第 5 章　隐私保护技术

内容提要：随着计算机、移动互联网等技术的发展和应用，用户的电子医疗档案、互联网搜索历史、社交网络记录、GPS 设备记录等信息的收集、发布等过程中涉及的用户隐私泄露问题越来越引起人们的重视。大数据场景下，多个不同来源的数据基于数据相似性和一致性进行链接，产生新的更丰富的数据内容，也给用户隐私保护带来更严峻的挑战。本章介绍围绕用户隐私的典型数据、隐私保护需求、相应的攻击和保护技术，包括传统人口统计数据中的用户身份攻击、社交网络中的用户社交关系和属性推测、位置社交网络中的用户隐私位置推测和活动规律挖掘，以及对应的隐私保护技术等。早期基于典型的数据库表结构数据的研究为新出现的社交网络数据和轨迹数据研究提供了经典模型，后续研究更针对后两者的独特数据特征和保护需求。差分隐私模型提出了目前最严格的隐私定义，并忽略了对数据内容、攻击者能力的假设，但对数据可用性具有一定影响。隐私保护技术需要立足于具体场景的数据构成，综合考虑用户的多种隐私信息间的相关性，结合多种技术，才能提供全面的隐私保护解决方案。

关键词：身份隐私；社交关系隐私；属性隐私；轨迹隐私；链接攻击；同质攻击；近似攻击；k-匿名；l-多样化；t-贴近；社交关系推测；马尔可夫模型；高斯混合模型；贝叶斯模型；活动建模；时空模型；差分隐私；本地差分隐私；Rappor 协议；SH 协议。

5.1　基本知识

大数据时代，人类活动前所未有地被数据化。移动通信、数字医疗、社交网络、在线视频、位置服务等应用积累并持续不断地产生大量数据。以共享单车为例，截至 2017 年 5 月底，国内共享单车累计服务已超过 10 亿人次，注册用户超过 1 亿个。面向这些大规模、高速产生、蕴含高价值的大数据的分析挖掘不但为本行业的持续增长做出了贡献，也为跨行业应用提供了强有力的支持。共享单车的骑行路线在交通预测、路线推荐、城市规划方面具有重要意义[1]。

而随着数据披露范围的不断扩大，隐藏在数据背后的主体也面临愈来愈严重的隐私挖掘威胁，例如根据骑行路线推理个人用户的家庭住址、单位地址、出行规律，或者匿名用户被重新识别出来，进而导致“定制化”攻击，等等，为用户带来了极大损失。2017 年 6 月 1 日起，最高人民法院、最高人民检察院联合发布的《关于办理侵犯公民个人信息刑事案件适用法律若干问题的解释》正式生效，其中对“非法获取、出售或者提供行踪轨迹信息、通信内容、征信信息、财产信息 50 条以上的”等 10 种情形明确入罪，体现了国家对个人信息保护的重视。

为满足用户保护个人隐私的需求及相关法律法规的要求，大数据隐私保护技术需确保公开发布的数据不泄露任何用户敏感信息。同时，隐私保护技术还应考虑到发布数据的可用性。因为片面强调数据匿名性，将导致数据过度失真，无法实现数据发布的初衷。因此，

数据隐私保护技术的目标在于实现数据可用性和隐私性之间的良好平衡。

1. 数据隐私保护场景

一般来说，一个隐私保护数据发布方案的构建涉及以下 4 个参与方：

(1) 个人用户：收集数据的对象。

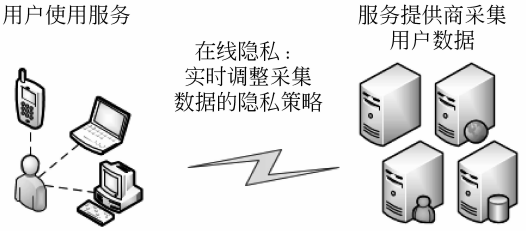
(2) 数据采集/发布者：数据采集者与用户签订数据收集、使用协议，获得用户的相关数据。数据采集者通常也负责数据发布(用户本地隐私保护情景除外)。根据数据发布的目的和限制条件，数据发布者对数据进行一定的处理并以在线交互或离线非交互方式提供给数据使用者，在进行数据处理时还须预防潜在的恶意攻击。

