ICS 35.240
CCS L 67

团　　体　　标　　准

T/ISC 0078—2025

匿名化技术互联网应用指南

Application guidelines of anonymization technology in Internet

（发布稿）

2025-07-18

2025-07-18 发布　　2025-08-18 实施

中国互联网协会　　发布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由中国互联网协会提出并归口。

本文件起草单位：深圳市腾讯计算机系统有限公司、腾讯云计算（北京）有限责任公司、联通数字科技有限公司、天翼安全科技有限公司、北京数牍科技有限公司、北京快手科技有限公司、郑州信大捷安信息技术股份有限公司、宏盟集团、北京市竞天公诚律师事务所、北京沃东天骏信息技术有限公司、秒针信息技术有限公司、国家广告研究院、尼洱市场研究（上海）有限公司、北京风行在线技术有限公司、利欧集团数字科技有限公司、中移动信息技术有限公司。

本文件主要起草人：吴以源、刘阳璐、罗素、陈昱、邓静恒、李朝霞、温源、康和、金银玉、落红卫、王昕、刘献伦、赵悦、唐青、周杨、李力、石钛戈、张泽华、林战刚、刘沛、于晓蕾、胡春磊、顾明毅、王其武、潘冲、周崧弢、李冠洲、王娅琼、高梦娇、滕飞。

引　言

互联网领域是数据收集、使用、加工、提供和委托处理密集的行业领域，在互联网广告、互联网医疗、互联网金融、互联网政务等场景中，都涉及到包括个人信息在内的海量数据的流转流通，这些应用模式面临一定的安全合规风险。而匿名化技术是互联网领域中重要的数据安全保障措施，已发展出许多成熟的技术解决方案。

在法律、法规、政策层面，《中华人民共和国数据安全法》、《中华人民共和国个人信息保护法》等法律，以及各部委的规章制度对各行业数据安全、个人信息保护提出明确要求。《个人信息保护法》提出个人信息经匿名化处理后所得的信息不属于个人信息，匿名化是指个人信息经过处理无法识别特定自然人且不能复原的过程。《关于构建数据基础制度更好发挥数据要素作用的意见》（“数据二十条”）提出创新技术手段推动个人信息匿名化处理，保障使用个人信息数据时的信息安全和个人隐私，以及促进数据要素有序流通。

在互联网业务中，参与机构多，生态链条长，很多机构无法直接获得个人同意或取得同意的成本巨大，这些数据处理行为在个人信息保护、商业秘密保护等方面，面临一定的安全风险。匿名化的合法路径将是互联网行业数据利用的重要主要方式。但是，目前互联网行业内，对匿名化、去标识化、假名化、数据脱敏等概念界定交叉、模糊，对于匿名化处理的合规要求、技术措施、管理措施、效果评价等方面的实践经验与理解差异较大，也缺少明确的数据匿名化实施定义与可行方案，因此有必要制定本团体标准，在互联网行业明确匿名化处理的目标与原则、管理措施、技术措施、操作流程、效果评价方法等，达成行业共识，从而促进互联网行业的合规、健康、稳定发展。

匿名化技术互联网应用指南

1 范围

本文件给出了互联网业务中个人信息匿名化处理技术的应用指南，包括匿名化技术应用的目标与原则、实施框架、需求分析、方案制定、技术处理、效果评价、评估审查、管理保障等。

本文件适用于指导互联网业务中的个人信息匿名化处理活动，也适用于对互联网业务中个人信息匿名化处理活动的监督、管理、评估。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 25069—2022 信息安全技术 术语

GB/T 35273—2020 信息安全技术 个人信息安全规范

GB/T 37964—2019 信息安全技术 个人信息去标识化指南

GB/T 42460—2023 信息安全技术 个人信息去标识化效果评估指南

T/TAF 137—2022 基于差分隐私的用户个人信息保护技术要求

ISO/IEC 20889:2018 隐私增强的数据去标识化术语与技术分类

3 术语和定义

下列术语和定义适用于本文件。

3.1

个人信息 personal information

以电子或者其他方式记录的与已识别或者可识别的自然人有关的各种信息。

[来源：GB/T 42460-2023，3.1]

注：不包括匿名化处理后的信息。

3.2

匿名化 anonymization

通过对个人信息的技术处理，使得个人信息主体无法被识别或者关联，且处理后的信息不能被复原的过程。

注1：个人信息经匿名化处理后所得的信息不属于个人信息。

注2：匿名化是去标识化的极端情况，即在受控环境下将特定主体被重标识的风险控制在可接受的范围内。

[来源：GB/T 35273-2020，3.14，有修改。]

3.3

去标识化 de-identification

通过对个人信息的技术处理，使其在不借助额外信息的情况下，无法识别或者关联个人信息主体的过程。

[来源：GB/T 35273-2020，3.15]

3.4

标识符 identifier

微数据中的一个或多个属性，可以实现对个人信息主体的唯一识别。

[来源：GB/T 37964—2019，3.6]

注：标识符分为直接标识符和准标识符。

3.5

直接标识符 direct identifier

微数据中的属性，在特定环境下可以单独识别个人信息主体。

[来源：GB/T 37964—2019，定义3.7]

3.6

准标识符 quasi-identifier

微数据中的属性，结合其它属性可唯一识别个人信息主体。

[来源：GB/T 37964—2019，定义3.8]

