中外法学 Peking University Law Journal
Vol. 36, No. 2(2024)pp. 326－345

个人信息匿名化的理论基础与制度建构

赵精武*

摘 要 匿名化技术作为兼顾数据利用和数据安全的技术始终难以真正落地，原因在于《个人信息保护法》第73条对“匿名化”概念设置了“无法识别”和“不能复原”两个限定条件，这一苛刻要求使得匿名化制度处于“存而不用”的状态。虽然匿名化处理无法实现完全匿名已成共识，但是对于相对匿名的判断标准以及重新识别风险的限定范围始终存有较大争议。现阶段匿名化制度的理论探讨普遍忽视了对该项制度的理论基础探究，而将解决思路限定在个体权益保护的范畴。在数据关系理论范式下，这种信息要素关联性指向的是横向数据关系层面的群体性特征，而这种群体性特征恰恰也是数据经济价值的关键所在。因此，匿名化制度的功能定位应当从单纯地保障个体信息私密性转变为在保障数据关联经济价值的基础上降低重新识别风险。在制度建构层面，匿名化信息所要求的“不能复原”应当被解释为经由事前的风险评估，复原的技术难度和时间成本远超一般主体所能接受的范围。

关 键 词 匿名化 数据关系理论 再识别风险 个人信息保护

一、问题的提出

在推动数据要素市场化配置的大背景下，个人信息如何安全地商业化利用成为一个重要问题。虽然《个人信息保护法》将匿名化处理的信息排除在个人信息范畴之外，为个人信息商业化利用提供了合法性的判断标准，但其第73条却将“匿名化”技术效果界定为“无法识别特定自然人”且“不能复原”。此种规定实际上与主流观点所认可的匿名化技术效果相去甚远，因为在信息技术创新发展的当下，所谓的完全匿名仅仅可能存在于理想的技术场景中，实践中只

* 北京航空航天大学法学院副教授。本文系国家社科基金重大项目“信息法基础”(项目编号：16ZDA075)的阶段性研究成果。

能实现相对匿名化的处理效果。如果严格按照文义解释的方法,《个人信息保护法》第73条所界定的匿名化无法在商业实践中找到与之匹配的技术方案。为了避免将匿名化制度的实际功能和价值架空,当下最稳妥的方案便是将“不能复原”进行补充解释,亦即匿名化处理的个人信息无法被复原或者复原的成本和难度远远超出复原之后所能获得实际收益。如此一来,匿名化制度的理论难题似乎得到了彻底解决。然而,问题远未就此止步,因为既然普遍承认法律所规定的匿名化处理是一种相对匿名化效果,那么这种“相对性”当如何解释仍有疑问。如果仅仅停留于按照具体的匿名化处理场景进行“具体问题具体分析”式的探讨,那么匿名化制度的适用问题并没有得到彻底解决,只不过是将一个概念模糊的问题转变为一个标准模糊的问题。国内外学者对于匿名化存在诸如“功能性匿名化”“主客观匿名化”以及“相对不可识别说”等解释方案,这也导致在制度层面迟迟难以正面回应何种匿名信息可以不受《个人信息保护法》调整。如此看来,匿名化制度的相关问题远没有就此得到妥善解决,尤其对于监管机构而言,零风险的偏好使其难以对匿名化技术保持信任。理想中的匿名化制度建构应当是一种风险可控的相对匿名化,即通过相应的匿名化处理能够有效切断个体信息与其他存在社交关联个体信息之间的关系。因此,为了正面回答匿名化信息的基本概念、匿名化效果以及重新识别风险等判断难题,同时澄清匿名化制度是否涉及再识别风险概率计算的问题,有必要从匿名化制度的基础理论出发,明确我国匿名化制度的内在逻辑和建构方向。

二、匿名化制度的学说争议澄清

(一)国内匿名化制度的理论共识与分歧

《个人信息保护法》颁布前后,学界曾就匿名化制度的理论基础和具体内容兴起过一阵讨论热潮,但是伴随着个人信息保护相关配套制度的出台,相关探讨反而有所衰减。这并非因为后续的配套制度已经基本解决了数据安全和数据利用的平衡问题,而是因为学界所达成的共识以及难以调和的分歧使得匿名化制度的相关研究陷入瓶颈。隐私计算技术的出现更是直接以“数据可用不可见”的技术方案解决了数据安全和数据利用的平衡问题,继续探讨匿名化制度的现实意义似乎有所不足。当然,此种观点混淆了匿名化技术与隐私计算技术之间的概念关系,隐私计算技术处理数据的性质依然是匿名化处理,〔1〕匿名化技术并不是一项内容固定、方式单一的数据处理技术。

在共识层面,学界普遍承认个人信息的匿名化是相对的,因为现有的数据分析技术以及不同渠道的数据公开或泄露均有可能导致匿名化信息再识别特定自然人的可能性,没有任何技术能够实现完全匿名化的技术效果。〔2〕此外,这种再识别风险客观上难以进行量化计算,并

〔1〕 参见方竞、周雍恺、卞阳等:“数据基础制度下隐私计算的实践与思考”,《信息通信技术与政策》2023年第4期,第52—53页。

〔2〕 参见张建文、程海玲:“‘破碎的隐私承诺’之防范:匿名化处理再识别风险法律规则研究”,《西北民族大学学报(哲学社会科学版)》2020年第3期,第77—78页。

且在很大程度上受到匿名化处理过程、技术方案以及应用场景的影响。即便有学者从技术层面对匿名化再识别风险进行整体性评估，意图划定再识别风险的概率区间，〔3〕但这种区间评估模式客观上无法解释个人信息处理者的匿名化技术标准，即究竟是按照区间下限的标准进行匿名化处理，还是只要保证匿名化再识别风险不超过最大值即可。基于“匿名化是相对的”和“再识别风险量化是不具可操作性的”两种共识，匿名化制度的理论探讨进入到下一个阶段：以风险可控作为匿名化处理的法定标准。既然再识别风险概率难以精准计算，并且该风险也不可能彻底根除，那么倒不如按照风险管理的基本逻辑解释个人信息处理者是否达到符合法定要求的匿名化处理效果。这种“基于风险”的治理思路强调应当在数据处理全生命周期的各个环节均采取预防再识别的保障措施，〔4〕并辅之以相应的安全技术标准，最大限度地降低再识别风险的发生概率。遗憾的是，学界有关“风险可控”的相关探讨大多是在制度建构层面予以回应，主流观点所提及的“设置禁止再识别义务”〔5〕“设置防范再识别风险义务”“设计匿名化行为准则”“设置再识别风险评估机制”〔6〕等主张并未能在可操作性层面解释这些机制能够实现何种程度的“风险可控”。

在分歧层面，鉴于再识别风险的不可根除性以及数据挖掘能力的跃迁式发展，部分学者开始质疑匿名化制度在个人信息保护立法框架下是否仍有讨论的必要性。因为《个人信息保护法》第 73 条所要求的“不能复原”和“无法识别”导致实践层面的匿名化处理技术无法使用，这使得个人信息保护义务履行机制陷入两难境地：一方面，倘若企业对个人信息进行匿名化处理，但由于匿名化后的信息仍然存在能够复原的可能性，故而匿名化后的信息仍然属于个人信息，企业采取匿名化处理技术似乎“多此一举”；另一方面，倘若企业真正实现了“不能复原”的匿名化处理，那么其采取的匿名化方案显然是删除、替换或加密大量的识别符或准识别符，由此导致数据的商业效用丧失，企业匿名化处理的商业目的无法达成。这种两难境地正在逐渐淡化匿名化制度的理论价值与现实意义，故而有部分学者提出改变“因匿名化后的信息不属于个人信息而可以自由处理”的立法逻辑，明确企业仍然需要对匿名化后的信息承担后续必要的安全保护义务。〔7〕但这种改良方案又会导致新问题——企业应当对匿名化后的信息承担何种程度的个人信息保护义务？在现有技术不可能达成完全匿名化的背景下，以“危险制造者”的归责逻辑要求企业继续对匿名化后的信息承担保护义务看似符合权利义务对等的基本内