(3) 数据使用者：任意可获取该公开数据的机构和个人。数据使用者希望获得满足其使用目的的尽可能真实有效的数据。

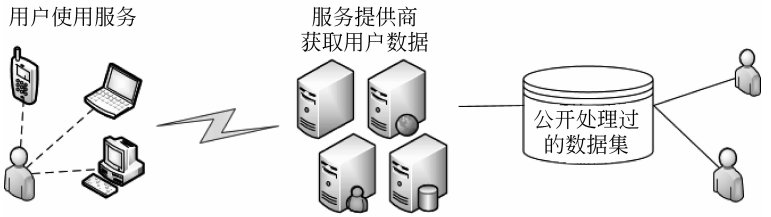
(4) 攻击者：可获取该公开数据的恶意使用者。攻击者可能具有额外的信息或者知识等，试图利用该公开数据识别特定用户身份，获取关于某特定用户的敏感信息，进而从中牟取利益。

攻击者的能力可分为两类。一类是背景知识(background knowledge)，通常是关于特定用户或数据集的相关信息。如攻击者可能知道 Amanda 是部门经理，Alice 是营业员，Bill 的出生日期是 1976 年 12 月 1 日。背景知识的获得完全基于攻击者对具体攻击目标的了解，攻击者可以利用其掌握的背景知识，在公开发布的数据中识别出某个特定用户。另一类是领域知识(domain knowledge)，指关于某个领域内部的基本常识，通常具有一定的专业性。例如，医学专家可能了解不同区域人群中某种疾病的发病率。当攻击者将目标范围缩小到有限的记录集时，攻击目标可能患有的疾病也仅限于记录集中的几种。具有医学知识的攻击者可以根据攻击目标的地域推理出其可能患有的疾病。

在实际场景中，数据采集/发布者隐私保护方案可选择在线模式或离线模式。在线模式又称“查询-问答”模式，对用户所访问的数据提供实时隐私保护处理。在在线模式(图 5-1(a))下，通过数据发布者的调控，数据被收集的个人用户和期望获得真实数据的使用者之间



(a) 简单的在线数据隐私场景



(b) 简单的离线数据隐私场景

图 5-1　数据隐私场景示意图

应能够就数据的使用目的、范围、限制情况达成一致。但在线模式对算法性能要求较高。离线模式(图 5-1(b))是指在对所有数据统一进行隐私保护处理后批量发布。数据一旦公开发布,数据发布者和数据被收集的个人用户就失去了对数据的监管能力。任意获得该公开数据的第三方,包括恶意攻击者在内,都可以对这些数据进行深入分析。因此,在离线模式下,数据发布者应力求提前预测攻击者的所有可能攻击行为,并采取有针对性的防范措施。即使无法对攻击者的所有行为进行预测,数据发布者也应重点关注个人用户最基本的隐私保护需求,并进行对应的保护方案设计和攻击预防,从而避免对个人用户的隐私造成严重侵害。本章主要讨论离线模式数据发布场景。

2. 隐私保护需求

用户隐私保护需求可分为身份隐私、属性隐私、社交关系隐私、位置与轨迹隐私等几大类。

(1) 身份隐私。它是指数据记录中的用户 ID 或社交网络中的虚拟节点对应的真实用户身份信息。通常情况下,政府公开部门或服务提供商对外提供匿名处理后的信息。但是一旦分析者将虚拟用户 ID 或节点和真实的用户身份相关联,即造成用户身份信息泄露(也称为"去匿名化")。用户身份隐私保护的目标是降低攻击者从数据集中识别出某特定用户的可能性。

(2) 属性隐私。属性数据用来描述个人用户的属性特征,例如结构化数据表中年龄、性别等描述用户的人口统计学特征的字段。宽泛地说,用户购物历史、社交网络上用户主动提供的喜欢的书、音乐等个性化信息都可以作为用户的属性信息。这些属性信息具有丰富的信息量和较高的个性化程度,能够帮助系统建立完整的用户轮廓,提高推荐系统的准确性等。然而,用户往往不希望所有属性信息都对外公开,尤其是敏感程度较高的属性信息。例如,某些视频观看记录被公开会对用户的形象造成不良影响。但是,简单地删除敏感属性是不够的,因为分析者有可能通过对用户其他信息(如社交关系、非敏感属性、活动规律等)进行分析、推测将其还原出来。属性隐私保护的目标是对用户相关属性信息进行有针对性的处理,防止用户敏感属性特征泄露。