3.7

重标识 re-identification

把去标识化的数据集重新关联到原始个人信息主体或一组个人信息主体的过程。

[来源：GB/T 37964—2019，定义3.9]

4 缩略语

下列缩略语适用于本文件。

ID：标识符（Identifier）

5 匿名化技术应用的目标与原则

5.1 匿名化技术应用的目标

匿名化处理的目的是在满足合规和安全风险可控的情况下，对数据资源进行合法有效利用，促进释放数据使用价值。

数据提供者开展匿名化处理活动宜符合以下的目标要求：

a) 安全合规目标

1) 依据相关法律法规规定，在未取得用户同意或其他法律依据、用户撤回同意、超范围处理数据、用户注销账号、数据保存期限届满等情况下，如需继续开展个人信息处理活动，需要事先对个人信息进行匿名化处理；

2) 对拟开展匿名化的主体、行为、对象及对应的同意授权及应用场景等条件，进行充分的风险评定与匿名化处理的必要性论证，满足合规性要求；

3） 符合《个人信息保护法》《数据安全法》《网络安全法》《民法典》等法律法规规定的合规性要求，即经过加工无法识别特定个人且不能复原。

b） 技术保障目标

1） 实施的匿名化处理技术能避免重识别的风险，且能控制重关联行为，所必需的辅助信息需要单独进行维护和管理；

2） 鼓励创新技术手段，通过匿名化处理技术保障促进数据安全和个人信息保护。

c） 数据使用目标

1） 匿名化处理并非追求完美、绝对的匿名化状态，现有的各项匿名化技术都无法彻底消除处理后的信息所残留的再识别风险。如果匿名化处理在满足合规合法的前提下，数据使用方或者可能获取匿名化信息的其他数据处理者无法访问或不掌握允许匿名化后的数据重新识别到数据主体的额外信息，或不具备合法重新识别数据主体的手段，或重新识别数据主体需要不合理的时间、努力或资源，则不视为是可复原的；

