。

由于信息专员办公室似乎没有能力或动力来监管匿名服务和应用程序,因此该行业自行监管;实际上,公司会标记自己的作业。这意味着逆向选择,因为最不认真的提供者将承诺最多的功能。正如我已经指出的,有许多公司出售细粒度的位置数据、社交数据等,他们声称它是匿名的,即使它显然不是。

即使组织是善意的,他们也很少在遇到麻烦之前真正理解问题,而且不止一次,我们有供应商在他们忍无可忍后向我们寻求建议.一旦遇到问题,数据用户通常不想与真正的专家交谈,因为他们意识到他们知道的越多,解决问题的成本就越高。至于加强监管,政府做得越多,其信息产业的竞争力就越弱。匿名化是一个如此棘手的问题的原因之一是它的安全经济学真的很糟糕。

11.4.2 奖励

即使是不完美的去识别化也可以保护数据免遭随意浏览和某些不安全甚至掠夺性的使用。然而,它可能会让流氓觉得自己有权做流氓事（尤其是自 UKAN 以来）。因此,在统计安全方面,是否应该让最好的成为好的敌人的问题可能需要比其他地方更精细的判断。正如我在经济学一章中所讨论的那样,在具有许多利益相关者的大型系统中,安全失败的最常见原因是激励错误 Alice 保护系统而 Bob 为失败付出代价。那么这里的激励措施是什么?

整体画面不好。例如,医疗隐私取决于人们如何支付医疗费用。如果你私下看心理分析师并支付现金,那么激励是一致的;分析师将锁定您的笔记。但在美国,医疗保健通常由您的雇主支付;在英国,政府支付了大部分费用。在这两种情况下,出于管理目的而集中控制的尝试都引发了医生和患者之间的冲突。虽然此类冲突可以通过匿名声明暂时掩盖,但不太可能通过任何可行的隐私技术来解决。一旦人们接受了这一点,就可以开始更现实的政治对话。

11.4.3 备选方案

一种方法是将弱匿名与访问控制相结合,无论是要求研究人员访问安全站点（如在新西兰,以及英国的税务数据研究)还是要求许可并入保密协议以及访问和使用控制禁止任何尝试识别主题（如在德国）。如果这样做的话,这可以是健壮的:

1. 胜任,具有良好的安全工程;
2. 诚实,没有虚假声明数据不再是个人的;和
3. 在法律范围内,这在欧盟将涉及赋予数据主体权利
   选择退出是受尊重的。

在医学中,黄金标准是在患者明确同意的情况下进行研究。这不仅允许完全访问数据,而且提供了积极的受试者和质量更高的临床信息,而不是仅仅作为正常临床活动的副产品。例如,肌萎缩侧索硬化（剑桥天文学家史蒂芬霍金所患的运动神经元疾病)的研究人员网络在患者及其家人的完全同意下,在十几个国家的医生和其他研究人员之间共享完全可识别的信息家庭。该网络允许在拥有非常严格的隐私法的德国和几乎没有隐私法的日本之间共享数据;即使在美国空军轰炸塞尔维亚时,美国和塞尔维亚的研究人员仍在继续共享数据。同意模型正在传播。第二个例子是英国的一个研究项目 Biobank,在该项目中,数十万名志愿者不仅让研究人员可以完全访问他们余生的记录,还回答了广泛的问卷调查并提供了血液样本,以便那些后来患上有趣疾病的人生命可以分析其遗传和蛋白质组学构成。不用说,完全同意的访问还需要强大的安全工程,因为同意将取决于仅限研究人员访问。

无论您走的是受信任的研究人员路线还是完全同意的路线,研究访问也将取决于伦理批准。在第 10.4.5.1 节中,我们讨论了医学伦理学的起源,美国的塔斯基吉实验和德国纳粹医生进行的实验,以及现在出现的保障措施:美国的机构审查委员会 (IRB) 和美国的伦理委员会欧洲。如果你是一名医学研究人员,除了在不稳定的法律基础上使用从医疗实践中收集的记录并使用有漏洞的去识别机制保护之外别无选择,那么你别无选择,只能依靠你的 IRB 或伦理委员会。尽管机构之间（和机构内部)的确切过程有所不同,但关键原则是此类研究必须得到独立于研究人员的人员的批准 通常是一位或多位匿名同事,他们评估调查的目的和拟议的方法.然而,存在一些严重的道德风险。