〔3〕 参见张艳、王璐瑶：“政府数据开放场景下匿名化数据的再识别风险防范”，《电子政务》2024 年第 1 期，第 89 页。

〔4〕 参见张涛：“欧盟个人数据匿名化治理：法律、技术与风险”，《图书馆论坛》2019 年第 12 期，第 100 页。

〔5〕 参见齐英程：“我国个人信息匿名化规则的检视与替代选择”，《环球法律评论》2021 年第 3 期，第 64 页。

〔6〕 参见张涛：“大数据时代个人信息匿名化的规制治理”，《华中科技大学学报(社会科学版)》2019 年第 2 期，第 82—83 页。

〔7〕 参见李润生：“个人信息匿名化的制度困境与优化路径——构建‘前端宽松＋过程控制’规制模式之探讨”，《江淮论坛》2022 年第 5 期，第 117—120 页。

涵，但忽视了匿名化处理的根本目的正是以非个人信息的形式最大化数据利用效率。更麻烦的是，再识别的发生往往并不是单纯因为个人信息处理者匿名化处理不符合法定要求或技术标准所导致的，而是受到“其他具有关联性数据集合的公开和使用”“第三方恶意识别”等外部因素的影响。〔8〕暂且不提匿名化信息的后续安全保护义务等类似主张存在二次划分义务履行标准的偏差逻辑，仅仅是在因果关系和过错要件层面就难以证成企业故意导致匿名化后的信息被再识别事件的侵权责任。

（二）国外匿名化制度研究的迟滞原因

国外有关匿名化制度的研究兴起于个人隐私保护领域，最早是与人口统计活动相关。这些研究普遍是以匿名化处理统计数据为范例，论证如何减少统计数据公开后可能造成的隐私泄露风险。并且，国外学者普遍将匿名化处理视为一种隐私保护过程，而不是一种纯粹的安全保障技术。基于此种立场，匿名化处理的理论研究也与具体的应用场景、数据处理目标等因素相关，故而也延伸出匿名化再识别风险是否可以量化评估的学术争议，如从统计学角度测算特定数据库经匿名化处理后的再识别风险低于0.1%，以此证明“公平的匿名化”确实存在。〔9〕在立法层面，学者们常将欧美中的匿名化法律标准模式分别总结为“穷尽所有可能性标准”“隐私权类型保护标准”以及“不可识别和复原标准”。〔10〕

欧盟作为匿名化技术应用的典型范例，其公布的匿名化技术标准和监管要求常被学者们予以引用与分析。而美国则倾向于在各个行业领域规定各自的“去标识化”技术要求，如《健康保险流通和责任法》（Health Insurance Portability and Accountability Act，HIPAA）第164.514(b)条规定“受保护健康信息的去标识化标准”，《加州隐私保护法案》（California Consumer Privacy Act，CCPA）第1798.140(m)条规定的“去标识化”（deidentified）。如若按照欧盟相关规范性文件的公布时间为节点，国外匿名化制度研究大致可以分为三个阶段：第一个阶段是早期匿名化研究，多以个人隐私保护为重心。第二个阶段则是以欧盟第29条工作组在2014年发布的《关于匿名化技术的意见》为起点，围绕匿名化技术的实际效果开始了形式主义与实用主义之间的争论。第三个阶段则是以欧盟《通用数据保护条例》（General Data Protection Regulation，GDPR）在2018年正式生效为起点，剖析欧盟模式的匿名化机制弊端，匿名化制度的研究重心转至如何在法律与技术交叉视角下解决再识别风险问题。在司法实践中，域外法院大多将再识别风险的认定标准限定为“采用所有可能合理措施方可识别”，如在英国的Common Services Agency v. Scottish Information Commissioner一案中，法院认为再识别的可能性应当考虑识别主体可能获取的所有外部信息并加以判断。〔11〕这类标准看似解决了再识别风险问题，但本质上依然未能为产业实践提供更为操作可行的具体标准或规范化业务

〔8〕 参见郑佳宁：“数据匿名化的体系规范构建”，《政法论坛》2022年第4期，第62—64页。

〔9〕 Zoltan Alexin, “Does Fair Anonymization Exist,” *International Review of Law, Computers and Technology*, Vol. 28, No. 1, 2014, pp. 40-41.

〔10〕 参见李悦、陈秋竹：“‘数据管税’背景下纳税人信息匿名化的法律标准探讨”，《税务与经济》2020年第4期，第75—76页。

〔11〕 Case [2008] UKHL 47, Judgments of The Lords of Appeal, 1 and 2 APRIL 2008, para. 17, 21.

流程。

在上述三个阶段中,类似于我国的“匿名化是否仍有实际意义”的探讨同样存在,尤其是在第二个阶段,国外学者逐渐发现有关匿名化的学理争论已经停滞不前,论证重心侧重于因“正确的外部信息而被去匿名化”〔12〕而存在的风险,相关研究活动对个人隐私保护的促进作用微乎其微。有学者对现有匿名化的争论进行了总结:因为辅助信息的存在,使得匿名化后的信息仍然存在再识别的风险,这也导致了两种针对匿名化是否仍有实践意义的对立学说——实用主义者认为在实践中,所谓的辅助信息并不是那么容易获得,却忽视了“数据主体是可区分的”和“数据主体能够被再识别”两个概念,其论证逻辑是通过预设各种场景去量化具体的再识别风险,其结果也是倾向匿名化仍然对隐私保护具有重要意义。形式主义者则认为量化匿名化的有效性或再识别风险是不科学的,因为这种量化的前提基础是预设匿名化破解主体会选择何种技术方案,其结论是不应当过高地期待匿名化在隐私保护领域的预期作用,重要的是如何尽可能降低再识别风险。关于这两种学说争论的根源,有学者言简意赅地指明是“学科视角的差异性导致的”,因为形式主义者对于量化技术有效性或风险概率的出发点是以“数学的严谨性”为基础,但其结论也因此缺乏相应的实践可操作性;而实用主义者则是将统计学的方法应用于具体场景,虽具有可操作性却面临着适用范围的局限性。〔13〕这阶段的匿名化制度的治理逻辑可以归结为“基于损害”“基于风险”和“基于流程”以及三者融合的中间路径,〔14〕但这些治理逻辑还是将再识别风险作为核心内容。

事实上,形式主义和实用主义的学说之争一直持续到第三个阶段也未能解决,这两种学说背后的学科视角差异也引发了学者们对方法论的审视与反思,进而开始寻求法律与技术结合的解释工具。事实上,国内外研究也均存在数据安全立法与信息技术创新之间的发展不同步问题,这种不同步的根源除了法律无法客观预测技术发展走向之外,还表现为法律与技术针对同一客体的概念基础存在显著差异。基于这种考虑,国外学者主张建构技术与法律规范的混合概念,并以欧盟 GDPR 序言第 26 条提及的“筛选”(singling out)为例。该概念原本是指一组数据记录中的特征数据组合能够唯一描述个人,并且不会在人群中偶然出现,那么就会出现“筛选”现象。因此,这些学者主张“数据处理者应当采取适当的保护性披露限制技术预防筛选风险”,其中的“保护性披露限制技术”是指,在法律层面应当至少满足“特征组合在匿名数据集合中出现频率最小化”和“匿名化处理不会显著增加原始数据与其他数据之间的区分可能性”中的一个条件;而在技术层面,“频率最小化”和“显著增加区分度”显然不可能以一个具体数值

〔12〕 Elizabeth A. Brasher, “Addressing the Failure of Anonymization: Guidance from the European Union's General Data Protection Regulation,” *Columbia Business Law Review*, Vol. 2018, No. 1, 2018, p. 228.

〔13〕 Ira S. Rubinstein and Woodrow Hartzog, “Anonymization and Risk,” *Washington Law Review*, Vol. 91, No. 2, 2016, p. 703.

〔14〕 Sophie Stalla-Bourdillon and Alison Knight, “Anonymous Data v. Personal Data—A False Debate: An EU Perspective on Anonymization, Pseudonymization and Personal Data,” *Wisconsin International Law Journal*, Vol. 34, No. 2, 2017, pp. 308-310.