2） 所选取的匿名化技术，以及匿名化处理的程度，需要能满足业务合法、可控的数据应用需求，促进数据要素的安全合规流通。

5.2 匿名化技术应用的原则

数据提供者开展匿名化处理活动，宜遵循如下的原则：

a） 主体权益保护：匿名化过程需要满足国家对数据安全和个人信息保护等的相关要求， 保障个人信息主体享有的法定权利，不应与个人信息主体的基本权利和自由相冲突；

b） 发挥数据价值：匿名化处理在满足合规合法的前提下，能够满足业务的合法、可控的数据应用需求，发挥数据价值；

c） 过程有序可控：匿名化的实施过程需要维护有限可控的数据利用环境秩序，并使得效果可评估、证据可保存；

d） 数据使用控制：匿名化的处理过程需要能通过技术方法实现对数据的使用控制。

6 匿名化技术应用的实施框架

匿名化处理的实施框架，如图1所示。

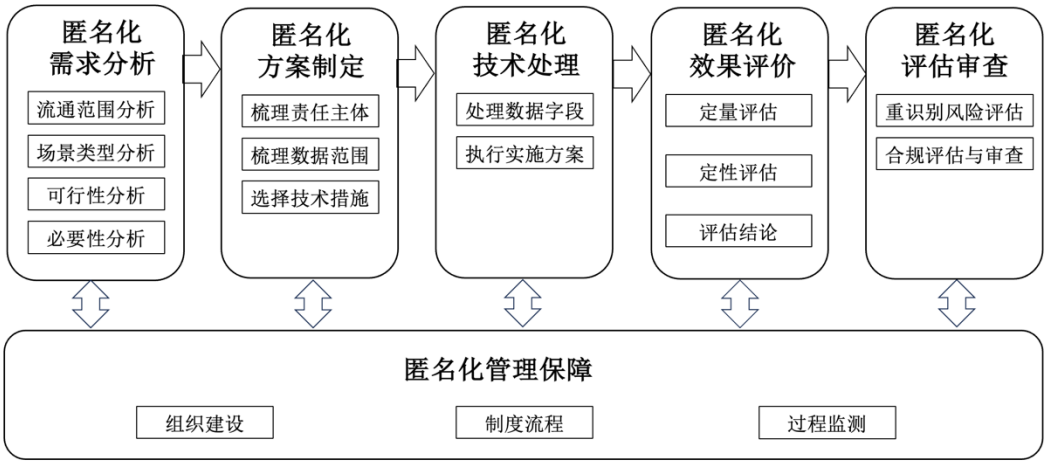

图1 匿名化技术应用的实施框架

其中，匿名化处理的实施框架主要包括匿名化需求分析、匿名化方案制定、匿名化技术处理、匿名化效果评价、匿名化评估审查等步骤。

匿名化需求分析主要包括流通范围分析、场景类型分析、可行性分析、必要性分析等步骤。

匿名化方案制定主要包括厘清责任主体、梳理数据范围、选择技术措施等步骤。

匿名化技术处理主要包括处理数据字段、执行实施方案等步骤。

匿名化效果评价主要包括K匿名效果评估模型、差分隐私评估模型、线性敏感度评估模型等方式。

匿名化评估审查主要包括重识别风险评估、合规评估与审查等。

在匿名化处理过程中，需对整个过程进行保障，包括组织建设保障、制度流程保障和过程监测保障等。

7 匿名化需求分析

7.1 流通范围分析

在数据流通场景中，数据的流通范围主要包括：

a) 组织内部的流通，比如公司内跨主体的数据流通；

b) 组织外部的流通，比如公司与某合作方之间的数据流通；

c) 对外公开，比如公司对外公开披露匿名化处理后的数据。

对于不同的数据流通范围，匿名化处理的要求和效果评价方法会有所不同。具体可参见第10章“匿名化处理效果的评估”的要求。

7.2 场景类型分析

在数据流通场景中，宜根据不同的场景类型，选用不同的匿名化处理技术。涉及到匿名化处理的场景类型主要包括：

a) 统计分析：数据处理者进行数据的挖掘分析时，将个人信息进行匿名化处理后再进行数据分析使用；

b) 开发测试：为了验证业务应用功能或业务算法模型，在系统测试、联调时，将个人信息匿名化处理后再进行开发测试使用；
c) 现网应用：在实时的现网应用场景中，可通过细粒度的访问控制，实现不同用户对同一个人信息访问时进行匿名化处理，实现脱敏展示；
d) 系统运维：运维人员在运维的工作中，直接连接生产数据库进行查询时，需要对查询的个人信息进行匿名化处理。

7.3 必要性分析

分析匿名化处理的必要性时，宜考虑如下方面：

a) 法律法规要求。业务侧宜结合国家、地区或行业的相关政策、法律、法规等规定，判断待收集、存储、使用、加工或向第三方提供的数据是否涉及匿名化的相关要求。
b) 监管设定要求。业务侧宜结合本行业、领域内相关监管部门要求，判断待收集、存储、使用、加工或向第三方提供的数据是否涉及匿名化的特别监管要求。
c) 履行约定承诺需要。业务侧宜结合自身是否与数据来源主体或者其他利益相关方存在数据匿名化处理的相关约定、声明或承诺，判断待收集、存储、使用、加工或向第三方提供的数据是否涉及匿名化的履约需求。
d) 业务经营策略需要。业务侧宜结合业务运行拓展过程中需要进行必要的数据统计分析、科学研究、训练测试等，综合考量数据特性和敏感程度，判断数据内外部应用时是否需要进行匿名化处理。

7.4 可行性分析

匿名化处理的具体实施可能会受限于数据类型、业务场景、技术成熟度、操作成本等因素，进行可行性分析宜考虑以下内容：

a) 待处理的数据属于结构化数据或静态数据，匿名化处理难度较小、成本较低，且风险可控；
b) 匿名化处理运行过程中，对相关信息系统不会造成不合理负荷，其数据存储及计算成本可控；
c) 数据经过匿名化处理，在现有技术能力下，满足标识符无法识别且不能复原的合规要求；
d) 数据经过匿名化处理，在具体业务场景下，不丧失数据的可用性，满足预设的业务用途要求。

业务侧经可行性分析后，可以考虑在受控环境下进行相对匿名化处理，以满足数据流通的需求。

8 匿名化方案制定

8.1 厘清责任主体

匿名化处理实施工作的相关责任主体宜包括：

a) 执行方：匿名化处理的发起者和策划者，通常是数据的控制者或管理者，负责确定匿名化的目标、范围和策略，并执行和实施匿名化处理过程，对匿名化处理的行为和结果负责。执行方通常是数据流通中的数据提供方；

b) 使用方：实际使用匿名数据的主体。使用方需要遵守相关承诺或协议，按照约定的用途和范围使用数据，并承担相应的责任，确保在使用数据的过程中不侵犯数据主体的隐私权益，并遵守相关的法律法规和行业规范；使用方通常是数据接收方；

c) 监督方：负责对匿名化处理过程进行监督和评估，确保数据处理过程合规、安全风险可控。可以对实施方的操作进行抽查或审计，对匿名化后的数据进行验证和测试，以确保数据的匿名性和可用性达到预期目标。监督方通常是组织的法务合规部门。

8.2 梳理数据范围

匿名化处理中，梳理数据范围的步骤宜包括：

a) 圈定后续业务所需数据范围。基于业务场景确定不同数据字段的匿名化需求，明确支撑具体业务所需处理的大致数据类型和范围，分析数据的性质、内容、格式、关系和体量等基本情况；

b) 移出无需处理的非敏感数据。为减轻匿名化操作的存储和计算成本，提升数据处理的效率，可以将不影响匿名化结果无需匿名化处理的数据集，先行移出至备用数据库表；

c) 移入可能关联识别的数据记录。抽取其他数据集中可能与待处理数据具有关联识别可能，或者其他可能影响匿名化效果的数据记录，一并纳入匿名化处理范围。

8.3 选择技术措施

在业务场景中，匿名化技术的选择，宜从可用性、时效性、安全性、合规性等方面进行综合评估：

a) 可用性：包括数据的特征保留、数据的真实性、数据的有效性等；

b) 时效性：包括匿名化过程的自动化执行、时间占用情况、计算资源的占用情况、存储资源的占用情况、成本等；

c) 安全性：包括抗重识别风险、可靠性、可控性等；

d) 合规性：包括对于法律法规的依从性等。

匿名化处理技术措施的选择要素，可参见附录A。

说明：匿名化和去标识化技术主要包括统计技术、密码技术、抑制技术、假名化技术、泛化技术、随机化技术、数据合成技术等；匿名化是去标识化的极端情况；安全隔离、数据沙箱、封闭域、数据专区、隐私计算、可信执行环境、联邦学习、同态加密、安全多方计算等属于构建数据流通环境的技术方案。