(3) 社交关系隐私。用户和用户之间形成的社交关系也是隐私的一种。通常在社交网络图谱中,用户社交关系用边表示。服务提供商基于社交结构可分析出用户的交友倾向并对其进行朋友推荐,以保持社交群体的活跃和黏性。但与此同时,分析者也可以挖掘出用户不愿公开的社交关系、交友群体特征等,导致用户的社交关系隐私甚至属性隐私暴露。社交关系隐私保护要求节点对应的社交关系保持匿名,攻击者无法确认特定用户拥有哪些社交关系。

(4) 位置轨迹隐私。用户位置轨迹数据来源广泛,包括来自城市交通系统、GPS 导航、行程规划系统、无线接入点以及各类基于位置服务的 APP 数据等。用户的实时位置泄露可能会给其带来极大危害,例如被锁定并实施定位攻击。而用户的历史位置轨迹分析也可能暴露用户隐私属性、私密关系、出行规律甚至用户真实身份,为用户带来意想不到的损失。用户位置轨迹隐私保护要求对用户的真实位置进行隐藏或处理,不泄露用户的敏感位置和行动规律给恶意攻击者,从而保护用户安全。

从数据类型角度看,用户隐私数据可表示为结构化数据或非结构化数据。通常,用户的属性信息(如年龄、性别、购物记录等)属于典型的结构化数据,可表示为数据库表;用户位

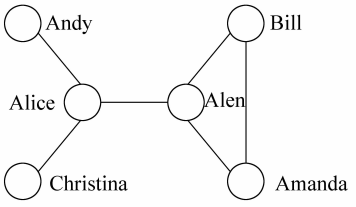
置、轨迹数据一般以点集的形式表示，也属于结构化数据。而用户社交关系数据则表现为相对复杂的网络关系，属于非结构化数据，一般用图结构表示。图 5-2 中展示了基本数据类型。为了表达两者之间的关联，后文中将用户隐私表示为“属性-图”结构。

姓名	年龄	性别	邮编	工资
Andy	42	M	100190	1000
Alice	22	F	100190	1100
Alen	53	M	100180	1200
Bill	42	M	100180	1300
Amanda	22	F	100170	1400
Christina	53	F	100170	1500

(a) 关系型表数据

Id	Time	Longitude	Latitude	Tid
Andy	2016.12.23	39.9777985	116.3353885	T000
Andy	2016.12.23	39.9777985	116.3351000	T000
Alice	2016.12.25	39.9674738	116.3392735	T000
Alice	2016.12.25	39.9675288	116.3392885	T000
Alice	2016.12.25	39.9675288	116.3392885	T000
Alice	2016.12.25	39.9708951	116.3214983	T000

(b) 轨迹数据



(c) 社交结构数据

图 5-2　基本数据类型

除了数据类型不同，用户的关系型表数据、位置轨迹数据、社交结构数据在各自的数据维度上也具有明显不同的特性。数据表中的一条记录通常只代表一个用户，用户间的相关性较弱。记录之间的相关性基本上只与其所处的统计分组有关，属性之间的相关性只与整个表呈现出的数据分布有关。个人的位置轨迹数据通常是一系列长度不定的点集序列，具有明显的时间顺序和周期重复特征，反映了个人运动规律，使得用户的运动轨迹易于被预测，而难以合理、高效地彻底隐藏。社交网络数据中除了属性数据，还具有复杂的边连接。在这种场景中，用户通过边连接进行影响力传播和相似性传递，最终导致“朋友的朋友也是我的朋友”的局部相似性日益凸显，使得用户的属性、社交关系甚至身份容易从局部社区中被推测出来。隐私保护技术必须针对不同数据的特征进行处理，才能实现期望的隐私保护效果。

3. 隐私保护技术分类

前面提到，数据隐私保护的目标在于实现数据可用性和隐私性之间的良好平衡。因此，一个隐私保护方案有明确的隐私保护目标与可用性目标。

当前的隐私保护模型有两大类：以 k-匿名为代表的基于等价类的方法和差分隐私方法。前者假设攻击者能力有限，仅能将攻击目标缩小到一定的等价类范围内，而无法唯一地准确识别攻击目标；后者则假设可能存在两个相邻数据集，分别包含或者不包含攻击目标，但攻击者无法通过已知内容推出两个数据集的差异，因此，也无法判断攻击目标是否在真实

数据集中。前者的优势在于,在攻击者能力不超过假设的前提下,能够以较小的代价保证同一等价类内记录的不可区分性。而如果攻击者能力超过了假设,攻击者就能够进一步区分等价类内的不同记录,从而实现去匿名化。后者的优势在于,攻击者不可能具有超过假设的攻击能力,因而不可能突破差分隐私方法提供的匿名保护。但是,由于数据集的差异性,差分隐私方法可能会对原始数据造成较大扰动,过度破坏数据可用性。