予以规范，而是需要结合具体的数据效用分别设置“0.1%—1.0%”的最小组合识别概率阈值和“5%—20%”的最大组合识别概率阈值，以便实现纳入立法中的“筛选”规范具有技术层面的可操作性。〔15〕当然，也有学者则认为法律层面的匿名化与技术层面的匿名化并不具有可协同性，因为很难将法律植入计算机代码之中，法律意义上的匿名化更像是一种风险缓解策略，降低隐私泄露的担忧，更重要的是，匿名化法定标准越高，意味着立法者希望隐藏更多的个人数据，而数据集合的效用会显著降低。〔16〕

（三）匿名化制度理论分歧的根源：为何匿名化？

从各国匿名化立法差异性来看，各国对于匿名化、去识别化乃至假名化的概念界定并不完全相同。中国《个人信息保护法》将匿名化信息的特征限定为“无法识别”和“不能复原”。欧盟GDPR第4(5)条强调了匿名数据在没有借助额外信息的情况下无法识别数据主体的技术效果，亦即匿名数据本身的无法识别。〔17〕我国与欧盟均选择将匿名化数据排除在个人信息范畴之外，但由于我国强调的是匿名数据的“不能复原”，即匿名数据如若能和其他数据结合“复原”则仍属于个人信息范畴，导致我国的匿名化制度陷入适用困境。而美国在其立法中更常采用去标识化概念，并且2020年的《加州隐私权法案》（The California Privacy Rights Act，CPRA）还规定了个人信息处理者负有禁止再识别的义务，承认了去标识化的数据存在再识别的安全风险。CPRA并没有选择“全有或全无”的方式将去标识化技术作为免于承担数据安全责任的“安全港”，故而也就不存在我国《个人信息保护法》中匿名化技术的法定概念与实际技术效果脱节的问题。由此可见，我国匿名化制度困局的根源之一在于匿名化技术的法定概念与“全有或全无”模式背后的义务内容存在冲突，解决路径也应当是对《个人信息保护法》的匿名化概念作出补充性解释。

从国内外匿名化制度的理论分歧来看，主要的争议焦点可以总结为“匿名化技术是否仍然具有保护个人信息的实践意义”“匿名化规范如何兼顾法律与技术”以及“匿名数据对应着何种义务内容”三个问题。这些问题彼此之间看似毫无关联，但是其内核均是以再识别风险为论证导向。其一，鉴于无法实现完全的匿名化技术效果，并且商业实践也并没有反映出匿名化能够在保障数据效用的前提下实现数据安全，尤其是出现了诸如“美国在线公司（AOL）公布的匿名化搜索记录被《纽约时报》反向识别特定个人”“美国医疗机构公布的医疗数据集合能够与其他机构公布的选民登记表进行关联并识别个人”等匿名化失败案例，部分学者将匿名化的失败

〔15〕 Micah Altman, Aloni Cohen, Kobbi Nissim and Alexandra Wood, “What a Hybrid Legal-Technical Analysis Teaches Us about Privacy Regulation: The Case of Singling out,” *Journal of Science and Technology Law*, Vol. 27, No. 1, 2021, pp. 55-57.

〔16〕 Liane Colonna, “Privacy, Risk, Anonymization and Data Sharing in the Internet of Health Things,” *Pittsburgh Journal of Technology Law and Policy*, Vol. 20, No. 1, 2020, p. 148.

〔17〕 “匿名化”指的是在采取某种方式对个人数据进行处理后，如果没有额外的信息就不能识别数据主体的处理方式。此类额外信息应当单独保存，并且已有技术与组织方式能够确保个人数据不能关联到某个已识别或可识别的自然人。

归因于不可控、不可量化的再识别风险，〔18〕这也就导致了第一个争议焦点。

其二，由于匿名化处理方案与数据结构、数据环境等外部因素相关，加之数据的外部共享、内部使用等差异化的数据处理目的，意图利用抽象凝练的法律规范对复杂多变的技术方案进行完整的描述存在一定难度，这也就导致了第二个争议焦点。并且，这种复杂性更使得匿名化信息的判断标准具有动态性、场景化特征，如在 Breyer v. Bundesrepublik Deutschland〔19〕案中，欧盟法院考量了动态 IP 地址的技术原理、数据采集方合理识别特定自然人可能性以及《德国国家安全法》规定的平台配合执法机构确认用户身份等要素，最终将涉案的动态 IP 地址认定为个人信息而非匿名化信息，这也恰恰反映了法定的匿名化效果同样需要结合个案判断。

其三，主流观点大多赞同个人信息处理者无需对匿名数据承担原有的个人信息保护义务，前提正是匿名数据已经不具有立法者所担忧的侵害个人信息权益之风险，但是客观存在的再识别风险却使得“匿名数据不属于个人信息”这个结论受到质疑，进而导致了第三个争议焦点。可以说，再识别风险问题的解决已经成为匿名化理论研究不可回避的关键问题，并且对于该问题的关注点也从最初的风险是否可量化评估转变成了风险成因及其预防可能性的探讨。

无论是制度差异，还是理论分歧，匿名化制度的相关研究大抵依循“是什么——为什么——怎么样”的论证逻辑，并根据预设的匿名化概念提出不同的匿名化制度规范。但是，在实际的研究过程中，学者们总是下意识地忽略或淡化“为什么要匿名化”这个关键问题，仅仅是以协调个人信息保护与数据效用作为理由，进而转向具体的制度方案探讨。即便有学者将匿名化处理的理论基础归纳为“必要性原则”“目的限制原则”和“区分对待原则”，〔20〕依然未能触及匿名化制度建构的理论根据。这种研究重心的遗漏本质上属于忽视了隐私时代和个人信息保护时代的匿名处理目的的差异。个人隐私和个人信息在特定情形下可能交叉重叠，但是从权益的角度来看，两者意味着不同的风险形式和权利保护方式。国内学者在援引国外观点学说时，未曾注意到美欧学者更侧重从隐私保护的角度探讨匿名化目标的实现，进而也未能区分个人隐私保护与个人信息保护语境下的匿名化差异。以个人隐私保护为导向时，由于个人隐私的保护逻辑是避免隐私信息被擅自公开，所以匿名化处理的目的是保障涉及隐私内容的信息保持私密性。而以个人信息保护为导向时，由于完全隐匿自己在网络上的数据活动根本不可能，并且事实层面的个人完全控制个人信息流向也近乎“乌托邦式幻想”，所以匿名化处理目的更侧重特定范围和条件下的原始信息不可见，亦即“相对匿名化”。并且，匿名化机制的功

〔18〕 Paul Ohm, “Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,” *UCLA Law Review*, Vol. 57, No. 6, 2010, pp. 1717-1725.

〔19〕 Case C-582/14, Patrick Breyer v. Bundesrepublik Deutschland, Judgment of the Court (Second Chamber), 19 October 2016, paras. 31-49.

〔20〕 参见郑曦：“匿名化处理：刑事诉讼被遗忘权实现的另一种途径”，《法治研究》2021 年第 5 期，第 48—50 页。

能是复合性的，在保障个人信息安全的同时，更要促进数据的高效利用，而不是限制数据处理活动。

三、匿名化制度的理论基础再思考

（一）匿名化处理的技术特征与法律性质

匿名化处理的技术原理实际上是将一组数据或数据集合中能够关联到特定自然人的识别要素予以筛除或隐藏，可以说是一种“隐藏个人身份”和“避免从数据集合中发现特定个人”的信息技术。在技术层面，匿名化处理的关键环节便是如何控制和阻断匿名化后的信息与其他数据之间的关联性，既包括基于家庭、工作等社会活动所形成的社会关联性，也包括基于共同偏好等形成的群体关联性。欧盟《关于匿名化技术的意见》总结了匿名化处理的四个基本特征：一是匿名化处理的目的在于防止识别数据主体的具体身份，且过程不可逆转；二是不同的匿名化技术存在不同的适用场景；三是匿名化处理需要考虑到数据控制者和任何第三方识别数据主体所可能采取的一切合理措施；四是匿名化处理存在固有风险，需要事前进行评估。美国的去标识化技术更侧重对于直接识别符、间接识别符的删除和隐匿，在其2010年发布的《个人信息保护指引》中直接将去标识化界定为“通过移除足够的个人可识别信息以至于剩余的信息不能识别特定个人，并且没有合理理由相信这些信息能被用于识别特定个人”。当然，技术层面的匿名化处理并非仅限于姓名、身份证号、联系电话等识别符的删减，随着隐私保护技术的创新发展，数据聚合、差分隐私、数据噪音添加等技术方案也逐渐成为主流。但无论匿名化技术方案如何设计，其底层逻辑大抵可以分为两类：一是直接修改数据组合与特定个人之间的关联程度，牺牲的数据效用主要以数据真实性为限；二是间接淡化或模糊特定个人对应的数据要素，如将个人的出生日期聚合为特定年份出生的群组，牺牲的数据效用主要以数据准确性为限。