9 匿名化技术处理

9.1 处理数据字段

需要处理的数据字段宜包括：

a) 直接标识符：在特定环境下，可以单独识别个人信息主体的属性；

b) 准标识符：结合其他属性，可以唯一识别个人信息主体的属性；

c) 其他属性：直接标识符、准标识符之外的属性信息。

直接标识符和准标识符的识别，可参见GB/T 42460-2023 中附录A、附录B和附录C。

9.2 执行实施方案

执行匿名化技术实施方案的步骤宜包括：

a) 针对不同的属性，指定不同的去标识化处理措施，比如统计技术、密码技术、抑制技术、假名化技术、泛化技术、随机化技术、数据合成技术等；具体的去标识化技术措施，参见 GB/T 37964—2019 附录 A；
b) 针对不同的匿名化处理措施，设定不同的算法参数，比如掩码的位数，掩码的字符等。

匿名化处理技术的适用场景，可参见附录 B。

10 匿名化效果评价

10.1 评价目标

对于匿名化技术应用，以数据接收方在受控环境中合法合规的条件下，将重识别风险控制在可接受的范围内为主要目标。

数据流通过程分为加工过程和结果应用两个阶段，应主要评价加工过程中的匿名化。

10.2 评价方法概述

匿名化效果评价方法包括但不限于：基于K匿名的效果评价方法、基于差分隐私的效果评价方法、基于线性敏感度模型的评价方法等。可根据具体适用的场景，选用不同的效果评价方法，也可以选用其他的效果评价方法。

10.3 基于 K 匿名的效果评价方法

基于K匿名的效果评估方法，宜遵照如下实施指引：

a) 对于离线库表结构的数据集，使用K匿名进行匿名化效果评估。即经过匿名化处理后数据集中某条记录如存在K条重复记录，则说明该记录的匿名化处理程度达标；
b) 在组织外部数据流通场景中，数据集的K匿名值设置为大于等于5；
c) 在组织内部数据流通或数据长期存储场景中，数据集的K匿名值设置为大于等于3；
d) 在外部公开的场景中，数据集的K匿名值设置为大于等于20；
e) 对于进行假名化处理之后的直接标识符，不在K匿名的评价范围之内，即不用计算重复度。

具体评价方法和示例，参见附录 C 和附录 D。

10.4 基于差分隐私的效果评价方法

基于差分隐私的匿名化效果评价方法，宜遵照如下实施指引：

a) 对基于差分隐私的匿名化处理方式，基于差分隐私来进行匿名化效果评价；
b) 差分隐私的隐私度量模型，用在特定统计分析的设计中，来提供数学上的保证：无论某个数据主体是否包含在输入数据集中，这种分析结果的概率分布仅相差一个指定参数。这个指定参数可用于衡量每次分析输出时所遭受的“隐私损

失”。即使攻击者能访问其他相关数据集，“隐私损失”也能被控制在一个特定水平内。

具体评估方法可参见T/TAF 137—2022《基于差分隐私的用户个人信息保护技术要求》附录B。

10.5 基于线性敏感模型的效果评价方法

基于差分隐私的匿名化效果评价方法，宜遵照如下实施指引：

a) 对于宏观数据发布的场景，采用线性敏感模型来进行匿名化效果评估；
b) 采用线性敏感模型度量数据主体贡献接近程度；
c) 使用特定参数来确保对于一个真实值有足够大的距离，使得攻击者难以根据另一个数据主体的知识精确地估计任何特定数据主体的贡献。通常结合不同的线性敏感度措施以解决不同潜在风险。

具体可参考ISO/IEC 20889中10.4节的内容。

11 匿名化评估审查

11.1 重识别风险评估

组织宜基于匿名化处理效果来识别重标识风险，并形成评估报告。其步骤宜包括：

a) 评估匿名化处理效果：在数据流通前，业务部门对匿名化处理的效果进行自评，判别重识别风险，并形成自评记录或报告，包括实施团队、过程行为、数据对象、数据结果等，评估过程和结果宜形成记录或报告；
b) 评估数据用途：业务部门管理层需要依据数据发布共享用途、重标识风险、数据有用性最低要求等因素，以及验证结果、匿名化各步骤实施过程中的监控审查纪律等因素，做出是否认可数据匿名化处理结果的决定；
c) 业务部门宜根据情况变化或定期进行重标识风险评估，并与预期可接受风险阈值进行比较，以保障个人信息安全性。情况发生变化是指重标识风险的相关要素发生变化，相关要素包括但不限于：数据使用者、目标信息系统、目标信息安全环境、新增数据处理行为、新增匿名数据；
d) 通过匿名化实施评估后，如涉及后续传输、提供、存储、加工等行为，控制数据内容与匿名化实施评估保持一致。业务部门不得在通过评估后另行再增加关联其他标识符或者技术回滚等导致与评估情况不符的不合规行为，如出现该情况应重新评估。