典型的隐私保护技术手段包括抑制(suppression)、泛化(generalization)、置换(permutation)、扰动(perturbation)、裁剪(anatomy)等。此外,也有人通过密码学手段实现隐私保护。

(1) 抑制是最常见的数据匿名措施,通过将数据置空的方式限制数据发布。

(2) 泛化是指通过降低数据精度来提供匿名的方法。属性泛化即通过制定属性泛化路径,将一个或多个属性的不同取值按照既定泛化路径进行不同深度的泛化,使得多个元组的属性值相同。最深的属性泛化效果通常等同于抑制。社交关系数据的泛化则是将某些节点以及这些节点间的连接进行泛化。位置轨迹数据可进行时间、空间泛化。

(3) 置换方法不对数据内容作更改,但是改变数据的属主。例如,将不同的个人用户的属性值互相交换,将用户 a 与 b 之间的边置换为 a 与 c 之间的边。

(4) 扰动是在数据发布时添加一定的噪声,包括数据增删、变换等,使攻击者无法区分真实数据和噪声数据,从而对攻击者造成干扰。

(5) 裁剪技术的基本思想是将数据分开发布。例如,对于表结构数据,首先将用户划分为不同的组,赋予同一组的记录相同的组标识符(group id),对应记录的敏感数据也赋予相同的组标识符,然后将准标识符(如地域、性别等)和敏感数据分别添加组标识符作为两张新表发布。恶意攻击者即使可以确定攻击目标的组标识符,但是无法有效地从具有相同组标识符的敏感数据中判定攻击目标对应的敏感数据。

(6) 密码学手段利用数据加密技术阻止非法用户对数据的未授权访问和滥用。

隐私保护方案需要引入可用性标准。一种通用的机制是度量数据失真程度,并不考虑发布的数据被如何使用。通过定义一系列数据集属性特征,比较真实数据和数据发布版本的特征变化来衡量数据损失程度。例如,对于关系型数据表中的数值型数据,计算其平均值的偏移量。如果数据有明确的应用领域,例如对数据进行统计分析、计算均值、找出 Top-k 对象等,那么可用性指标可以更具体化,表示为计算结果的准确度。

5.2 关系型数据隐私保护

2002 年,Sweeney[2,3] 提出了 k-匿名模型,这是第一个真正意义上完整的隐私保护模型。这一方案能够杜绝攻击者唯一地识别出数据集中的某个特定用户,使其无法进一步获得该用户的准确信息,能够提供一定程度的用户身份隐私保护。在 Sweeney 提出的隐私方案中明确了对数据可用性和用户隐私性的保证。此外,人们还关注表结构数据中的用户敏感属性的隐私保护需求。根据敏感属性的分布情况,人们提出了 l-多样化、t-贴近模型。这些方法为后续社交网络隐私保护与位置轨迹隐私保护奠定了基础。

本节主要介绍早期的表结构数据研究中的身份匿名和属性匿名方法、一些常见的攻击方法以及数据连续发布场景中的问题与解决方案。

5.2.1 身份匿名

1. 链接攻击与身份匿名

简单地去标识符匿名化仅仅去除了表中的身份ID等标志性信息，攻击者仍可凭借背景知识，如地域、性别等准标识符信息，迅速确定攻击目标对应的记录。此类攻击称为记录链接(record linkage)攻击，简称链接攻击。如表5-1所示，原始用户医疗记录表中包含了Name(用户姓名)这一标识符，简单删除标识符列之后可以得到如表5-2所示的匿名记录表。如果攻击者持有公开的选民记录表作为背景知识(表5-3)，与公开发布的匿名记录表对比，通过Z(邮编)、Age(年龄)等若干项属性信息，攻击者仍可以唯一地识别出某些用户。例如，可推断出第2条记录对应的用户是Bob。

表5-1 原始的用户医疗记录表

	Identifier	Quasi-identifier			Sensitive Data
#	Name	ZIP	Age	Nationality	Condition
1	Kumar	13053	28	Indian	Heart Disease
2	Bob	13067	29	American	Heart Disease
3	Ivan	13053	35	Canadian	Viral Infection