前述技术原理仅仅反映了匿名化处理达成匿名效果的基本路径，但实践中还需要考量具体的匿名化数据类型和匿名化目的。因为结构化数据与非结构化数据对应着不同的技术方案：结构化数据主要包括常见的各类统计表等，因为该类数据能够反映完整全面的个人状况，且无需企业进行数据归集、清洗和挖掘等处理活动，故而匿名化处理的基本逻辑主要是通过删除直接识别符、准识别符、根据场景限定特定数据字段〔21〕等方式消除数据再识别的可能性。

〔21〕 美国2015年发布的《个人信息去标识化》列举了基于场景的去标识化评估——“网飞奖”，即一个人看过的电影数量可以是一种标识符。网飞公司发布了一个包括一些客户看过的电影数据集，并将其列为“网飞奖”竞赛的一部分。虽然该数据集中没有直接标识符，但研究人员发现，观众观看的一组电影（尤其是不太受欢迎的电影，如邪典和外国电影）作为网飞数据集中的用户资料经常可以匹配到互联网电影数据库（IMDB）中的单个用户资料，这些用户资料没有被删除标识，并包括用户名，其中许多是真名。这一威胁的具体情形是，通过在IMDB上对几部电影进行评级，一个人可能会无意中透露他们看过的所有电影，因为他的IMDB个人资料可能与网飞奖的数据相关联。参见美国国家标准与技术研究院（NIST）官网，https://nvlpubs.nist.gov/nistpubs/ir/2015/nist.ir.8053.pdf，最后访问日期：2024年2月24日。

如美国 HIPPA 法案所列举的包括姓名、电话号码、社保号、银行账号等在内的 18 类直接识别符，该法的调整对象只要删除了这些直接识别符就即可被认定为完成了数据的“去标识化”。非结构化数据则是指数据要素之间并不存在直接的逻辑关系和固定结构，其匿名化处理方式则面临着诸多问题。以汽车数据的匿名化为例，一是智能网联汽车生成的包含图片、视频等非结构化数据脱敏难度较高，二是匿名化处理过程中难以把控脱敏的个案标准/程度，因为在脱敏的过程中还需要保留用于安全驾驶的基础信息，例如行人、车辆的信息。中国汽车工业协会发布的《汽车传输视频及图像脱敏技术要求与方法》和国家标准《智能网联汽车 数据通用要求》则对匿名化处理结果提出相应的评估方案，即“敏感区域不可恢复”“多帧无法还原信息”“脱敏区域和实际人脸/车牌区域的交并比应当满足 50%—75%”等。

无论是国内外有关匿名化技术概念的界定差异，还是技术层面匿名化处理模式的场景化特征，其实都说明一个关键事实：法律与技术层面的匿名化概念存在“鸿沟”，而且匿名化处理并不是一个固定的技术方案或者总能保持同等效果的技术措施。探讨匿名化技术原因的目的在于重新理清法律与技术话语体系下匿名化概念的实质区别，学界久争不决的根源恰恰是将法律话语体系下的匿名化效果强行捆绑于技术话语体系下的匿名化技术特征。从前述事实不难发现，法律文本中多习惯以“不可复原”“不可再识别”等修饰词作为匿名化处理的基本要求，殊不知这种修饰词天然地与匿名化处理的个性化特征格格不入，因为以同一个标准去规范不同场景、不同类型的匿名化处理不可避免地存在“削足适履”的问题，而这也是我国《个人信息保护法》中匿名化概念界定的问题之一。进一步而言，“根据法定要求实施满足相应标准的匿名化措施”和“采取具有相同安全效果的匿名化措施”本质上是两个层面的问题，而理论争议中往往存在将前者异化为后者的逻辑误区，匿名化措施所遵循的技术原理是尽可能消除数据之间的关联性，而不是彻底切断数据之间的关联性，这也是缘何需要重新审视匿名化处理义务理论基础的原因之一。

（二）匿名化制度的理论基础证成：数据关系理论

现有的匿名化制度研究争议之一是数据自由使用与个人信息安全利益之间的平衡问题，倘若仅在纯粹的法律价值层面进行讨论，无助于解决匿名化后再识别等实践问题，故而问题的解决思路则回到了如何控制数据结合所产生的不确定风险。在用户画像模式中，用户标签虽然不属于个人信息，但数个用户标签的组合却有可能识别到特定自然人。国外学者为了应对大数据分析技术对个人信息、个人隐私保护模式的影响，在爱德华·布鲁斯汀（Edward Bloustein）提出的“群体隐私”（group privacy）基础上延展论证了具有数字时代特征的群体隐私理论。〔22〕国外学者也发现集体行动对传统个人隐私保护效果的影响，提出“集体隐私”（collective privacy）的概念。因为基于数据分析技术，单一来源的保密信息（如脸书等平台的

〔22〕 Luciano Floridi, “Group Privacy—A Defense and an Interpretation (June 17, 2017),” pp. 19-20, https://papers.ssrn.com/sol3/papers.cfm? abstract_id=3854483, last visited on 24 February 2024.

用户标签)能够解释多个个人隐私信息,进而对这些信息传播的合法性存在诸多争议。〔23〕部分学者将群体隐私划分为两个群体隐私类型,一种是基于特定社会关系(如家庭等)形成的群体,这些群体成员对该类隐私共同享有权利,但个体成员无法直接主张该权利;另一种则是基于算法分组所形成的群体,这些群体共同具有的社会交往活动信息,也被称为"推导类隐私"。〔24〕不过,国外学者认为经由数据分析获取特定群体的社会交往行为规律或者个人行为模式选择属于数字社会必要的技术活动,故而"推导类隐私"并不能像布鲁斯汀所主张的群体隐私权那般作为一项绝对权存在,更适宜作为一种道德权利。进一步而言,推导类隐私在某种程度上与能够反映特定群体在社会交互活动方式的群体特征信息(如用户标签等)具有相似性,两者均是通过数据结合分析的方式还原特定群体或特定个人的社会交往活动。核心区别在于,国外学者的推导类隐私的内容是以隐私信息为主,群体特征信息所结合推导的信息内容是以个人信息为主。不过,推导类隐私作为一项道德权利之论断也能够为匿名化制度提供一种观察视角:匿名化后再识别风险的预防虽然需要重点关注数据与数据之间的结合分析,但是这种结合分析的干涉应当存在相对明确的边界,法律不可能为了控制再识别风险而禁止所有的数据分析行为。

从商业实践的角度来看,所有的数据收集、处理行为都只存在一个经济性目的,即建构自然人与自然人之间的群体性特征关系。具备一定经济价值的数据集合必然包括自然人一定数量的身份属性值,"少于三个属性点的数据集合毫无价值"。〔25〕事实上,国外学者已经注意到个人信息保护领域的赋权弊端,如美国萨洛梅·维尔约恩(Salomé Viljoen)基于"将信息法益弱化为个人立场的权利诉求",提出数据治理的关系理论。其理论价值在于回答两个数据治理领域的核心问题,一是数据之间的关联性在数据治理领域有何特殊意义,二是这种基于关联的数据社会关系包含哪些内容。对于第一个问题,维尔约恩认为个人数据权利保护模式忽视了数字经济中数据收集行为的主要目的,即单一个体的数据经济价值微不足道,企业收集用户个人信息的根本目的是尽可能发现数据的关联性,而这种关联性可以转变为以用户画像、个性化推荐、同类客户挖掘等常见的盈利业务模式。对于第二个问题,萨洛梅·维尔约恩将数据关联划分为纵向与横向的数据关系,纵向的数据关系是指个体层面的数据主体与数据处理者之间的数据关系,用以解释为何需要设置个人数据权利以及特殊的数据处理义务,横向的数据关系则是指"不同数据主体之间的关联性以及与其具有相同群体特征的其他主体"之间的数据关

〔23〕 Lior Jacob Strahilevitz, "Collective Privacy," in Saul Levmore and Martha C. Nussbaum (eds.), *US: The Offensive Internet: Speech, Privacy, and Reputation*, Massachusetts: Harvard University Press, 2010.