11.2 合规评估与审查

数据流通参与方宜建立匿名化合规管理机制，采取必要的技术、合规及管理措施、在风险相对可控的前提下开展数据使用流通。具体措施宜包括：

a) 制定匿名化技术实施方案并实施；
b) 限定数据处理环境，控制环境内的数据处理行为；
c) 对数据导出等高危行为加强审核；
d) 数据流通的数据提供方和数据接收方清晰界定各自的权责边界，通过协议或合同明确各自的权利和义务，以约束数据流通行为；

e） 数据接收方建立抽检机制，检验匿名数据的合规情况，并采取相应补救措施（如删除或重新匿名化处理等）。

12 匿名化管理保障

12.1 组织建设

组织机构在匿名化处理的组织建设方面，宜遵守如下的实施指引：

a） 设立专门负责个人信息的团队或岗位，并由组织高级管理层负责；
b） 明确相关人员在匿名化处理流程中的职责和权力；
c） 定期对员工开展个人信息保护和匿名化处理相关的培训；
d） 明确组织机构对匿名化工作的管理要求。

12.2 制度流程

组织机构在匿名化处理的制度流程方面，宜遵守如下的实施指引：

a） 建立匿名化处理过程中标识符识别和处理的规则库，供组织内部参考实施；
b） 建立匿名化处理过程中合规评估的流程与要点；
c） 建立匿名化处理效果评价的方法与流程；
d） 建立匿名化处理的控制体系，通过限定和控制数据使用中的数据内容、处理形式和约束条件，形成受控环境和安全边界，通过对关联的控制而约束数据的合规使用，并有效留存相应的合规性证据。

12.3 过程监测

组织机构在匿名化处理的过程记录方面，宜遵守如下的实施指引：

a） 对匿名化处理的步骤进行监测，并记录日志；
b） 日志内容包括时间、处理人、处理措施、处理结果等内容；
c） 在匿名化处理的步骤完成时，对相关记录进行不定期抽查，检查输出记录是否齐全和内容完备；
d） 及时发现已经出现或可能出现的错误或偏差，并采取适当控制措施，监督各步骤执行过程得到完整和有效地执行。

附 录 A
（资料性）
匿名化处理技术措施的选择要素

在业务场景中，匿名化技术的选择，宜从可用性、时效性、安全性、合规性等方面进行综合评估。

A.1 可用性

可用性的评估，主要包括如下几个方面：

a) 特征保留：匿名化后的数据尽可能地体现和保持原始数据的原有特征，且尽可能多的保留原始数据中的有意义的信息，以减小对使用该数据的系统的影响；
b) 真实性：匿名化后的数据，与原始数据的符合度；
c) 有效性：保证数据的开发、测试、使用等过程中不会受到匿名化的影响。

示例：

a) 可用性相对较高的技术包括：同态秘密共享、确定性加密、保序加密、同态加密等。
b) 可用性相对较低的技术包括：数据聚合、抑制技术、统计技术、随机技术、泛化技术、数据截断等。

A.2 时效性

时效性主要包括如下几个方面：

a) 自动化执行：保证匿名化的过程可通过程序自动化实现，可重复执行；
b) 时间占用：匿名化处理过程所占用的时间；
c) 计算资源的占用：匿名化处理过程所占用的计算资源；
d) 存储资源的占用：匿名化处理过程所占用的存储资源；
e) 成本：匿名化处理过程所需要的成本。

示例：

计算性能相对较低的技术包括：同态加密、同态秘密共享等。

A.3 安全性

安全性主要包括如下几个方面：

a) 数据使用环境：数据使用环境为公开环境、受控环境等；
b) 抗重识别风险：保障匿名化后的数据重识别的风险在可接受的范围内；
c) 可靠：保障匿名化算法不被同质属性、概率、知识推断等手段攻击，确保匿名化技术安全可靠；
d) 可控：匿名化的数据仍需采取合适的方式控制知悉范围，通过恰当的安全管理手段，防止数据外泄。避免因扩散范围过广导致的多源串联交叉比对分析导致风险增加。

示例：

相对而言，抗重识别的技术包括：数据聚合、同态加密、同态密码共享等。
易受重识别攻击的技术包括：确定性加密、保序加密、抑制技术等。
假名化技术容易受到对假名分配表的攻击。

采用了隐私计算环境、可信执行环境、数据沙箱、封闭域等受控的数据流通环境时，安全性相对较高。

A.4 合规性

合规性主要体现在对《个人信息保护法》等法律法规的依从性等方面。

附 录 B
（资料性）
匿名化处理技术的适用场景

B.1 数据统计分析场景

B.1.1 场景描述

数据统计分析场景指数据分析工程师进行数据挖掘分析时，将敏感数据脱敏后再进行数据分析使用。将数据导出到数据文件或者大数据平台，来进行分析。

B.1.2 场景特点

在数据分析场景下，需要重视数据之间的关联性、分析结果，脱敏后的数据应保留原有的数据关系与格式，确保数据脱敏后不会影响分析结果。

B.1.3 技术方案

宜采用泛化、均化等技术。

B.2 系统开发测试场景

B.2.1 场景描述

数据从生产环境导出到测试环境时，一般是为了验证业务应用功能或业务算法模型。是指系统测试、联调时，将敏感数据脱敏后再进行开发测试使用。

B.2.2 场景特点

为了防止敏感数据泄漏，在满足测试环境业务验证的前提下，只提供保持最小化数据特性的匿名数据。在开发测试场景下，需要重视数据的可用性，因此匿名化处理可以采用相同含义的数据替换原有的敏感数据。