表5-2 匿名后的用户医疗记录表

	Quasi-identifier			Sensitive Data
#	ZIP	Age	Nationality	Condition
1	13053	28	Indian	Heart Disease
2	13067	29	American	Heart Disease
3	13053	35	Canadian	Viral Infection

表5-3 选民记录表

Name	ZIP	Age	Sex	Vote
Natalia	13053	28	Female	Yes
Bob	13067	29	Male	Yes
Lisa	13053	35	Female	No
Umeko	13067	36	Female	Yes

2. k-匿名基本模型

为避免攻击者通过链接攻击从发布的数据中唯一地识别出特定匹配用户，导致用户身份泄露，Samarati和Sweeney最早提出了适用于关系型数据表的k-匿名(k-anonymity)模型[2,3]。这一方案按照准标识符将数据记录分成不同的分组，且每一分组中至少包含k条记录。这样，每个具有某个准标识符的记录都至少与$k-1$个其他记录不可区分，从而实现用户身份匿名保护。

定义5-1(k-匿名) 令$T(A_1,A_2,\cdots,A_n)$为一张行数有限的表，属性集合为$\{A_1,A_2,\cdots,A_n\}$。QI_T为表中的准标识符$QI_T=\{A_i,A_{i+1},\cdots,A_j\}$。表$T$满足$k$-匿名，当且仅当每一组准标识符的取值序列在$T[QI]$中出现至少$k$次。

为了让发布的数据满足k-匿名需求，Samarati和Sweeney给出了相应的数据处理方法，提出了一种通过元组泛化实现k-匿名的解决方案。

属性A的泛化函数可表示为$f:A\rightarrow B$。属性A的持续泛化过程可表示为域泛化层次

结构(domain generalization hierarchy)DGH_A，通过一组函数 $f_h(h=0,1,\cdots,n-1)$ 的作用，实现从属性 A 的所有取值泛化到“任意”或者“*”的完整泛化路径：$A_0 \xrightarrow{f_0} A_1 \xrightarrow{f_1} \cdots \xrightarrow{f_{n-1}} A_n$。其中 $A_0=A$，$|A_n|=1$。例如，ZIP 编码可由具体的 02138 逐步或直接泛化为不具体的 0213^*、021^{**}、02^{***}、0^{****}、*****。出生年份可由精确的 1965 泛化为 1960—1970、1950—1970。泛化路径的属性值之间存在偏序关系。对于属性 A 的两个泛化值 v_i 和 v_j，若 $i \leqslant j$ 且 $f_{j-1}(\cdots f_i(v_i)\cdots)=v_j$，那么 v_i 和 v_j 存在偏序关系，表示为 $v_i \leqslant v_j$。

显然在泛化层次树中，离树根越近的节点泛化程度越高，对数据的破坏越大。为了在数据处理过程中尽可能保持数据可用性，同时，尽快满足 k 个相同记录的需求，Sweeney 等人提出了 k-匿名最小泛化的概念。

定义 5-2(k-匿名最小泛化) 令 $T_1(A_1,A_2,\cdots,A_n)$ 和 $T_m(A_1,A_2,\cdots,A_n)$ 分别为两张表，其准标识符均为 $QI_T=\{A_i,A_{i+1},\cdots,A_j\}$，且 $T_1[QI_T] \leqslant T_m[QI_T]$。称 T_m 是表 T_1 的 k-匿名最小泛化，当且仅当满足以下两个条件：

(1) T_m 在定义的准标识符 QI_T 上符合 k-匿名模型。

(2) $\forall T_z$：$T_1 \leqslant T_z$，$T_z \leqslant T_m$，如果 T_z 也满足 k-匿名模型，那么必然有 $T_z[QI_T]=T_m[QI_T]$。

在存在多种符合 k-匿名模型的最小泛化的场景中，需要进一步比较泛化过程中的数据扰动来选取最优的泛化方案。为此，Sweeney 等人定义了数据准确度 Prec 来衡量泛化过程中的信息变化以及定义最小扰动的概念。

定义 5-3(数据准确度 Prec) 令 PT 为原始数据表。表 PT 的准标识符由 N_a 个属性 $\{A_1,A_2,\cdots,A_{N_a}\}$ 组成，共包含 N 条记录，tp_j 为表 PT 中的第 j 条记录。RT 为 PT 的一个泛化表，tr_j 为与表 PT 中 tp_j 对应的泛化后记录。h_{ji} 为 tr_j 中属性 A_i 的泛化结果 $tr_j[A_i]$ 处于该属性的泛化层次结构的路径深度。DGH_{Ai} 为属性 A_i 泛化层次结构的高度。RT 的数据准确度由下式确定：

$$\mathrm{Prec}(\mathrm{RT})=1-\frac{\sum_{i=1}^{N_a}\sum_{j=1}^{N}\frac{h_{ji}}{|\mathrm{DGH}_{A_i}|}}{N \cdot N_a}$$