〔24〕 Michele Loi, "Markus Christen, Two Concepts of Group Privacy," *Philosophy and Technology*, Vol. 33, No. 2, 2020, p. 207.

〔25〕 参见蒋洁、兰舟、祁怡然:"个人信息去标识化的类型结构与治理方案",《图书与情报》2021年第3期,第81页。

系。〔26〕举例而言,IP 地址信息具有唯一标识性,能够与更少的信息要素组合识别到特定自然人,如“IP 地址信息+姓名”“IP 地址信息+某天上网时间”等;但是,IP 属地信息在横向数据关系层面需要更多的信息要素组合才能识别到特定自然人,如“IP 属地信息+姓名+联系电话”才能准确识别到特定自然人。

与国内个人信息保护的主流观点相比,数据关系理论将横向的数据关系缺位作为个人信息安全与使用法益失衡的根本原因。并且,国内少数学者已经注意到个人信息识别背后的数据关系问题,进而将个人信息保护制度解释为“由私主体及公私主体间霍菲尔德法律关系构成的法律关系网络”。〔27〕因为个人信息保护对应的法益并不是财产性利益,而个人信息之间的关联性才是商业活动所需要的财产性利益,所以导致在论及法益平衡时,纵向层面的个人信息权利与横向层面的商业利益被强制置于同一维度内予以权衡比较,其结果也必然是笼统层面的价值比较。进一步而言,相对匿名化的判断标准问题实际上也是基于数据关联性进行数据分析活动的法律干涉问题。例如,在交通管控领域,分析特定路段行车路线、行车习惯、起点与终点等数据进而针对性地进行道路疏导,这类数据分析行为虽具有识别个人信息的可能性,但从降低交通事故、解决道路拥堵等社会公共利益层面来看,又是合理且必要的。之所以将匿名化制度的理论基础解释为数据关系理论,原因有三:第一,匿名化制度的技术原理即尽可能消除具有关联第三方的群体性特征,淡化借由社会关系再识别自然人的可能性,而数据关系理论也恰恰是将数据主体关联性所对应的数据关系作为数据治理的核心内容。第二,数据关系理论为安全与利用的法益平衡提供了相对具体的解释标准。倘若将个人信息自决权保护作为匿名化制度的理论基础,其结论必然是“消除所有再识别可能的匿名化效果符合法律规定”,这也意味着匿名化信息丧失了企业所需的经济价值,因为达到这一匿名化效果必然需要清除所有可能关联到特定自然人的识别符。然而,在数据关系理论中,匿名化制度满足安全与利用双重法益的原因被解释为匿名化处理能够保留企业所需的基本经济价值,而不是清除所有具有社会关联意义的识别符。进一步而言,匿名化义务的履行同样可以将“横向数据关系的显著程度”作为认定标准。只有满足“基于显著的数据关系能够再次识别特定自然人”以及“识别难度和识别成本属于合理预见范围”这两个条件,才可以认定匿名化处理符合《个人信息保护法》的基本要求。第三,数据关系理论能够解释在特殊场景中“匿名化信息为何不是个人信息”。横向数据关系的存在使得“个人信息”概念在商业实践中“举步维艰”,由于数据分析技术的迭代优化,使得任何存在关联的不同数据在理论层面具有识别特定自然人的可能,所谓的“不属于个人信息的匿名化信息”也被视为技术空想。但在数据关系理论下,以“识别”为关键要素的个人信息认定模式则转变为以“识别可能性”与“显著关联程度”的综合认定模式。换言之,匿名化信息之所以不属于个人信息,不是因为“不具有识别性”,而是因为“不具有识别特定社会关联的能力”。

〔26〕 Salomé Viljoen, “A Relational Theory of Data Governance,” *The Yale Law Journal*, Vol. 131, No. 2, 2021, pp. 573-613.

〔27〕 参见戴昕:“数据界权的关系进路”,《中外法学》2021 年第 6 期,第 1571—1572 页。

(三) 数据关系理论在匿名化制度的适用

数据关系理论在解释匿名化制度目标时,其核心作用在于能够提供相对明确的判断标准。既然再识别风险不可能彻底消除已成共识,那么匿名化制度的建构重心便成为"法律认可的匿名化处理效果如何解释"或"如何将再识别风险控制在可接受范围内"。基于隐私权保护的匿名化处理与基于个人信息保护的匿名化处理实属两种治理路径:对于前者,因为隐私信息的特性在于高度的私密性,故而匿名化处理应当满足"完全不可复原"之效果。对于后者,因为在万物互联的网络空间中,数据一旦集聚,就有可能产生重新识别的可能性,故而匿名化处理应当满足"不存在能够识别到特定自然人的显著关联信息"。结合数据关系理论来看,横向数据关系决定了匿名化处理方案具有显著的"个性化特征",即在技术层面需要考虑所处理个人信息的信息系统环境,阻断能够轻易发现群体性特征的横向关联数据,实现"最适合"而非"最先进"的匿名化处理效果。那么。在后者的治理路径中,"不具有识别性"这一判断标准难以为商业实践提供明确清晰的指引,而"不具有识别特定社会关联的能力"则能够在技术方案层面解决这一问题:其一,根据数据关系理论,个人信息处理者在选择匿名化技术方案之前,需要明确待处理个人信息所包含的群体性特征,并且将群体性特征所反映的"共同行为特征"或"社交关联属性"纳入再识别风险影响因子。其二,明确信息系统环境安全性、其他数据处理者的数据安全保障能力等场景化要件,进而分析上一步的风险影响因子是否会使风险作用力显著提升。例如,待处理的个人信息为"用户 ID 号码+近期消费购物类型+购物时间段",那么,"购物类型"和"购物时间段"则属于群体性特征。倘若电商平台打算以此分析近 5 年的购物偏好,那么"购物类型"和"购物时间段"就会因为 5 年内的用户消费数据聚集而导致再识别风险显著增加。但是,倘若电商平台仅仅打算以此分析用户一年内的消费频率,那么"购物时间段"而非"购物类型"才会增加再识别风险。在明确风险增加的来源之后,则需要阻断或降低横向数据关系的显著性。继续以前述例子为基础,在分析近 5 年购物偏好的目标下,匿名化处理的效果应当表现为阻断和降低"购物类型"和"购物时间段"两个横向数据的关联性:如存在"购物类型"信息要素时,则需要选择消除、泛化"购物时间段",将原来的"晚上 10 点购物"这类信息内容处理为"晚上购物"等。

进一步而言,从阻断和降低横向数据关联性的目标来看,匿名化处理模式至少应当满足隐藏数据属性、分离数据关系、扰乱数据排序三个效果,即实现遮蔽、切断和打乱横向数据关系的效果。隐藏数据属性主要表现为泛化和抑制特定数据属性,泛化是指将原有的数据属性值用父值替代,如个人的居住地在海淀区清河街道,采用数据泛化的结果是用北京市替代海淀区清河街道;抑制是指对具有唯一标识性的数据属性值直接予以删除,例如在教育部门公布录取名单时,不公布身份证号码,只公布准考证号末尾数。分离数据关系主要表现为分离其他数据属性值与敏感属性值之间的关联程度,既包括无法与匿名化信息中的其他属性值相互结合再识别,也包括无法将匿名化信息的部分属性值与其他数据集合中的敏感属性值予以直接关联。扰乱数据排序则是指采用增加数据噪声、平均值或者合成数据等技术措施导致第三方无法判断原始数据的属性与数值的对应关系。当然,这些技术效果主要还是面向匿名化处理方案的选择,此外,匿名化处理过程中还需要在处理环境、处理溯源以及方案比较层面进行明确相关

的技术标准的明确，即在安全可控的数据处理环境下进行匿名化处理，并且匿名化处理所涉及的原始数据、匿名数据流动应当能够溯源和记录，在多个匿名化处理方案中比较并选择最优化方案。