B.2.3 技术方案

宜主要采用替换、变形、扰乱、泛化、均化、时间随机、数值随机等技术。

B.3 系统日常使用场景

B.3.1 场景描述

系统的日常使用场景主要包括：

a）数据从生产环境导出到租户环境：一般是用于统计分析和数据挖掘；

b）用户访问生产环境中的数据：为实现敏感数据保护，可通过细粒度的访问控制，实现不同用户对同一敏感数据访问时进行不同的展示，对低权限用户进行匿名化展示；

c）运维人员访问生产环境中的数据：运维人员访问生产环境中的数据是为了验证平台或业务的正常运行，不需要获取真实的敏感数据。为防止敏感数据泄漏，只需提供为运维人员匿名化处理后的数据。

B.3.2 场景特点

系统的日常使用场景的特点主要包括：

a） 对于用户访问生产环境中的数据的场景，业务系统访问时效性要求较高；
b） 对于数据从生产环境导出到租户环境，匿名化操作发生在数据批量迁移时，时效性要求低；
c） 运维人员访问生产环境中的数据的场景，访问业务系统时效性要求较高。

B.3.3 技术方案

对于数据访问的场景，宜采用不可逆脱敏算法对数据进行匿名化展示，防止运维人员接触敏感信息。

附 录 C
（规范性）
基于 K 匿名的匿名化效果评价方法

C.1 概述

对于离线库表结构的数据集，宜使用K匿名进行匿名化效果评估。即经过匿名化处理后数据集中某条记录如存在K条重复记录，则说明该记录的匿名化处理程度达标。

K匿名是计算数据集重标识风险水平的一种简单方法。根本上说，它是指在一个数据集中可以分组在一起的最小数量的相同记录。在评估数据集的总体重标识风险时，通常采用最小的组来表示最坏的情况。K匿名值为1意味着该记录是唯一的。

K匿名值越高意味着重标识风险越低，相反，K匿名值越低意味着风险越高。通常而言，应该尽可能设置更高的K匿名阈值，以最大限度地减少任何重标识风险。

度量方法:计算数据集中每个等价类的记录数,确保每个等价类至少包含K个记录。K值越大，隐私保护程度越高。基于K匿名值进行定量评估时，建议结合场景系数和环境系数来进行评估。场景系数表示数据匿名化后使用场景的安全系数，如组织内部使用、向特定第三方提供、公开发布等；环境系数表示数据流通时，数据流通环境的技术保障能力和管理保障能力。

通常而言，在外部数据共享场景中，应设置尽可能高的K匿名值（例如：5或更大）；而内部共享和数据长期存储场景的K匿名值可以更低（例如：3）。

C.2 计算方式

对于离线库表结构的数据集，建议基于K匿名进行效果评估。建议采用如下的计算方式来计算匿名化程度：

匿名化程度 = 数据集K匿名值＊场景系数 ＊ 环境系数

要求：匿名化程度>=1，匿名化程度越高越好。

其中：

a) 数据集K匿名值：表示数据集中，经过匿名化处理后，具备相同的准标识符字段组合的记录的条数的最小值；
b) 场景系数：表示数据集被使用的具体场景，如组织内部流通、组织外部流通、对外公开等；
c) 环境系数：表示数据流通时，数据流通环境的技术保障能力和管理保障能力。

基于K匿名的效果评价的示例，可参考附录D。

C.3 K 值计算

对于给定的离线库表结构的数据集，定量评估主要是对K匿名的最小K值的计算。计算方式如下：

a) 先识别出直接标识符和间接标识符;具体可参见参考GB/T 42460-2023附录A、附录B和附录C。由于评估方与匿名化处理方通常不同，因此需要重新识别；
b) 将数据集中的直接标识符假名化、加密或剔除；直接标识符假名化或加密之后，作为敏感属性来处理；
c) 针对匿名化处理后的间接标识符的组合，找出数据集中所有的等价类；即数据集中与某个数据项具备相同的间接标识符的数据项的集合；

d) 针对每个等价类，找出等价类大小，即K匿名中的K值；
e) 找出数据集中的K值。

C.4 对场景系数的评估

定量评估还包括场景系数的设定。K匿名值的场景系数需要根据不同的应用场景，进行设置。

通常而言，在外部数据共享场景中，应设置尽可能高的K匿名值（例如：5或更大）；而内部共享和数据长期存储场景的K匿名值可以更低（例如：3）。

建议的场景系数，如下表C.1所示：

表C.1 场景系数建议

场景	建议的场景系数
组织机构内部的数据流通	1/3
组织机构外部的数据流通	1/5
对外公开	1/20

注：在实际使用时，可根据具体的使用场景，适当调整。

C.5 对环境系数的评估

定性评估主要是对环境系统的评估。环境系数包括对技术保障能力和管理保障能力的综合评估。

a) 技术保障能力，具体包括：
   1) 安全和隐私控制能力。比如，是否采用权限管理、访问控制策略等；
   2) 数据流通技术保障能力。比如，是否采用安全隔离、数据沙箱、封闭域、专区、隐私计算等技术；
   3) 重识别攻击的动机和能力。比如，是否采用技术手段，抗重识别攻击等。
b) 管理保障能力：包括组织建设、制度流程、人员能力、协议合同约束、审计机制、事件与应急管理、风险控制能力等管理保障能力的综合评估。

