arXiv:2303.00845v1 [stat.AP] 1 Mar 2023

目录

1 21^{st} 世纪统计披露限制：动机与挑战

3 John M. Abowd 和 Michael B. Hawes 1.1 动机.4 1.1.1 历史视角下的统计披露限制.5 1.1.2 21^{st} 世纪的披露风险.7 1.2 考量因素.9 1.2.1 官方统计的三重权衡 . . . . . 9 1.2.2 评估可接受风险时保持一致性的必要性 . 10 1.3 挑战.10 1.3.1 用例的识别与优先级排序 . . . . . 11 1.3.2 确定“最优”的隐私损失预算 . . . . 12 1.3.3 统计披露限制在数据生命周期工作流程中的应用 . . . . 12 1.3.4 关于统计披露限制方法及其影响的沟通 . 13 1.3.5 在共享行政数据和第三方数据的背景下治理披露风险.14 1.4 原则与最佳实践.14 1.4.1 首先承认披露风险客观存在 . 15 1.4.2 应在数据产品发布的整个生命周期中考虑统计披露限制，而非零敲碎打.15 1.4.3 在数据生命周期早期开始统计披露限制规划 . . . . 16 1.4.4 统计披露限制的治理（但不一定是实施）需要集中化.16 1.4.5 让关键内部利益相关者参与决策并为其提供培训.16 1.4.6 在作出决策之前与外部合作伙伴和数据用户沟通并为其提供培训，以支持关键决策 17 1.4.7 认识到逐步现代化披露控制是可行的，而且可能是必要的.17

参考文献 **19**

1

21^{st} *Century Statistical Disclosure Limitation: Motivations and Challenges*

John M. Abowd 和 Michael B. Hawes 美国
人口普查局

目录

1.1 动机.4 1.1.1 统计披露限制的历史视角.5 1.1.2 21^{st} 世纪的披露风险.7 1.2 考量因素.9 1.2.1 官方统计的三重权衡.9 1.2.2 评估可接受风险时保持一致性的必要性 . .. 10 1.3 挑战.10 1.3.1 用例的识别与优先级排序.11 1.3.2 确定“最优”的隐私损失预算 . .. . 12 1.3.3 数据生命周期工作流程中的统计披露限制 . .. . 12 1.3.4 关于统计披露限制方法及其影响的沟通 . .. 13 1.3.5 在共享行政数据和第三方数据的背景下治理披露风险.14 1.4 原则与最佳实践. 14 1.4.1 首先承认披露风险的现实存在 . .. 15 1.4.2 应在数据产品发布的更广泛生命周期中考量统计披露限制，而不是零散地处理.15 1.4.3 在数据生命周期早期开始统计披露限制规划 . .. . 15 1.4.4 统计披露限制的治理（但不一定是实施）需要集中化.16 1.4.5 让关键内部利益相关者参与决策（并对其进行培训）.16 1.4.6 在作出决策前与外部合作伙伴和数据用户接洽并对其进行培训，以便为关键决策提供信息 17 1.4.7 认识到对披露控制进行渐进式现代化是可行的，也可能是必要的.17 致谢与免责声明.17

在未来十年，国家统计局很可能会对其数据保密计划进行重新设计，其规模可与20世纪70年代开始的统计披露限制（SDL）转型相媲美。Fellegi [13] 和 Delenius [7] 引入了基于原则和科学的统计披露限制方法，从根本上重塑了统计机构在公开数据发布中评估和控制披露风险的方式。在随后的几十年中，机构持续改进并强化统计披露限制的实施，但这些变化主要是对20世纪70年代和90年代开创的方法进行渐进式调整和扩展。如今，计算能力的提升、强大优化算法的发展，以及丰富的第三方数据日益普及，共同形成了新的数据保护环境，使过去几十年广泛使用的统计披露限制方法日益脆弱。面向 21^{st} 世纪的统计披露限制现代化并不容易，也不会毫无争议。这不仅要求统计机构重新思考其整体统计披露限制方法及其在更广泛数据生命周期中的定位，还要求机构和数据用户共同就官方统计的内容与形式，以及数据用户如何访问这些统计数据，作出艰难决定。

本章探讨机构现代化官方统计统计披露限制方法的动机与必要性，并讨论这一现代化对机构更广泛的数据治理和决策的影响，同时指出机构在这一过程中可能面临的挑战。最后，我们提出一些原则与最佳实践，希望借此指导机构推进其保密计划的转型。

1.1 动机

任何国家统计局的核心使命，按定义而言，都是生产官方统计数据。关于一国人民和经济的准确、客观统计数据，是“民主社会信息系统中不可或缺的要素”，也体现了“公民享有公共信息的权利”（经济合作与发展组织，2005年，原则1） [21]。官方统计数据如能得到恰当使用，便可为众多有益活动提供关键依据，包括政府决策、私人投资和科学研究等。但要描绘出丰富的社会图景，就需要有关一国人民、企业、健康、安全和组织的详细信息；如果这些信息遭到滥用，对于信息所收集自或涉及的数据主体而言，可能造成极大伤害。为促进获取统计信息所带来的社会效益，各国通常授权统计办公室获取并发布非统计（或行政）机构的汇总数据，并往往鼓励将这些行政数据与统计机构直接收集的信息合并使用，