需要说明的是，基于数据关系理论所确定的匿名化判断标准与传统的匿名化、去标识化相比，优势在于能够结合再识别风险成因和作用机制形成较为明确的业务合规模式。一方面，相较于传统的匿名化，《个人信息保护法》所规定的“无法被识别或关联”“不能被复原”之标准与实践中的技术效果并不完全贴合，在法律适用过程中仍然需要进一步解释。另一方面，《个人信息安全规范》将“去标识化”界定为“使处理后的个人信息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”，其内在逻辑是降低信息区分度，而这些信息一旦与其他信息一并处理，并不排除重新识别的可能性。此外，还需要予以澄清的是，匿名化技术是一个宽泛的概念，并不存在唯一的技术标准。因此，在法律制度层面探讨匿名化技术更多的是探讨如何选择更为恰当的匿名化技术方案，而不是对现有技术工具提出一个完全无法实现的技术目标。换言之，在数据关系理论框架下，匿名化处理所采用的具体技术仍然以泛化技术、随机化技术、K-匿名模型等为主，唯一不同的是，相应的匿名化制度内容则表现为一套匿名化处理流程行为规范。例如，在团体标准《互联网广告匿名化实施指南》中，匿名化处理则表现为包括“环境维护、确定目标、技术处理、效果评估、行为控制和过程监督”等在内的综合性匿名化业务流程。

（四）匿名化再识别风险问题的解释路径

匿名化处理的再识别风险判断问题实质上并非一个纯粹概率量化问题，讨论风险概率的多少、风险可接受程度以及是否能够做到完全预防再识别风险无助于解决现实问题，匿名化处理的法定标准应当从完全风险控制转向至相对风险预防。匿名化处理针对的是横向数据关系，在消除这类数据关系显著识别性的同时，确保横向数据关系的经济价值，故而匿名化处理是否符合法定要求的判断依据应当兼顾处理过程与处理结果。匿名化处理过程应当采取尽可能安全可靠的技术方案消除具有显著识别功能的标识符，匿名化处理结果应当满足数据处理者无法通过可以预见的技术措施或其他辅助数据集合重新确定匿名化信息存在的横向数据关系。事实上，国内外也有不少学者主张匿名化信息再识别风险应当限定为相对风险的控制，但大多因为无法解释所谓的“相对”程度而难以自圆其说。即便以“采取所有合理、可能的手段仍无法识别”〔28〕“再识别成本难度较大”等细化标准予以补充说明并限定“相对风险”区间，在适用过程中仍难以同时满足特定场景下特定主体匿名化处理的特殊需求。此外，也有学者以欧美等模式为范本，提出包含“可以防止数据集中再次识别”“可以防止通过链接同一自然人的数据属性再次识别”和“可以防止从数据集合中单独推断识别”三项有效匿名化标准，〔29〕但是这仍属于最基本的判断标准。

这其中部分原因在于学者们仅仅作出了相对风险的学理性解释，忽视了匿名化制度本身

〔28〕 参见李润生：“论个人医疗信息的匿名化处理制度”，《交大法学》2022 年第 4 期，第 133 页。

〔29〕 参见张涛：“欧盟个人数据匿名化的立法经验与启示”，《图书馆建设》2019 年第 3 期，第 63 页。

兼具技术属性与法律属性，未能同时作出与学理性解释配套的一般性技术标准，最终导致匿名化制度的再识别风险迟迟未能得到真正解决。在数据关系理论框架下，匿名化信息相对再识别风险的预防逻辑表现为尽可能使得第三方无法或难以明确具体的横向数据关系，包括数据集合所反映的社交关系、群体性特征等关系类型。

在法律话语体系下，匿名化信息识别风险的预防效果应当分别从识别难度、识别成本以及识别来源三个角度综合考量，部分国外学者也提出了身份识别的货币成本、所需时间、可用技术以及技术发展等类似的判断标准，〔30〕其目的也是为了协调法律与技术层面有关再识别风险的不同认知差异。首先，识别难度是指第三方再识别出特定自然人需要采取较高标准的技术手段和辅助数据集合。因为法律层面的识别不仅仅包括人为地主观推断匿名化信息对应的个人信息，还包括机器结合唯一识别符与其他标识符形成的特定数据关系，从数据集合中区分特定个体，所以识别难度主要体现在“利用常见的数据分析技术难以识别”。其次，识别成本则是指再识别所需要的成本远高于再识别所能获得的经济利益。通常而言，第三方之所以恶意再识别匿名化信息，其根本目的是通过将再识别的信息进行售卖等活动获取相应的经济利益，但是如果识别成本过高，例如需要花费一定的技术成本破解加密方案、用于辅助间接识别的其他数据获取难度大、可再识别的数据规模较小等，那么行为人就会缺乏主动进行数据分析和重新识别特定个人信息的经济动机。相对应的，大范围、规模化的匿名信息再识别安全事件发生概率也会随之降低。最后，识别来源主要针对的是意图再识别的行为主体类型。在既有的学说假设中，常存在重视再识别可能性而忽视再识别主体的现象，这会导致泛化可能发生的再识别风险。因此，匿名化处理模式的选择以潜在的再识别行为主体为基础，充分评估可能存在的再识别风险类型，这也是为何需要引入数据关系理论作为匿名化制度的理论基础。

四、基于数据关系理论的匿名化制度建构路径

（一）《个人信息保护法》中匿名化制度的体系定位

匿名化处理无法彻底预防再识别风险并不意味着“匿名化已死”，即便是隐私计算等新兴技术也无法满足理想状态下的完全匿名效果。匿名化技术仅能实现相对范围的再识别风险预防，基于数据关系理论的匿名化制度则是以促进数据商业化利用和减少对个人信息权益的损害为基本目的，其功能定位并非传统的意义上的个人信息保护制度，而是要求数据处理者以最大化、最优化的方式履行个人信息保护义务，借由个人信息的安全保障实现无障碍的数据流动和交易。恰如前文所提及的，对于监管机构而言，最大的担忧莫过于匿名化制度成为数据处理者不履行个人信息保护义务的“避风港”。但是，既然明确了匿名化技术预防再识别风险仅以

〔30〕 Michael Kolain, Christian Grafenauer and Martin Ebers, “Anonymity Assessment—A Universal Tool for Measuring Anonymity of Data Sets under the GDPR with a Special Focus on Smart Robotics,” *Rutgers Computer and Technology Law Journal*, Vol. 48, No. 2, 2022, p. 222.

相对风险为限，那么这种担忧也将得到充分解决：一方面，匿名化处理效果未能满足法定义务标准和技术安全标准，即便信息经过匿名化处理，也不能当然认定该信息属于匿名信息，而非个人信息；另一方面，匿名化处理所能够预防的再识别风险主要是以精准识别横向数据关系为主，在绝大多数情况下，匿名化信息能够满足此种技术效果，再识别的难度、成本无疑会远超能够获得的实际成本，能够有效降低第三方再识别的经济动机，故而也能在一定程度上解决匿名化制度成为过度处理个人信息的“免责事由”。

《个人信息保护法》将匿名化信息排除在个人信息范畴之外，虽然意味着匿名化信息的处理活动不必遵守《个人信息保护法》的强制性规定，但是这种制度效果不应当简单解释为“全有或全无”的规定模式。因为匿名化制度的制度目的并不是提供数据商业化利用的免责事由，而是提供一种兼顾安全和利用的技术方案，以制度的形式要求个人信息处理者按照最优化方式充分履行匿名化制度所要求的个人信息安全保护义务，也有学者将匿名化的功能与价值归结为“排除个人信息以发挥信息效用”和“控制信息风险以履行法律义务”。〔31〕 进一步而言，匿名化处理并没有免除个人信息处理者的法定义务，而是将这种法定义务嵌入匿名化处理过程中。前文提及的匿名化处理标准就是以阻断横向数据关系的识别为核心内容，在实现法律与技术双重标准的匿名化处理活动后，个人信息处理者的法定义务也得到充分履行，故而《个人信息保护法》也无必要重复要求履行相同目的的个人信息保护义务。并且，倘若个人信息处理者未能充分履行匿名化制度的基本要求，那么即便处理后的信息在事实层面具有匿名效果，且暂时未被第三方再次识别，也不等同于个人信息保护义务的履行完毕。因为不符合法定要求和技术标准的匿名化信息仍然存在相当程度的再识别风险，风险尚未发生并不意味着风险不会发生，故而《个人信息保护法》的相关内容仍得以适用。