定义 5-4(最小扰动) 令 $T_1(A_1,A_2,\cdots,A_n)$ 和 $T_m(A_1,A_2,\cdots,A_n)$ 分别为两张表，其准标识符均为 $QI_T=\{A_i,A_{i+1},\cdots,A_j\}$，且 $T_1[QI_T] \leqslant T_m[QI_T]$。$\forall x=i,i+1,\cdots,j$，$DGH_{A_x}$ 是准标识符 QI_T 的域泛化层次结构。称 T_m 是表 T_1 符合 k-匿名模型的最小扰动，当且仅当满足以下两个条件：

(1) T_m 在定义的准标识符 QI_T 上符合 k-匿名模型。

(2) $\forall T_z$：$\mathrm{Prec}(T_1) \geqslant \mathrm{Prec}(T_z)$，$\mathrm{Prec}(T_z) \geqslant \mathrm{Prec}(T_m)$，如果 T_z 也满足 k-匿名模型，那么必然有 $T_z[QI_T]=T_m[QI_T]$。

根据定义，若 PT 中的记录未经过泛化，则任意记录的准标识符属性 $h=0$，$\mathrm{Prec}(\mathrm{PT})=1$。在另一种极端情况下，RT 中的准标识符属性均泛化到层次结构的根节点，那么 $h=|\mathrm{DGH}|$，$\mathrm{Prec}(\mathrm{RT})=0$。在实际数据隐私处理的过程中，数据发布者希望获得较高的数据准确度，就必须尽可能少地进行数据泛化，也就是说，使得数据泛化的位置尽可能离泛化层次结构的根节点更近，以实现最小扰动。

Sweeney 等人设计了一种最小扰动的 k-匿名泛化算法。该算法包括如下两个步骤：

(1) 判断 PT 是否符合 k-匿名模型，如果是，输出 PT，否则进入第(2)步。

(2) 执行如下操作：

(2.1) 生成 PT 的所有可能的泛化表集合，记为 allgens。

(2.2) 检测 allgens 中符合 k-匿名模型的泛化表，将该集合记为 protected。

(2.3) 保存 protected 中符合最小扰动的泛化表，记为 MGT。

(2.4) 根据用户定义的偏好，从 MGT 中输出唯一的符合用户偏好的最小扰动输出。

在基本 k-匿名算法的基础上，Lefevre 等人[4]提出了一个基于贪心算法的改进方案，重点优化了寻找最小扰动的过程，算法的效率有了很大提高。Bayardo 等人[5]给出了基于数据拆分发布和元组抑制的解决方案。

3. *k-匿名模型的局限性*

用户购物历史、观影历史等数据虽然也可以用数据表的形式表示，但是，这类数据中不存在严格的准标识符信息。因为数据发布方无法准确界定哪一条购买记录和用户评价信息是用户的准标识符信息，任何非特定记录都可能被攻击者用来重新识别出用户身份。很显然，基础的 k-匿名模型的适用范围并不包括这类数据，而是仅限于能准确定义准标识符属性的关系型表结构数据。

2006 年 Netflix 的用户隐私泄露事件就是由于公开的用户观影记录匿名程度不足而导致部分用户的身份泄露。Narayanan 等人随后在 2008 年的 S&P 会议上公开了他们利用 IMDB 数据库对 Netflix 数据进行链接攻击的方法[6]。该文直观地展示了 k-匿名模型的不足。

首先，该文定义了一个简单的打分比较算法。假设当前攻击者获得了关于某个特定攻击目标的额外信息，需要根据这些信息判定攻击目标与当前待定用户 r'的相似度。打分算法就是用来计算当前掌握的关于攻击目标的额外信息 aux 和待定用户 r'的所有属性的相似程度：

$$\mathrm{Score}(\mathrm{aux},r') = \min_{i\in \mathrm{supp}(\mathrm{aux})} \mathrm{Sim}(\mathrm{aux}_i, r'_i)$$