这些行政数据与统计机构直接收集的信息合并使用。为防止滥用为官方统计目的而生成的信息，各国通常会设立两项立法制度。首先，制定严格的保密法律，限制获取为统计目的而收集的可识别信息，并禁止发布此类信息。其次，在政府范围内（包括统计办公室），对所收集的信息类型及其允许用途实施监管。

因此，为履行编制官方统计数据的使命，机构面临着相互制衡的义务：既要编制符合允许的数据使用规定的统计摘要，又要确保在机构发布的出版物中妥善保护数据主体的可识别信息。为此，机构会对拟发布的统计数据进行项目审查和披露风险评估。它们将统计分析限制在获批项目范围内，并在必要时应用 SDL方法，以降低披露风险。

1.1.1 从历史视角看统计披露限制

统计披露限制的目标是在不揭示总体中任何特定个人/实体的特征，或与这些特征相关的身份的前提下，生成对总体特征的详细而准确的摘要。统计披露限制方法的复杂程度差异很大，但即使是聚合或删节这类最简单的统计披露限制机制，也可能因具体情境而有效。然而，通常而言，机构拟发布的汇总统计越细致，或个体级记录越详尽，就越容易识别出特定数据主体，而这些简单的统计披露限制机制降低披露风险的效果也越差。几十年来，随着官方统计数据在规模、范围和细节程度上不断增长，机构通常不得不依赖日益复杂的统计披露限制机制来管理这种风险。

从历史上看，统计机构所遵循的大多数保密法，以及机构为履行这些保密义务而依赖的许多统计披露控制框架，都建立在这样一种假设之上：一旦对可识别数据妥善实施统计披露限制，所得数据便会“去标识化”，不再存在披露风险。同样，从历史上看，权威技术手册也强化了这一观点，尽管其中恰当地提醒了风险限制 [9] 或控制 [23]方面的注意事项。两本教科书都没有提供一种方法，用于核算来自同一机密来源的多次发布（包括表格数据和微观数据）所产生的全局披露风险。

这种将数据二分为机构必须保护的可识别数据，以及无需进一步考虑即可向公众发布的去标识化数据的做法，十分便利。它使统计数据的法律保密框架显得界限分明、绝对明确，也简化了机构对披露审查流程的治理，促使数据保管者针对每种数据产品单独采用统计披露限制，遵循“一经发布，便不再过问”的模式。遗憾的是，认为只要正确运用

统计披露限制，就能将数据去标识化的假设是错误的 [14]。发布任何源自保密来源的统计数据，都会使可识别信息遭披露的风险有所增加。自20世纪70年代以来，人们原则上一直承认这一令人遗憾的事实 [12]，并且该事实已于 2003 [8] 得到数学证明。

在官方统计数据发布中应对披露风险这一现实并非易事。尽管大多数保密法仍然维持着可识别数据与去标识化数据之间错误的二分法，但过去五十年来，大多数统计机构已在不同程度上转向披露风险管理框架。

在传统统计披露限制方法的背景下，并结合机构的数据生产周期，为官方统计采用披露风险管理框架远非易事。统计机构通常会基于同一份机密数据，依次发布多个数据产品。然而，大多数传统统计披露限制方法无法客观量化多次数据发布所带来的披露风险；而能够做到这一点的方法，如果未能毫无例外地应用于机密数据源中发布的每项统计数据，也会失效。此外，机构快速而紧凑的数据生产计划等实际情况，更倾向于在发布前的短时间内，依据规则逐次应用统计披露限制。

因此，在过去半个世纪的大部分时间里，大多数统计机构主要依赖主观评估，来判断不同类型数据产品所涉及的总体披露风险，并制定规则和手册，说明在各种情境下应如何应用统计披露限制。这些做法倾向于对大多数汇总表格统计数据采用相对简单的舍入和／或抑制规则，而对微观数据发布则采用更复杂的抑制、粗化或噪声注入规则。

一些统计机构，例如美国人口普查局，通过定期对特定数据产品开展重新识别研究，并利用研究结果调整或补充后续数据发布所遵循的统计披露控制规则，以加强这种披露风险管理方法。然而，这些重新识别研究的问题在于，它们高度依赖于对哪些类型的攻击可行，以及攻击者可能掌握哪些资源（例如计算基础设施和外部信息）的具体假设。因此，这些研究至多只能提供某一特定时刻数据产品总体披露风险的有限概况。