对于环境保障能力，建议采用定性评估的方式。环境保障能力的中间值建议为1。当环境保障能力较低时，建议调高对于K匿名值的要求，以提升整体的匿名化程度。

C.6 形成评估结论

基于定性评估中的环境系数，和定量评估的K值和场景系数结果，使用C.2节中的计算方式，可以形成评估结论，评定给定的数据集是否满足匿名化条件。

对于匿名化程度评价的结果，可以有如下的运用方式：

a) 可以对整个数据集给出评价值，看是否满足预定的匿名化程度要求；
b) 针对数据流通场景，可以先设定出K匿名值，根据评价结果，剔除出低于预定K值的数据项；
c) 可以对低于预定K值的数据项继续作匿名化处理（如泛化、抑制等），直到数据集中的数据项全部满足匿名化程度要求。

附 录 D
（资料性）
基于 K 匿名的匿名化效果评价示例

D.1 数据集

本附录参考GB/T 42460—2023《信息安全技术 个人信息去标识化效果评估指南》中附录D.2中的数据集。

某组织内部共享的一批用户的广告投放记录数据集，已经对姓名、年龄等属性进行去标识化处理，如下表D.1所示。

表D.1 某组织内部的去标识化数据集

性别	年龄	业务编码
男	36 ~ 40	700225
女	36 ~ 40	355421
男	51 ~ 55	355611
男	36 ~ 40	455641
女	46 ~ 50	355421
男	41 ~ 45	255456
男	51 ~ 55	355421
男	36 ~ 40	756987
女	36 ~ 40	700227
男	51 ~ 55	379044
女	36 ~ 40	455641
男	41 ~ 45	355459
女	46 ~ 50	700225
男	41 ~ 45	487792
女	46 ~ 50	437562
男	51 ~ 55	736920

D.2 计算 K 匿名值

针对上述的数据集，计算等价类的大小，即K值。如表D.2所示：

表D.2 数据集中的等价类大小

等价类	准标识符		等价类大小（K值）
	性别	年龄	
1	男	36 ~ 40	3
2	男	41 ~ 45	3
3	男	51 ~ 55	4
4	女	36 ~ 40	3
5	女	46 ~ 50	3

可见，在上表中，该数据集的K匿名值为3。

D.3 场景系数评估

由于该数据共享的场景是该组织在其组织内部的数据共享，其场景系数为1/3。

D.4 环境系数评估

经评估，该组织内部的技术保障能力和管理保障能力较高，其环境系数为1。

D.5 评估结果

基于10.3节的评价方法，匿名化程度 = 3 * 1/3 * 1 = 1，满足匿名化要求。

附 录 E
（资料性）
大语言模型场景中的匿名化应用案例

E.1 场景概述

通用大模型训练阶段，需要使用大量的数据集开展模型预训练，以提升模型性能或者为模型开展安全对齐，保证模型能力提升和输出安全。预训练阶段使用的数据类型可能涉及涉及公开数据集，专业数据集等无标注数据。

E.2 匿名化需求

大模型场景涉及大量非结构化文本数据，为符合法律法规要求，需要对其中涉及个人信息进行匿名化处理。通过人工分析法，在对业务处理、数据集结构、相互依赖关系和可用数据等要素分析的基础上，综合判断数据集重标识风险，参考《网络安全法》《数据安全法》《个人信息保护法》《信息安全技术 个人信息安全规范》（GB/T 35273-2020）等现行有效的法律法规及国家标准中的定义与示例，识别目标数据集中最有可能存在的典型个人信息类型。

鉴于个人信息类型宽泛，客观上技术无法完全保障所有个人信息均已识别及排除，企业可以根据数据内容识别出最有可能存在的典型个人信息类型，可参考的重点范围如下：真实姓名、用户ID、邮箱、电话号码、用户IP地址、身份证信息、车牌号、银行卡号等。

E.3 技术方案

技术方案的流程如下：

a）大语言模型客户端对用户提交的文本进行命名实体识别。

数据集中包含敏感信息的待处理文本，如示范文本：

“如果你对婺源的印象还只停留在上半年的油菜花，那么你就错过了下半年的一大美景——枫叶！秋天的婺源另有一番美艳，尤其 10 月底至 11 月，满山争艳的红叶，与灰白的民居交相辉映，让无数摄影发烧友如痴如醉！ 欢迎随时咨询，0755-82233606 美行国际摄影文化俱乐部旅游摄影，中国深圳旅游摄影俱乐部美行摄影会员火热召集中咨询方式：【1】请添加微信 czwphoto888 或 13603063441 中国深圳旅游摄影俱乐部；【2】电话报名：姓名张三 18938040678；姓名李四 13603063441；【3】现场报名：地址：深圳市罗湖区笋岗东路百汇大厦北座 2101 室美行国际摄影文化俱乐部旅游摄影【4】电子邮箱：szzqm@126.com【5】新浪微博：美行摄影。”