11.4.4 阴暗面

伦理审查程序为研究人员提供了两个层面的责任保护。
首先,如果出现问题并且研究人员因疏忽而被起诉,这将使用“行业标准”作为衡量标准进行评估。如果您遵循与其他人相同的流程,并且每个项目都由包含“独立”成员(实际上是指来自其他大学的教授,而不是真实数据主体的代表)的伦理委员会批准,那么您可以做出强有力的如果您遵循这些标准。

其次,如果最坏的情况发生并且您面临刑事起诉的可能性,在涉及双重测试的英美法系国家:“犯罪意图”或错误意图,以及“犯罪行为”或被禁止的行为。道德批准程序旨在提供没有犯罪意图的证据。如果你按照你说的去做,并且出于独立人士认可的原因,那怎么会是错误的意图呢?简而言之,优化伦理审查流程是为了保护研究人员和机构,而不是数据主体。

这并没有逃过大数据的关注。在第 11.2.5 节中,我提到了 Google DeepMind 的道德委员会及其未能防止丑闻;谷歌设法逃脱了信息专员的谴责(不像医院交出所有医疗记录)。毫不奇怪,道德委员会正在激增,尤其是当公司开始在大型数据仓库中使用人工智能和机器学习技术时,他们几乎不清楚结果可能是什么。人工智能伦理是学术界的热门话题,也是快速增长的就业来源。愤世嫉俗的运营商将采取行动遵守 UKAN 的一些建议,然后雇用一些失业的哲学家来谈论道德哲学和智力的本质,同时继续将您最私密的个人信息出售给垃圾邮件发送者的业务.道德清洗和数据滥用现在齐头并进。

更重要的是,公开宣传的隐私机制的存在可能会转移人们对潜在个人数据滥用的注意力。 2007 年 3 月,历史学家玛戈·安德森 (Margo Anderson) 和威廉·塞尔策 (William Seltzer) 发现人口普查的机密性在 1942 年被暂停,居住在华盛顿特区的日裔美国人的微观数据在 1943 年被提供给特勤局 [1699]。块级数据被提供给加利福尼亚州的官员,他们在那里围捕日裔美国人进行拘留。那里的单点故障似乎是人口普查局局长 JC Capt,他应财政部长 Henry Morgenthau 的要求向特勤局发布了数据。该局此后公开道歉[1319]。但这并不是什么新鲜事。 1914 年第一次世界大战爆发时,英国政府利用 1911 年的人口普查将外国人驱逐出境; 1941 年的人口普查被提前到 1939 年,作为征兵、配给和拘留的基础;直到 1980 年代,安全部门一直对人口普查有后门。在其他地方,德国人利用人口普查数据不仅在德国而且在荷兰和其他被占领领土上围捕犹太人。最近,Cambridge Analytica 及其母公司 SCL 被一些国家授予秘密访问完整的国家人口普查数据的权限,他们帮助现任政府赢得连任 [2052]。

有许多公开宣传的隐私机制的例子

不如他们看起来那么有效。英国正在建立一个“智能电表”系统，该系统通过中央票据交换所报告每个人的燃气和电力消耗情况，这些数据会从中央票据交换所发送到您的公用事业公司，以便他们向您收费；其他公司需要获得批准的隐私计划才能访问数据。然而，当我们查看典型的隐私计划时，我们会看到配电网络运营商可以访问其配电区域、米德兰兹、西南和威尔士 [2011] 的半小时电表数据。目的是预测何时必须更换变电站变压器。分销商承诺将每条馈线的馈电汇总为半小时总计 这些是离开变压器并为许多房屋供电的电缆。但是查看数据，我们发现 0.96% 的饲养者只为一所房子服务，2.67% 的饲养者为 3 所或更少。一个更强大的隐私监管机构会告诉他们只在自己的变压器上安装自己的仪表。事实上，更明智的公共政策应该是根本不做智能电表项目；我将在第 14 章讨论这个问题。