Cynthia Dwork 在微软研究院的工作 [10] 为解决通过引入首个形式化隐私的SDL框架——差分隐私——来控制累积披露风险这一问题打开了大门。她用简洁易懂的语言概括了形式化隐私：“我们将首先定义隐私目标，然后探讨在满足这些隐私目标的前提下能够实现怎样的效用”（第2页）。形式化隐私系统为保密性泄露提供了数学定义。能够被证明满足该定义的SDL方法被称为实施机制。例如，差分隐私领域的奠基性论文 [11] 将保密性泄露定义为

用户在最坏情况推断中判断某个实体是否被纳入统计制表时，判断结果增加超过 ϵ 。这一框架如今被称为纯差分隐私，或 ϵ差分隐私。他们证明，加入尺度参数取决于 ϵ 以及统计制表中最坏情况下可行变化量的双指数噪声，即可满足纯差分隐私的定义。我们如今称之为通过拉普拉斯机制实施纯差分隐私。本文所称的形式化隐私方法，是指能够为保密性泄露提供数学定义，并至少有一种满足该定义的可行实施机制的SDL框架。

去标识化这一方便的假设，以及随后采用基于规则的方法进行披露风险管理，所带来的另一个后果是：公众和机构负责人普遍认为，关于统计披露限制的决策属于机构内技术型统计披露限制从业者的职权范围。因此，历史上公众很少讨论或辩论可接受的披露风险程度，以及统计披露限制方法对数据效用的影响。[1] 尽管有效统计披露限制的设计与应用是一门高度技术性的学科，但披露风险评估与缓解背后的决策具有重大的公共政策影响。本文将探讨，关于披露风险的审议如何能从以下方面获益匪浅：在机构层面协调制定适用于全机构的统计披露限制框架和参数；与内部主题专家协调；以及参考外部利益相关者反馈制定指导方针。

1.1.2 21 世纪的披露风险

对于过去几十年来经验丰富的统计披露限制从业者而言，上文讨论的机构传统统计披露限制方法的种种局限，实际上并不令人意外。Fellegi 本人也承认，在不同表格或数据发布中应用基于规则的统计披露限制方法存在脆弱性 [13]。但机构普遍认为，按数据类型制定基于规则的方法，并在每次发布时采用，既有益又简单，足以证明由此增加的披露风险是合理的。而且，已开展的重新识别研究大多支持这一结论。

然而，过去十年的技术发展使传统统计披露限制方法的局限性变得更加清晰，并揭示了基于数据类型的规则式统计披露控制存在的根本问题。以往的披露风险评估主要侧重于针对表格数据的简单相减攻击，以及针对微观数据产品的基于记录链接的重新识别攻击。这导致机构几乎普遍对表格数据产品采用一套统计披露控制规则，并采用

[1]后一点在某些统计披露限制（SDL）方法的背景下进一步恶化（例如，对于某些置换和噪声注入机制），因为传统做法是对这些方法的参数及其对数据准确性的影响予以保密，以避免削弱这些机制。

第二套统计披露控制规则（通常但并非总是更为严格）用于微观数据发布，因为其潜在的披露风险通常被认为更高。

大规模云计算平台和强大的优化算法的可用性，使得数据库重构攻击成为可能并广泛普及；这种攻击最初由 Dinur 和 Nissim [8], 预测，为官方统计开辟了一种传统披露风险评估未曾考虑过的新攻击途径。人口普查局对已发布的2010年人口普查汇总数据进行的模拟攻击表明，可以生成用于编制这些表格的微观数据的高质量重构记录图像（对于超过65%的普查区块，图像完全准确）。根据当时生效的统计披露控制规则，这些微观数据被认为不可发布。 2010 [1]这次模拟攻击表明，至少对于大规模表格数据发布而言，汇总表格数据的统计披露限制不能与微观数据的统计披露限制分开看待。数据量足够大时，汇总统计可能等同于微观数据。

传统的披露风险评估低估了表格数据相关的披露风险，因为其基本假设没有考虑到基于重构的攻击途径。同样，如果机构认为，在新兴或未来攻击途径出现时，借助重构的再识别研究不会低估总体披露风险，那就未免太过天真。例如，推断攻击和成员资格攻击是机构应予以考虑的新兴攻击途径。Nissim[20] 提出了一个框架，可将其中许多攻击与机构保密法的法律要求联系起来。

从数据库重构成为攻击官方统计的可行手段这一现象中，我们唯一应汲取的教训是：只有对多次发布（无论发布的是统计数据、表格还是微观数据产品）所带来的增量披露风险进行量化和控制，才能有效管理披露风险。传统统计披露限制方法大多缺乏这种对多次数据发布的可组合性，因此随着时间推移，它们越来越可能失效。然而，基于形式化隐私框架的统计披露限制方法能够让机构精确量化每项已发布统计数据带来的增量披露风险，因为其数学定义具有可证明的组合性。这种组合性质意味着，此类框架提供了一套核算系统，可用于对多个数据产品的增量披露风险进行量化。这并不是说，采用形式化隐私就能消除官方统计中的披露风险。恰恰相反，形式化隐私统计披露限制可以在从完全规避披露风险到完全接受风险的整个范围内实施，具体取决于机构作出的决策。对增量风险进行精确核算，正是为决策提供信息并使决策成为可能的关键。