这个算法比较了攻击目标的额外信息 aux 和待定用户 r'的所有属性，并将属性相似性分值最小的记为两者的相似性打分。这里采用的 Sim 函数求得的是余弦相似性。在这种思想下，如果两个“用户”aux 和 r'在某个属性上差异特别巨大，那么这两者基本不可能是同一个用户。但如果额外信息 aux 或者待定用户 r'中的某个属性出现错误，就很容易导致两者的相似性打分非常低，所以将相似性打分公式更新为

$$\mathrm{Score}(\mathrm{aux},r') = \sum_{i\in \mathrm{supp}(\mathrm{aux})} \mathrm{wt}(i)\mathrm{Sim}(\mathrm{aux}_i, r'_i)$$

其中，$\mathrm{wt}(i)=\dfrac{1}{\log |\mathrm{supp}(i)|}$，$\log |\mathrm{supp}(i)|$为 r'所处的数据集中具有属性 i 的用户数。在这种情况下，越稀有的属性权重越高，两个“用户”的加权相似性最高，那么他们就可能是同一个用户的两个 id。

基于这个打分算法，Narayanan 等人选取了 IMDB 数据集中的 50 个用户和 Netflix 公开数据集的用户进行了打分匹配。他们利用 IMDB 数据集中的用户观影打分作为额外信息。实验发现，如果用户在 Netflix 和 IMDB 发布的影片评分相同，并且日期相差不远，此类

评分越多，用户账户越容易匹配。实验同时还发现，如果用户评分的电影越小众，他也越容易被识别，也符合打分公式中较少的人具有的属性权重较大的设置。在该文中，Narayanan等人指出，在实际的多维数据发布场景中，数据通常很稀疏，攻击者可能只需要掌握很少的属性(5～10个非热门电影)，就能识别出大量用户。实际上也就是说，与用户具有相同属性的人越少，用户的唯一性越强，该用户越容易被识别。

Narayanan等人的研究实际上也表明，受限于攻击者掌握的额外信息，只要用户能够和k个其他用户具有相同的观影历史，实际上攻击者是没有办法区分他们的。虽然攻击者无法确定到底哪一个id是他的攻击目标，但是实际上他已经获得了该用户的所有观影历史，也达到了一定的攻击目标，即使其达到的攻击目标与用户身份无关。

除了需要解决k-匿名模型本身的缺陷导致数据匿名不足的问题，当前的数据隐私保护方案还需要抗衡数据去匿名算法的攻击。随着大数据技术的不断发展，数据持有者自然地希望获得更多用户数据以综合分析并发掘其中的价值。在这种场景下，首先需要实现多源数据中的用户重识别，进而实现用户数据融合。多源数据融合场景中的用户重识别实际上就是根据异源数据的额外信息确定用户身份的去匿名化攻击过程。根据异源数据的来源和精确程度不同，去匿名化攻击可分为3种：基于特定模式精确匹配的去匿名、基于种子匹配的去匿名和基于相似度匹配的去匿名。

基于特定模式精确匹配的去匿名算法无法抵抗噪声影响。一旦数据经上述某种匿名化算法引入噪声，就不再有效了。

上文提到的针对Netflix数据的攻击实际上是一种基于种子匹配的去匿名攻击。在这类方案中，攻击者首先需要了解一定数量的用户在两个图之间的节点对应关系(种子匹配)。算法从种子匹配出发，计算不同网络中的连接节点间的相似度，并将相似节点进行匹配，从而实现多网络间用户身份的重识别。

基于相似度匹配是在不具有先验知识(种子数据)的情况下普遍采用的去匿名方法。Cao等人[7]基于MapReduce框架进行异源轨迹数据的用户重识别。数据预处理把轨迹处理为停留点(stay point)集合，然后对比潜在用户的SIG(signal based similarity)判断这些用户是否为同一个人。在这个模型中，将用户停留点分为核心地点和普通地点，核心地点发出刺激信号，普通地点不发出信号，而是收到随距离衰减的刺激信号。两个用户轨迹中的点的SIG相似性越高，越可能是同一个人在不同数据源留下的轨迹。

综上所述，可以看到，k-匿名模型的相关研究实际上陷入了很大的困境。正如上文所述，k-匿名模型仅适用于存在明确准标识符的数据，而不适用于当前大数据时代规模庞大的非表结构数据，其使用范围有限。其次，大量的去匿名算法试图通过模糊的种子匹配和相似度匹配算法识别出最相近的用户，从而避免了k-匿名算法对精确匹配算法造成的干扰，仍旧泄露了用户的特征，大大削弱了k-匿名算法的保护能力。但k-匿名模型作为经典的身份隐私保护模型仍在实际隐私保护应用中发挥作用，可为用户提供一定的隐私保护。