在医药方面，美国 HIPAA 系统授权 DHHS 监管健康计划、医疗保健信息交换所和医疗保健提供者，但将处理医疗数据的许多其他组织（例如律师、雇主和大学）排除在其范围之外。大型科技公司可能会逃避监管，具体取决于他们声称为谁处理数据。在英国，正如我们已经指出的那样，无论是患者选择退出还是广告宣传的去识别机制都无效。在许多国家/地区，比您想象的更多的组织可以访问完全可识别的数据。

11.5 总结

许多人愿意相信，您可以通过剥离 o 公开的标识符（例如姓名）来将敏感的个人数据变成工业原材料。这只适用于一些定义明确的特殊情况，例如全国人口普查 我们有一个以差异隐私形式存在的可靠理论。在大多数情况下，数据过于丰富，重新识别数据主体很容易。

然而，政策制定者、营销人员、医学研究人员和其他人都很难相信匿名为使用个人数据提供了一种神奇的解决方案，因此很难阻止他们滥用。围绕大数据和机器学习的不断炒作使教育任务变得更加困难，就像这些技术使匿名变得更加困难一样。随着隐私法违法的规模和范围对公众越来越清楚，我们可能会遇到严重的麻烦。可能需要丑闻才能带来真正的改变，而当这最终发生时，破坏可能是不小的。

研究问题

目前，围绕匿名和隐私有几个活跃的研究线索。首先，有实用的研究人员正在寻找从现有公共数据中获取敏感数据的新方法，或者试图了解营销人员和网络犯罪分子正在实施的漏洞利用。其次，有数学家正在研究在

各种背景,例如从相互不信任的公司持有的数据中学习。
第三,有隐私法学者试图研究如何缩小法律与实践之间的差距。第四,有实际的活动家（如 EPIC、Privacy International 和 Max Schrems)提起诉讼,试图阻止那些正在变得普遍但似乎违反我们已有法律的做法。随着我们周围越来越多的研究变得“聪明”,这个由理论、实践、学术和竞选组成的生态系统无疑将继续发展。 “智慧城市”是否仅仅意味着更普遍的监控?在极限情况下,是否会有如此多的可用上下文信息,以至于只有差异化隐私才能发挥作用?或者社会最终会说适可而止,并对数据的收集、分析和使用施加根本限制哪些限制可能会奏效?最后,最新的魔药是保护隐私的联邦机器学习。毫无疑问,人们可以找到边缘案例,在这些案例中,类似的东西可以发挥作用,就像差异隐私一样。但我怀疑它最终会变成过去 40 年来我们一直被灌输的关于匿名化的万金油的一种变体。 （嘿,如果你用氢氧化钠煮蛇油,你应该得到蛇皂。)揭穿它的最好方法是什么?

延伸阅读

如果你想深入了解差异隐私的细节,一个很好的起点可能是 Cynthia Dwork 和 Aaron Roth [594] 的一篇长篇调查论文。关于推理控制的经典参考资料是 Dorothy Denning 1982 年的书 [538]； Adam 和 Wortman 于 1989 年发表的调查论文很好地总结了当时的技术水平 [17]。参与美国政府工作的统计学家的一个重要参考是联邦统计方法委员会的“统计披露限制方法报告”,它介绍了美国各部门和机构使用的工具和方法[667];这可以追溯到 2005 年,所以有点过时了,目前正在重写。如果您要为在英国管辖范围内运营的客户进行匿名化处理,则必须阅读 UKAN 书 [626]。作为一个完全不同的应用示例,Mark Allman 和 Vern Paxson 在 [42] 中讨论了网络系统研究的匿名 IP 数据包跟踪问题。最后,可以在 [52] 中找到 Margo Anderson 和 William Seltzer 关于美国滥用人口普查数据的论文,尤其是在第二次世界大战期间。