此外，尽管按照惯例或出于便利，统计披露限制的传统方法大多是技术官僚式的，且对数据用户而言并不透明，但正式采用隐私保护型统计披露限制方法所能实现的透明度不应被低估。统计学作为一门学科所遵循的原则之一，是强调

应当公开数据或分析中已知局限的重要性 [5]。任何统计披露限制方法的选择与实施，必然会影响最终数据的适用性。统计机构应向数据用户透明说明这些影响。

1.2 考量因素

要将统计披露限制的选择从幕后决策带入公众和专业人士关于适当权衡的讨论，就必须承认，全球保密风险管理涉及三项选择，而非两项。

1.2.1 官方统计的三重权衡

关于统计披露限制，尤其是形式化隐私机制的公开讨论，往往聚焦于隐私与准确性／效用之间权衡的核心地位 [16]。这种权衡无疑至关重要，也是解读形式化隐私统计披露限制实施中的隐私损失核算（以及相应的适用性）的基础。但若只关注隐私与准确性之间的权衡，就会同时过度简化并使决策变得复杂。在统计披露限制的背景下，准确性或数据效用并非单一维度的特征——不存在评估数据效用的单一通用指标。更根本地说，从隐私与准确性相权衡的视角来进行统计披露限制决策，会使另一个重要维度变得复杂：待发布统计数据的数量。同一隐私损失预算可以分配给少量高准确性的统计数据，也可以分配给大量准确性较低的统计数据。两种选择的效用，取决于这些统计数据旨在支持哪些应用。

Dinur 和 Nissim [8] 证明，发布过多准确的统计数据会损害保密性。理论上，统计机构只要发布少量统计数据，就可以在几乎没有披露风险的情况下发布高度准确的统计数据。因此，在官方统计背景下，围绕统计披露限制的公共讨论和机构决策，不应聚焦于固定维度统计输出中的隐私与准确性权衡，而应关注更广泛的三重权衡：隐私与准确性、可用性之间的权衡 [15]。在三重权衡中，最大化任何一个维度都很容易，最大化其中两个维度也办得到，但必然会对第三个维度造成重大影响。因此，机构可以发布大量准确的数据，但这会带来侵犯保密性的重大风险。或者，它们可以在充分保护保密性的同时发布大量数据，但代价是显著降低由此产生的数据在至少某些应用中的效用。又或者，它们可以开展彼此独立的调查来估算不同的统计数据，从而确保高度保密性，但

如果调查之间分摊相同的运营成本，准确性就会降低。最后，它们可以发布高度准确且得到高度保护的数据，但只能大幅削减所发布统计数据的数量。采用恰当实施的主要抑制/互补抑制方法的机构 [6], ，例如美国人口普查局发布的经济普查数据产品，已经大幅削减了发布数量。数据数量这一维度还涉及“对谁有用？”的问题。尽管这些维度的研究仍处于起步阶段，但在形式化隐私框架下，可以不均等地将预算分配给不同查询，这意味着针对某些推断，数据的不同特征会获得不同的隐私保证 [18]。在更细粒度的表格中决定优先考虑数据的哪些维度，势必会限制发布数据的效用。

机构要在相互制衡的目标之间取得平衡：既要生成准确且与社会决策相关的统计数据，又要保护保密性，因此必须在隐私、准确性和数据数量这三个维度上作出妥协。这三个维度之间的相互作用凸显了高级领导层参与统计披露限制相关决策的重要性，而不应将其视为仅属于负责设计和实施保护措施的统计披露限制从业人员的专属领域。

1.2.2 评估可接受风险的一致性需求

统计披露限制从所谓的数据去标识化演变为披露风险管理，其核心在于假定存在一个可接受的披露风险水平。在不承认任何数据发布都固有披露风险的保密法背景下，并没有衡量可接受风险程度的客观标准。机构必须自行作出判断，权衡拟发布数据的公共价值、对自身保护保密性的法律和道德义务的理解，以及自身对潜在披露事件的容忍或规避倾向。由于披露可能造成的危害因所涉数据类型（例如财务、人口统计和健康数据）而有很大差异，要确保不同数据产品的披露风险评估与缓解具有可比性和一致性，就需要通过全机构治理来规范披露风险评估的实施方式。

1.3 挑战

如上文所述，要应对官方统计的三重权衡，机构必须就其数据产品的范围和内容作出重要决策，权衡数据用户的需求和官方统计的公共价值，以及保护数据主体信息保密性的法律与伦理要求。

随着机构对其统计披露限制的实施进行现代化改造，并转向采用形式化隐私核算来评估披露风险，它们将面临一系列与三重权衡下所需决策相关的挑战。

这些挑战并非形式化隐私所独有。数十年来，负责设计和实施各机构统计披露限制机制的SDL技术实践者一直以某种方式应对着其中每一项挑战。然而，形式化隐私解决方案在即将发布的全部数据产品中的量化与可组合性，使许多挑战更加突出，也使以原则为基础的决策——即如何应对这些挑战——变得更加重要。与一些传统的SDL方法相比，形式化隐私所带来的透明度机遇，也使人们能够以公开评估和讨论的方式，在这些挑战的背景下进行决策；这种方式既可以为相关决策提供信息，也可能使决策更加复杂。