因此，匿名化制度的功能定位在于最大程度保障匿名化技术的可靠性与安全性，并且，该项制度以“重新识别横向数据关系”为内容，恰好能够与其他个人信息保护制度予以衔接。其一，匿名化制度与个人信息保护影响评估、数据安全风险评估机制衔接。尽管匿名化信息不属于个人信息，但匿名化处理过程作为能够对“个人权益有重大影响的个人信息处理活动”，属于《个人信息保护法》第 55 条第 5 项情形，故而在匿名化处理之前应当进行《个人信息保护法》第 56 条规定的个人信息保护影响评估。并且，对于规模化的匿名化处理活动，因为待处理的海量个人信息可能构成“重要数据”，符合《数据安全法》第 30 条提及的“对数据处理活动定期开展风险评估”的情形，故而需要进行数据安全风险评估。其二，因为匿名化处理的技术原理是淡化或消除显著的横向数据关系，故而需要考量公开可获取的辅助数据集合。而在实践中，最能够反映横向数据关系的数据集合属开放的公共数据，囊括了各类社会活动所形成的社会关系。因此，在进行匿名化处理活动中，公共数据的脱敏处理方案以及公开范围和方式均属于匿名化处理需要考量的重点事项。其三，匿名化制度也与数据泄露通知义务相关，因为数据泄露事件的发生往往意味着具有关联性的原始数据可获取，第

〔31〕 参见韩旭至：“大数据时代下匿名信息的法律规制”，《大连理工大学学报（社会科学版）》2018 年第 4 期，第 72—73 页。

三方再识别匿名化信息的风险显著增加，故而需要将这类风险可能导致的损害作为泄露通知的重要事项。

（二）匿名化制度的建构模式选择：单行立法抑或补充解释

在匿名化制度实施层面，存在单行立法和补充解释两种路径用以解决现阶段《个人信息保护法》中匿名化制度“存而不用”的尴尬境地。单行立法路径的优势在于，通过细化匿名化处理的具体规则，将《个人信息保护法》规定的法定义务以预防部分再识别风险的形式转化为同等效果的匿名化处理义务，通过体系化的匿名化规则满足个人信息安全监管的要求。补充解释的优势则在于，通过对《个人信息保护法》第 73 条第 4 项“匿名化”概念进行补充解释，填补匿名化技术的实际功能与法定要求中“无法识别”和“不能复原”之间的不适配。不过，从目前立法现状而言，有关匿名化制度的具体内容少之又少，似乎单行立法这一路径或将成为最佳选择。然而，这一路径也面临着诸多难题：一是现有的匿名化规则体系并不足以支撑一部单行立法的法条容量，并且相关条款内容会与《个人信息保护法》相互重叠。二是我国目前已经制定诸如《互联网广告 匿名化实施指南》等匿名化相关的技术标准，面对更为复杂的数据安全风险，技术标准的灵活性与专业性可能更符合当下匿名化制度适用的实践需求。三是匿名化制度具有较强的技术属性，该项技术的更新迭代可能无法从内部稳定的法律制度得到体现。综合来看，现阶段以“补充解释＋技术标准＋实施指南”的方式更能满足灵活性与实践性的制度需求。

此外，在少之又少的匿名化信息相关案例中，法院认定匿名化信息的依据还是以“是否可识别”为核心，更确切地说，法院实际上是将“个人信息”与“匿名化信息”作为一对非此即彼的二分化概念关系。例如，在“美景信息科技有限公司、淘宝（中国）软件有限公司商业贿赂不正当竞争纠纷案”〔32〕中，法院认定涉案“生意参谋”数据产品所使用的用户信息经过匿名化脱敏处理后已无法识别特定个人且不能复原；在“上诉人北京百度网讯科技有限公司与被上诉人朱烨隐私权纠纷一案”〔33〕中，法院认定涉案数据信息的匿名化特征不符合“个人信息”的可识别要求，其认定的逻辑可以总结为，按照法定的个人信息概念界定，涉案的 cookie 信息等并没有与用户身份予以绑定，无法确定具体的信息规划主体，故而可以认定涉案数据信息所实现的匿名化效果不满足“个人信息”的概念界定。这种二分法的解释方式除了存在“用静态思路假设数据匿名化结果”〔34〕的弊端之外，还在混淆了“个人信息”和“匿名化信息”之间的差异性之问题，其潜在的逻辑是，只要不符合个人信息概念所要求的“识别性”就属于匿名化信息，而这一逻辑的必然推论是，因为匿名化信息不满足个人信息的概念要求，那么就无需遵守《个人信息保护法》的所有内容，进而再度陷入“全有或全无”的制度困境之中。

因此，匿名化制度的实施路径更应当对《个人信息保护法》第 73 条第 4 项规定的“匿名化

〔32〕 浙江省杭州市中级人民法院民事判决书，(2018)浙 01 民终 7312 号。

〔33〕 江苏省南京市鼓楼区人民法院民事判决书，(2013)鼓民初字第 3031 号。

〔34〕 参见衲钦、张慧春：“数智环境下匿名数据治理创新对策研究”，《科学管理研究》2022 年第 2 期，第 128 页。

信息”概念进行法教义学解释。第一,“经过处理”泛指各类匿名化处理技术,而不单一限定某一类特定技术方案。这里的“处理”与个人信息“处理”所包含的加工、传输、共享等常见环节并不相同,更侧重体现匿名化处理的技术特征,即应当以横向数据关系为对象的处理活动。有意思的是,在国外司法实践中,法院对于匿名化处理的认定是排除一般的数据处理行为,如英国法院结合彼时 1998 年的《个人数据保护法案》中宽泛的“个人数据处理行为”定义,以匿名化处理导致个人信息丧失与特定个人之间的关联而不受立法调整为由,将该处理活动排除在法定的“处理行为”之外。〔35〕 第二,“无法识别”与“不可能识别”“完全不能识别”并不相同,实际强调的是一种“个人信息经过匿名化处理后难以再次被识别”的技术状态。需要注意的是,“无法识别”的内涵除了再识别的可能性属于可接受范围之外,还包括采用了所有合理可行的匿名化技术用以保障匿名化信息无法被再次识别。此外,“无法识别特定自然人”这里应当作扩大解释,如果以《个人信息保护法》第 4 条有关“个人信息”概念界定作为解释标准,则会使得匿名化信息与个人信息在概念界定层面具有二分化的对应关系。因此,“无法识别特定自然人”既包括无法识别特定个体,也包括无法识别与特定个体显著相关的数据关系。第三,“不能复原”是指匿名化信息无法恢复到原始状态,不包括部分信息要素的复原。从立法目的来看,倘若匿名化信息能够复原至原始状态,意味着匿名化处理并不符合数据安全监管要求,且更能够轻易再次识别到特定自然人。相对地,如果将“不能复原”解释为“任何形式的部分复原或全部复原”,则显然超过了匿名化处理客观能够实现的技术效果。

(三) 匿名化制度的治理功能衔接:法律与技术

与个人信息保护影响评估、知情同意等个人信息保护规则相比,匿名化制度的特殊性表现为法律话语体系与技术话语体系的融合,单纯以笼统抽象的价值评断只能解决“匿名化义务是否履行”的判断问题,而无法解决“匿名化义务是否充分履行”的判断问题。因为匿名化制度的再识别风险预防逻辑是一种相对风险预防,为了尽可能降低再识别风险可能导致的安全事件,个人信息处理者的义务履行方式需要以“充分履行”为标准。进一步而言,匿名化制度需要兼顾法律的规范性作用和技术的实用性功能,这也是将技术标准、实施指南等作为匿名化制度实施路径的原因之一。由于法律话语体系的简洁性、抽象性与技术话语体系的具体性、明确性,简单地在规范性文件中事无巨细地设置具体的匿名化操作流程和技术方案并不能实现法律与技术兼顾的目标。更确切地说,法律与技术的话语体系融合应当表现为“以立法目标作为技术措施是否充分的判断依据,以技术原理作为条款内容的创设依据”,也就是说,匿名化制度的内容建构并不需要以具体的技术细节作为核心内容,而是提供一种技术方案选择和实施的规范化流程,至于具体的技术细节则由义务履行者根据自身的技术能力、经营成本等因素进行选择。

具体而言,在实施指南层面,匿名化制度的体系内容主要以淡化和消除横向数据关系为目标,同时结合相对再识别风险的预防逻辑,可以将匿名化处理流程划分为 5 个主要业务流程,

〔35〕 John Devereux, “R v Department of Health: ex parte Source Informatics Ltd,” *Journal of Law and Medicine*, Vol. 8, No. 1, 2000, pp. 27-28.