5.2.2 属性匿名

1. 同质攻击

在5.2.1节中讨论的k-匿名模型能够用来防止链接攻击，避免攻击者唯一地识别出攻击目标。那么，在发布的匿名数据满足k-匿名模型的情况下，是不是攻击者就不能从中推

测出用户的其他隐私信息？在经过 k-匿名处理后的数据集中，攻击目标至少对应于 k 个可能的记录。但这些记录只满足准标识符信息一致的要求，而非准标识符数据和敏感数据保持不变。正如在 5.2.1 节分析 Netflix 隐私泄露事件时所讨论的，如果这 k 个用户的观影记录相同或非常接近，攻击者也能够获得用户的所有观影历史，分析用户的隐私属性。例如，这 k 个用户都喜欢看海洋纪录片，分析的结果是攻击目标可能是环保主义者。在 k-匿名的数据记录中，如果记录的敏感数据接近一致或集中于某个属性，攻击者也可以唯一或以极大概率确定数据持有者的属性。这类攻击称为同质攻击。

2. k-匿名模型的变体

人们首先在 k-匿名模型的基础上进行了一系列改进，试图抵抗同质攻击。

Zhang 等人[8]提出了(k,e)-匿名模型，主要处理数值型敏感属性数据。(k,e)-匿名的思想是：要求每个等价类中元组个数至少是 k 个，同时等价类中敏感属性取值范围不能小于给定的阈值 e，也就是要求等价类中敏感属性的最大值与最小值的差至少是 e。

Wang 等人[9]提出了(X,Y)-匿名的概念。其中，X、Y 为不相交的属性集。在这种方案中，讨论了数据库表中多条记录代表同一个数据持有者的情况。在此类情况下，多条记录的准标识符值相同或者基本相同，很有可能被划分到同一等价类中。简单的 k-匿名要求难以实现对用户隐私的保护。为此，他们提出，在属性组 X 中的属性均相同的情况下，每一组 X 均需对应至少 k 个不同的敏感属性组 Y 中的值。这种方案在普通 k-匿名的基础上增加了对敏感数据的限制条件。因此，能够提供比 k-匿名更好的保护。

为避免用户敏感属性被推测，在社交网络中出现大量基于 k-匿名聚类的改进算法。Ford 等人[10]提出了 p-sensitive k-anonymity 方法，要求聚类中节点数大于或等于 k，并且不同敏感值属性个数大于或等于 p。Sun[11]在此基础上提出了 $p+$sensitive k-anonymity 方法，该方法采用敏感属性值的类别概念，要求敏感属性值的类别至少出现 p 类。

3. l-多样化模型

Machanavajjhala 等人[12]提出了 l-多样化(l-diversity)这一新的模型，要求在准标识符相同的等价类中，敏感数据要满足一定的多样化要求。他们通过熵来定义数据的多样化程度，提出了熵 l-多样化(entropy l-diversity)的概念。

定义 5-5(熵 l-多样化) 如果对每一个泛化的 $q*$ 条记录组，满足 $-\sum_{s\in S} p_{(q^*,s)}\log p_{(q^*,s)} \geqslant \log l$，那么该表满足熵 l- 多样化。

其中 $p_{(q^*,s')}=\dfrac{n_{(q^*,s)}}{\sum_{s\in S} n_{(q^*,s')}}$ 为 $q*$ 记录组中敏感值等于 s 的记录所占的比例。但是，这一要求过于严格。如果表格中 90%的用户敏感属性都是“健康”，$q*$ 记录组的熵 l-多样化很可能只有极少数不是“健康”，从而使得该 $q*$ 记录组无法满足熵 l-多样化的要求。

递归(c,l)多样化(recursive(c,l)-diversity)在此基础上降低了多样性的要求，并假设不会影响到用户隐私的属性可以公开，不将其作为敏感值进行保护，例如用户“健康”这一属性值。

定义 5-6(递归(c,l)-多样化) 将每一个 $q*$ 元组中用户敏感值按照出现的频繁程度降序排列，其出现次数分别为 $r_1,r_2,\cdots,r_m$，如果对每一个 $q*$ 元组，存在 $r_1<c(r_l+r_{l+1}+\cdots+r_m)$，即最频繁的属性频率 r_1 不超过最不频繁的 $m-(l-1)$ 个属性的频率之和的 c 倍，那么