1.3.1 用例的识别与优先排序

任何SDL框架及其实施机制的选择与实施，都必然会降低最终数据的潜在效用。选择使用哪种统计披露限制机制，以及在不同机制参数之间进行取舍，都会直接影响最终数据对于不同用例的效用。要使每一种潜在用例的效用最大化，唯一的办法就是完全放弃SDL，在不采取任何保密保护的情况下发布数据 [8]。因此，任何降低披露风险的尝试，都必然意味着某些可以设想的用例会受到负面影响。尽管官方统计被广泛视为一种公共产品，人们也常常高度重视其对于尚未确定的分析和研究问题的价值，但要以原则为基础对SDL的框架和实施进行决策，机构就必须识别并确定其数据产品的预期用例及其优先级。为支持这些决策而进行有效的优先排序，将产生总体效用和社会价值最高的数据产品。然而，如果未能识别并优先考虑重要用例，最终的数据产品可能会对任何可以设想的用例都表现平庸。

全面识别并确定各类用例的优先级，尤其是针对旗舰官方统计产品的用例，可能是一项艰巨的任务。有些重要用例显而易见，但机构往往不了解其数据产品那些不太显眼却同样重要的下游用途。尽管如此，要最大限度地提升官方统计的整体社会价值，机构就必须齐心协力，全面了解其数据目前或未来的各种用途。然而，识别这些用例还不够。根据用途的多样性，机构必须做出选择，优先满足某些用途的效用，而非其他用途。考虑到三重权衡，在保密性保护水平既定的情况下，所发布统计数据的数量以及这些统计数据对不同用途的效用都是有限资源，

需要谨慎分配和管理。例如，评估用于城市规划的统计数据与用于公共卫生的统计数据各自具有的社会价值可能并不容易；这些不同的用例对数据效用的衡量方式可能大相径庭，因此，在权衡三重取舍时，可能有必要判断哪种用途更为重要。

1.3.2 确定“最优”的隐私损失预算

尽管统计保密法往往依赖于可识别数据与去标识化数据之间简单却不切实际的二分法，但现实是，任何基于保密的可识别数据得出的统计数据，一旦公开发布，都必然会带来一定的披露风险 [8]。从风险管理视角着手设计和实施统计披露限制，意味着承认一定程度的披露风险是可以接受的，以支持拟发布官方统计数据所具有的社会价值。与许多传统统计披露限制方法中常见的、更具主观性的风险评估方法相比，形式化隐私核算有助于在这一风险管理框架下，对披露风险进行审慎、可量化的评估和缓解。

尽管形式化隐私能够精确量化披露风险，却无法帮助回答“可接受的披露风险有多大？”这一问题[3]。机构再次面临三重权衡下的艰难决策。加强保护会降低披露风险，从而减少数据主体可能遭受的损害，以及机构因披露而承受的法律或声誉损害。但这必然会降低数据的社会价值，因为可发布数据的数量和/或效用会减少。另一方面，接受更高的披露风险容忍度能够提升数据的整体社会价值，但也会增加数据主体可能遭受的个人损害。若缺乏有关可接受披露风险的法律或监管标准，以及有效的用例优先级排序，机构应接受多大风险这一问题便没有明确答案。归根结底，统计披露限制决策中公共利益与个人损害之间的适当平衡，属于公共政策决策。与其他有争议的公共政策决策（例如“枪炮还是黄油”、控制通胀还是实现充分就业）一样，机构若能在自身专业判断的基础上，听取公民社会各界的意见，将会更有裨益。

1.3.3 数据生命周期工作流程中的统计披露限制

官方统计数据的生产是一个复杂但一体化的工作流程，涵盖数据收集、处理、统计计算和传播系-统。在这一数据生态系统中，统计披露限制通常被作为统计计算与传播阶段之间的一道过滤环节。尽管从数据产品设计和运营效率的角度来看，这通常对机构较为便利，但将统计披露限制的决策和实施孤立起来

在数据生产周期中相对较晚的这一阶段，可能会显著限制SDL方法在三重权衡中为优先用例优化数据效用的灵活性。如果仅在数据产品设计规范完成后才考虑SDL（即在表格框架和报告类别最终确定之后），那么统计披露限制从业者就已经失去了许多原本可用于应对必须作出的隐私—效用—数量权衡的手段。然而，在产品生命周期早期纳入SDL框架的组成部分，可能会提高SDL解决方案灵活应对这些权衡的能力，并优化最终数据的整体社会价值和保密性。这样做也可能需要中断、修改，甚至重新设计机构经过精细调校的数据生产系统。对于在持续而紧凑的生产计划下运行的数据产品而言，这一点尤其可能带来问题。不过，即使是这些系统也会定期进行重新工程设计——例如，作为当前信息技术倡议的一部分，所有人口普查局系统都必须进行重新工程设计 [22]。适合企业级披露风险管理的披露限制框架可被纳入此类重新工程设计之中。