以此满足我国个人信息保护以及数据安全领域“全生命周期安全”的治理理念。

第一步，进行数据安全评估，其目的是充分预见和评估可能的再识别风险来源以及再识别所采取的技术方案，并以此作为是否充分预防再识别风险的判断依据。具体的评估事项包括已经公开且与匿名化信息显著相关的辅助数据集合、匿名化处理的个人信息是否属于重要数据、采取特定匿名化处理方案可以被哪些技术措施进行反向识别、是否存在与匿名化信息相关的数据泄露事件等。

第二步，预先确定数个匿名化处理技术方案，并评估和比较各个处理方案的优劣势。该环节看似增加了企业匿名化处理的业务成本，但匿名化处理本身具有针对性的技术特征，并不是采用同一种匿名化方案即可满足所有个人信息保护需求。因此，要求个人信息处理者确定数个匿名化处理技术方案本身即“充分履行义务”的一种直观体现，并且也满足了匿名化制度所要求的“选择最优化方案”。

第三步，匿名化处理内部管理制度实施情况评估，其目的是预防因内部员工操作不当、信息系统安全漏洞等内部原因导致的数据安全事件，以封闭式的处理环境保障匿名化信息的不可复原。客观而言，匿名化处理的安全性除了技术本身的安全可靠之外，还表现为处理环境的安全可信，如信息系统设置访问权限、为匿名化处理设置内部安全合规审计、匿名化处理的原始数据与外部网络隔离、匿名化处理操作流程及其业务人员身份记录与留档等。简言之，主要包括处理方式、操作人员、内部泄露、访问设置以及结果公开五个事项的安全。

第四步，模拟再识别事件发生可能产生的负面影响以及应对措施，其目的是进一步判断匿名化处理所能预防的再识别风险是否属于社会可接受范围。事实上，英国 2012 年发布的《匿名化：数据保护风险管理实践准则》所采用的“蓄意侵入者检验”标准〔36〕就是一种假设模式，即假设侵入者的识别动机、识别能力，并将具体的识别标准设定在介于“普通公众”和“具有一定专业知识和技能的专业人士”之间。〔37〕前述步骤本质上还是在事前阶段采取各种管理措施或技术措施消解潜在的安全风险，模拟再识别安全事件导致的负面影响则是验证采用的匿名化处理技术是否能够最大限度减少规模化的数据安全事件，并且提前确定应对措施也能够尽可能减少再识别可能导致的二次损害。

第五步，定期审查匿名化处理技术方案的安全可靠性，其目的是应对经优化迭代的再识别技术致使原有的匿名化技术方案不再安全的问题，同时也是为了应对其他数据源开放数据可能导致的横向数据关系再识别问题，亦即“处理完成后的再识别风险”。〔38〕匿名化制度所确立的匿名化效果应当是一种动态匿名效果，〔39〕即匿名化的判断标准并不以匿名化处理完毕

〔36〕 程海玲：“个人信息匿名化处理法律标准探究”，《法律与科学》2021 年第 3 期，第 30 页。

〔37〕 Information Commissioner's Office, “Anonymisation: Managing Data Protection Risk Code of Practice,” https://ico.org.uk/media/1061/anonymisation-code.pdf, last visited on 20 February 2024.

〔38〕 参见王立梅：“大数据视角下的个人信息匿名化规则构建”，《云南大学学报（哲学社会科学版）》2021 年第 5 期，第 146 页。

〔39〕 参见夏庆锋：“网络空间个人信息保护的通知义务完善与动态匿名化”，《江汉论坛》2022 年第 3 期，第 102 页。

时的实际效果为限，同时也包括匿名化处理是否能够应对未来潜在的再识别风险。

表 1

匿名化制度主要内容	法律目标	技术目标
1. 数据安全风险评估	安全风险预防	明确风险类型，确定匿名化方案
2. 选择最优匿名化方案	充分履行义务	确保匿名化处理方案的针对性
3. 内部管理制度实施评估	全生命周期安全	预防和解决内部安全风险
4. 再识别风险模拟	判断风险是否可接受	验证处理效果，设置应急预案
5. 定期审查匿名化效果	预防未来再识别风险	更新和优化匿名化技术方案

五、结 语

既然绝对匿名化根本无法实现，那么为了解决数据使用与数据安全之间固有的利益冲突，则需要以匿名化的最核心风险为基础，阻断匿名化后信息与其他群体信息之间的社会关联性。在实际的数据交易活动中，匿名化制度能够成为避免企业过度担心数据安全不合规、未履行个人信息保护义务的重要技术处理路径。我国目前的数据治理模式正在从单一强调数据安全转型至以数据安全促进数据流动，故而不能简单地将匿名化制度功能解释为保障个人信息安全，而是应当以匿名化处理所针对的数据关系为起点，将匿名化制度的功能定位为实现更大范围的数据商业化利用。从数据关系理论来看，数据的经济价值来自数据与数据之间所能反映的社会活动关系以及群体特征，并且这些数据关系能够应用于用户画像、个性化推荐、客户挖掘、产品定制化升级等诸多业务活动。匿名化处理的对象正是这些具有经济价值的数据关系，因而匿名化制度才能够在理论和实践层面达到安全和利用的法益平衡。并且，在数据关系理论下，匿名化信息的再识别风险争议同样值得重新审视。既然现在以及未来的任何一种信息技术均无法实现完全意义上的不可再识别，那么继续以再识别风险作为“匿名化制度已死”或“匿名化信息无法实现匿名化”的论证理由显然毫无意义。更重要的是，如果以匿名化信息不适用《个人信息保护法》且该类信息仍然有可能被再识别为由，限制匿名化制度的实施，则会导向更为严重的偏差结论，即任何以匿名、去标识等为目的的信息技术均不能构成个人信息处理行为的免责事由，并且也不能以匿名处理为由交易包含个人信息的匿名数据集合，因为这些信息技术无法保障绝对无法再识别。因此，需要明确的是匿名化处理仅能够预防可以合理预见且具有一定发生可能性的再识别风险，而不是停留于理论假设层面的所有风险类型。在我国匿名化制度的实施过程中，这种相对风险预防的可接受性并不直接表现为在制度或技术标准层面设置可以量化的风险概率区间，而是将匿名化技术效果作为匿名化制度的主要内容，通过将匿名化处理过程规范化和模块化，将再识别风险预防目标拆解成具体的个人信息安全保护目标，

以此达成匿名化制度的立法目标。此外,与再识别风险更为相关的是公共数据开放制度,这些数据集合往往包含了更为系统全面的数据关系,模式单一的公共数据开放路径可能会加剧匿名化信息再识别的可能性,故而匿名化制度的实施并不单纯是匿名化处理的业务流程和技术标准问题,同样也是该项制度与其他数据安全保护制度之间的体系衔接问题。

Abstract: Anonymization technology, as a technology that takes into account both data utilization and data security, has always been difficult to be effectively put into practice, because Article 73 of the Personal Information Protection Law sets two qualifying conditions for the concept of "anonymization"—"unidentifiability" and "irrecoverability", which is a harsh condition that puts the anonymization system in a state of being just on paper. Although there is a consensus that anonymization cannot achieve complete anonymity, there have been debates about the criteria for determining relative anonymity and the scope of re-risk. At the present stage, the theoretical study of the anonymization system generally ignores the discussion of theoretical basis but only keep eyes on the protection of individual rights and interests. Under the paradigm of data relationship theory, the correlation of information elements focuses on the group characteristics of horizontal data relationship, which is precisely the key to the economic value of data. Therefore, the function of anonymization system should be changed from simply guaranteeing the privacy of individual information to reduction of the risk of re-identification on the basis of the economic value of data association. At the level of system construction, the "irrecoverability" requirement of anonymized information should be interpreted in the following way: with *ex ante* risk assessment, the technical difficulty and time cost of recovery are far beyond the acceptable range of general public.

Key Words: Anonymization; Data Relationship Theory; the Risk of Re-Identification; Personal Information Protection

(责任编辑:彭　錞)