本步骤采用命名实体识别(Named Entities Recognition, NER)算法提取命名实体及其类型。以上述提交文本为例，提取的结果如下：

人物: ["姓名张三", "姓名李四"], "

社交账号: ["czwphoto888", "szzqm@126.com"],

地点: ["婺源", "深圳市罗湖区笋岗东路百汇大厦北座 2101 室"],

"电话": ["0755-82233606", "18938040678", "13603063441"],

"邮箱": ["szzqm@126.com"],)

b）大语言模型客户端生成一次性实体替换表。

对步骤 a）中涉及的命名实体进行同类型替换或用“*”替换。替换时从同类型的所有实体词中随机选取一个或直接用“*”，且保证文本的逻辑结构、正负面情绪等不变。以示范文本中的命名实体为例，所随机得到的一个替换表如下：

"姓名张三", "姓名李四" → 小李、小王或“*”，“*”

婺源 → 杭州或“*”

0755-82233606 → 021-12345678 或“*”

czwphoto888 → photo123 或“*”

13603063441 → 13987654321 或“*”

18938040678 → 18812345678 或“*”

深圳市罗湖区笋岗东路百汇大厦北座 2101 室→广州市天河区天河北路天汇大厦北座 2101 室或“*”

szzqm@126.com → user1@mail.com 或“*”

c）大语言模型客户端按照实体位置进行对位实体替换：

根据步骤 b）得到的替换表对原文进行对位实体替换。

以示范文本为例，对位替换后得到的脱敏文本如下：

“如果你对（杭州/“*”）的印象还只停留在上半年的油菜花，那么你就错过了下半年的一大美景——枫叶！秋天的杭州另有一番美艳，尤其 10 月底至 11 月，满山争艳的红叶，与灰白的民居交相辉映，让无数摄影发烧友如痴如醉！ 欢迎随时咨询，（021-12345678/“*”）美行国际摄影文化俱乐部旅游摄影，中国广州旅游摄影俱乐部美行摄影会员火热召集中咨询方式：【1】请添加微信（photo123/"*")或(13987654321/"*")中国广州旅游摄影俱乐部【2】电话报名：（18812345678 小李/“*” “*”）；（13987654321 小王/“*”*“）【3】现场报名：地址：广州市天河区天河北路天汇大厦北座 2101 室美行国际摄影文化俱乐部旅游摄影【4】电子邮箱：（user1@mail.com/”*“）【5】新浪微博：美行摄影”。

E.4 指定脱敏类型的示例

在大语言模型中，可以灵活性指定脱敏的数据类型。比如，指定脱敏类型为联系方式的示例如下：

处理前数据样例（原始数据）	您好！近期我在贵部门的在线系统中发现我的个人信息（如住址、联系电话等）尚未更新。为了确保后续业务能够顺利办理，我想向贵部门咨询并确认以下事项： 1）.个人信息更新进度 • 我于 2025 年 1 月 10 日通过线上平台提交了个人信息更新申请，请问该申请目前的处理进度如何？ • 若仍在审核中，预计还需要多长时间才能完成更新？ 2）.补充材料或手续 • 在审核过程中，是否还需要我提供其他补充材料？

	• 例如收入证明、学历认证、房屋租赁合同等。若需要，请告知具体要求以及提交方式。 3）.联系方式更新 • 我在申请更新时也提交了新联系方式（手机号码：13812345678，邮箱：zhangxiaoming@example.com），请问相关系统已完成同步更新吗？如果还未更新，需要我做哪些操作来加快处理？
匿名化处理对象	手机号码、邮箱
处理后数据样例（匿名化）	大模型将只对指定的类型进行限定范围内的脱敏，例如识别到原文中所有的联系方式如下： "联系方式": ["13812345678" "zhangxiaoming@example.com"] 然后进行脱敏替换： 13812345678 -> 13987654321 或”*“ zhangxiaoming@example.com -> user1@mail.com 或”*“ 同时保证其他类型，如人名、身份证号、地点等没有被指定的类型不被替换。

E.5 技术优势

采用大模型进行匿名化处理，实现个人信息脱敏的技术优势如下：

1）.可以通过自然语言描述灵活指定待脱敏的类型，没有被指定的类型不会被脱敏以保证数据的可用性；

2）.支持将脱敏对象替换为同类型但不同的实体，不影响后续对脱敏后数据的使用。

E.6 管理措施

根据大模型训练等实际需要，评估匿名化需求，划定需经匿名化处理的必要数据范围，并对评估过程和结果进行留档。参照本指南要求抽查核验经处理的数据，以核验匿名化效果。如出现不符合要求的数据字段，则针对该类字段重新制定标注清洗规则，以符合要求。

参 考 文 献

[1] 《中华人民共和国网络安全法》 2016 年 11 月 7 日第十二届全国人民代表大会常务委员会第二十四次会议通过

[2] 《电信和互联网个人信息主体个人信息保护规定》 2013 年 7 月 16 日中华人民共和国工业和信息化部令第 24 号公布，自 2013 年 9 月 1 日起施行

[3] GB/T 35273—2020 《信息安全技术 个人信息安全规范》

[4] 中国信息通信研究院产业与规划研究所、北京国际大数据交易所 数据清洗、去标识化、匿名化业务规程（试行）

[5] T/CAAAD 004—2022 T/CCSA 424—2022 互联网广告 数据匿名化实施指南