1.3.4 关于统计披露限制方法及其影响的沟通

统计学伦理实践要求统计机构识别并传达其所生成数据中任何可能存在或已知的限制，因为这些限制可能影响数据用户对数据的分析或解读；因此，在沟通SDL决策时提高透明度，必须成为任何现代化工作的一部分。 [5],然而，历史上统计机构往往只提供相对笼统或抽象的信息，说明所采用的统计披露限制方法，而几乎不提供这些方法对数据效用的影响信息。 [2]在某些情况下，例如采用抑制或粗化方法时，数据用户可以发现或推断所用方法的大致参数。然而，对于某些方法，例如数据交换和其他形式的扰动，所得数据对普通数据用户而言可能根本看不出经过了统计披露限制处理。无论选择何种框架和机制，应用统计披露限制总会降低数据对特定用例的适用性。即使是抑制程序——可以说是所有统计披露限制方法中最显而易见、透明度最高的方法之一——也可能带来数据用户不易察觉的数据使用限制，例如不可忽略的缺失。 [19, 2]另一方面，数据交换和其他传统上所谓的扰动方法在这方面甚至更具隐蔽性，因为按照惯例，为了维护机制的完整性，其参数（例如交换率）以及对最终数据的影响都应予以保密。

形式化隐私方法为机构提供了一个机会，使其能够就机制的设计、规格以及对数据效用的影响实现彻底透明，但这种透明度也带来了自身的挑战。这些方法的技术复杂性以及相对新颖性，使得难以向非技术受众有效地传达相关信息。事实上，

即使是经验丰富的技术用户，要将此类信息纳入分析也并非易事。在许多情况下，要有效传达最终数据的局限性，首先必须有效地介绍这些方法并开展相关教育。

1.3.5 在共享行政记录和第三方数据的背景下治理披露风险

统计机构始终关注其信息收集工作的相对负担，包括财务成本和受访者所耗费的时间。随着数据收集成本上升、调查响应率下降，机构越来越多地转向使用行政记录和第三方数据，以补充或替代其直接数据收集活动。美国旨在支持加强循证决策的倡议，也推动了行政记录的扩大使用以及统计机构之间更多的数据共享，尽管无论是相关法规（《美国法典》第44编第3561至3564条）还是配套条例（未完成），都未能消除此类数据共享面临的其他法律限制（例如《美国法典》第13编第8(b)条、第9条），也未能统一有关数据访问的规则。行政记录和第三方数据的扩大使用，以及机构之间的数据共享，对统计披露限制总体而言构成了重大挑战，也对形式化隐私框架和传统统计披露限制框架构成了挑战。当多个机构使用并发布来自同一来源的数据时，即使将决策过程扩展为由所有贡献机构充分参与，如何才能有效评估和降低披露风险？在基于规则的统计披露限制框架下，答案很直接：提供数据的机构会规定发布时必须采用的统计披露限制机制（例如，美国国税局出版物 1075 [17]）。然而，传统的基于规则的统计披露限制机制通常无法量化多次发布之间的增量披露风险——它们无法进行组合。形式化隐私披露风险核算或许能够提供解决方案，因为它可以在各机构的数据发布之间进行全局隐私损失核算，但这要求两个机构都同意使用兼容的形式化隐私框架和实施机制。对总体隐私损失预算的治理也可能构成挑战，尤其是在各机构对不同数据产品和/或用例的相对重要性持有不同看法的情况下。

1.4 原则与最佳实践

面向 21st 世纪的统计披露限制现代化，要求统计机构转向形式化隐私方法，以便量化并考虑多次数据发布带来的增量披露风险。若不这样做，

官方统计将越来越容易受到已知和新兴攻击途径的威胁。然而，统计机构转向形式化隐私很可能是一个漫长而艰巨的过程。美国人口普查局近期在将十年一次的人口与住房普查的统计披露限制转向形式化隐私方面积累的经验，提供了许多重要经验教训，可帮助各机构顺利完成这一转型。

1.4.1 首先承认披露风险确实存在

官方统计无法回避的三重权衡意味着，统计机构必须在所发布数据的数量、效用和保密性之间作出艰难抉择。作出这类决策的前提是承认：官方统计中的披露风险确实构成威胁，只能加以管理而无法彻底消除；任何缓解披露风险的措施，都必然会影响可公开发布的数据数量和／或效用。对于可接受的披露风险程度，人们难免会有合理的争论；对于应优先考虑哪些用例，也难免会产生冲突；而对于应在何处以及如何实施保密保护，人们的意见也可能不尽相同。然而，若忽视这一根本事实——公开发布大量源自机密可识别信息的统计数据，在不同程度上必然会造成信息披露——只会妨碍人们依据原则作出决策，无法妥善管理披露风险，也无法在官方统计的社会价值与披露可能对个人造成的伤害之间取得平衡。

1.4.2 在数据产品发布的更广泛生命周期中考虑统计披露限制，而非零敲碎打地处理

在多次连续数据发布的情境下，任何统计披露限制方法的灵活性本质上都取决于此前的选择。针对较早期数据产品所作的选择，无论是选用何种统计披露限制方法，还是决定支出多少隐私损失预算，都会限制后续数据产品可采用的统计披露限制方案的灵活性。采用形式化隐私披露风险核算，有助于机构评估并缓解多次数据发布中的披露风险，但机构需要谨慎管理这些发布所涉及的隐私损失预算。在机构的全部数据产品发布之后很久，研究人员往往才会发现机密数据具有高价值、创新且未曾预见的用途。支持这些用途不应自动以降低机密数据保护水平为代价（即超出此前设定的全局隐私损失预算）。相反，机构应在整个数据生命周期内，围绕三重权衡主动考虑这些未曾预见用途的可能性：选择合适的隐私损失预算适用期限，并为未预见的项目预留部分预算。

1.4.3 在数据生命周期早期启动 BeginSDL 规划

数据产品规范的细微变化，可能对数据最终产生的披露风险造成巨大影响。减少（或增加）报告类别，以及移除（或添加）交叉表，都会极大影响缓解披露风险的难易程度。1995年，隐私领域开始在商业技术背景下倡导“隐私保护设计” [4]。其理念很简单：在产品开发流程早期纳入有效的隐私保护措施，远比在已开发完成的产品中发现隐私问题后再加以解决容易。同样，如果在数据产品生命周期早期、尚未对数据产品规范作出最终决定之前就着手处理隐私性／效用／数量权衡，有效管理这一权衡也会容易得多。

1.4.4 统计披露限制的治理（但不一定是实施）需要集中化

长期以来，官方统计中有关统计披露限制的许多决策都被视为机构披露审查委员会（披露审查委员会）和统计披露限制从业者的技术职权范围。披露审查委员会和训练有素的统计披露限制专家的重要性怎么强调都不为过：评估和缓解披露风险是一门高度专业的技术学科，要有效开展这项工作，需要专门的专业知识。尽管如此，上文指出的诸多统计披露限制现代化挑战中，一个反复出现的主题是：有关统计披露限制的决策具有公共政策属性。关于可接受的披露风险程度，以及如何妥善平衡相互竞争的数据用例，应由机构领导层以协调、集中的方式作出决策，并为统计披露限制专家提供明确指导，以便他们予以实施。这并不意味着统计披露限制的所有应用都必须在机构内集中进行——许多统计机构在不同业务部门设有多位统计披露限制专家，他们各自专门了解所属领域的数据产品——而只是说，披露风险评估与缓解的治理应当集中化，以确保机构的优先事项在整个组织内得到一致贯彻。

1.4.5 让关键内部利益相关者参与决策（并对其进行培训）

由于存在三重权衡，围绕统计披露限制（SDL）的决策与有关数据生产和传播的决策密不可分。因此，在平衡官方统计的社会价值与保障数据主体信息的法律和伦理责任时，任何决策都应充分听取机构内不同利益相关者的意见。机构的政策制定者、隐私事务官员、法律顾问、数据管理专家和数据主题专家都将发挥重要

作用，阐述如何平衡机构相互冲突的责任。然而，要通过这些利益相关者之间的有效对话为机构领导层的决策提供依据，就必须让这些不同领域的人员具备关于披露风险评估、风险缓解，以及统计披露限制（SDL）所涉及的数据可用性与效用权衡的基础知识和理解。

1.4.6 在做出决策之前，与外部合作伙伴和数据用户沟通并向其普及相关知识，以为关键决策提供依据

官方统计只有得到使用才有价值。管理官方统计中的披露风险，必然会影响其对社会的价值。同样，机构在选择和实施统计披露限制（SDL）方面所做的决策，可能会对不同群体产生分配影响，影响数据产品发布所带来的价值或潜在危害。对这些问题做出有效决策，应尽可能全面地了解这些影响。为确保在做出这些艰难决策时能够考虑到各种重要观点，机构应努力在做出任何SDL决策之前，征询各类外部合作伙伴和数据用户的意见。与这些外部利益相关者进行有效对话时，他们可能并不熟悉披露风险评估和缓解这一技术领域，因此可能需要先向他们普及相关问题，以及这些问题与数据可用性和效用之间的关系。

1.4.7 认识到披露控制的渐进式现代化是可行的，而且可能是必要的

披露风险格局始终在变化，统计披露限制的现代化应被视为一段旅程，而非一个终点。形式化隐私解决方案为应对许多统计披露限制传统方法固有的一些棘手漏洞，提供了巨大的潜力。但形式化隐私也是一门相对较新的学科，要充分发挥其在种类繁多的统计数据产品中的潜力，还需要大量研究与开发。因此，机构不应将统计披露限制的现代化视为“全有或全无”的举措。有时，考虑到生产计划、运营要求或资源限制，可能有必要推迟对机构统计披露限制制度进行更全面的转型，转而逐步改进现有方法。只要这一决定以有原则的方式作出，依据现有最佳研究成果以及内部和外部利益相关者的意见，并充分理解其可能带来的风险与收益，那么这样做绝不意味着机构放弃了保护数据主体信息的义务。

致谢与免责声明

我们感谢 Gary Benedetto、Monique Eleby 和 Sallie Keller 提供的宝贵意见。本章所表达的观点属于作者本人，并不代表美国人口普查局。本章即将收录于《CRC 形式化隐私与合成数据统计披露控制方法手册》。

参考文献

[1] John M. Abowd 和 Michael B. Hawes。《2020年美国人口与住房普查中的保密性保护》。《统计学及其应用年度评论》，10(1)：null，2023年。[2] John M. Abowd 和 Ian M. Schmutte。经济分析与统计披露限制。《布鲁金斯经济活动论文集》，第 221–293 页，2015年。[3] John M. Abowd 和 Ian M. Schmutte。作为社会选择的隐私保护与统计准确性的经济分析。《美国经济评论》，109(1)：171–202，2019年1月。[4] Ann Cavoukian。隐私保护设计：权威研讨会。Ann Cavoukian 博士作序。《信息社会中的身份》，3(2)：247–251，2010年8月。[5]美国统计学会专业伦理委员会。《统计实践伦理指南》。https://doi.org/10.5281/zenodo.7092386，2022年。访问日期：2022-09-13。[6]Lawrence Cox。互补单元格抑制的网络模型。《美国统计学会会刊》，90(432)：1453–62，1995年。[7] Tore Delenius。迈向统计披露控制方法论。《Stat Tidskrifts》，15：429–44，1977年。[8] Irit Dinur 和 Kobbi Nissim。在保护隐私的同时披露信息。载于《第二十二届 ACM SIGMOD-SIGACT-SIGART 数据库系统原理研讨会论文集》，PODS ’03，第 202–210 页，纽约州纽约市，2003年。美国计算机协会。[9] George T. Duncan、Mark Elliot 和 Juan-Jos´e Salazar-Gonz´alez。《统计保密：原理与实践》。施普林格纽约，纽约州纽约市，2011年。[10] Cynthia Dwork。差分隐私。载于 Michele Bugliesi、Bart Preneel、Vladimiro Sassone 和 Ingo Wegener 编，《自动机、语言与编程》，第 1–12 页，柏林、海德堡，2006年。Springer Berlin Heidelberg。

[11] Cynthia Dwork、Frank McSherry、Kobbi Nissim 和 Adam Smith。为隐私数据分析中的敏感度校准噪声。载于 Shai Halevi 和 Tal Rabin 编，《密码学理论》，第 265–284 页，柏林、海德堡，2006年。Springer Berlin Heidelberg。[12] 联邦统计政策委员会。《统计政策工作论文 2：统计披露与披露规避技术报告》。技术报告，1978年。[13] Ivan P. Fellegi。论统计保密问题。《美国统计学会会刊》，67(337)：7–18，1972年。[14] Simson Garfinkel、Phyllis Singer、Joseph Near、Aref Dajani 和 Barbara Guttman。《政府数据集去标识化》。技术报告，2023年。第三稿访问日期：2023年2月15日。最终文件预计于2023年3月提交。[15] Michael B. Hawes。促进数据访问与数据保护：新兴技术如何帮助机构应对数据三重约束。
https://www.bea.gov/system/files/2021-03/Michael-Hawes.pdf，2021年。向联邦证据建设数据咨询委员会所作的报告，2021年3月19日。[16] V. Joseph Hotz、Christopher R. Bollinger、Tatiana Komarova、Charles F. Manski、Robert A. Moffitt、Denis Nekipelov、Aaron Sojourner 和 Bruce D. Spencer。在联邦统计系统中平衡数据隐私与可用性。《美国国家科学院院刊》，119(31)：e2104906119，2022年。[17] 美国国税局。《出版物 1075：联邦、州和地方机构税务信息安全指南；保护联邦纳税申报表和申报信息的保障措施》。www.irs.gov/pub/irs-pdf/p1075.pdf，2021年。[18] Daniel Kifer、John M. Abowd、Robert Ashmead、Ryan Cumings-Menon、Philip Leclerc、Ashwin Machanavajjhala、William Sexton 和 Pavel Zhu-ravlev。《差分隐私常见变体的贝叶斯与频率学派语义：在2020年人口普查中的应用》。技术报告，2022年。[19] Roderick J.A. Little 和 Donald Rubin。《缺失数据的统计分析 2^{nd} 版》。John Wiley & Sons，霍博肯，新泽西州，2002年。[20] Kobbi Nissim。隐私：从数据库重构到法律定理。载于《第40届 ACM SIGMOD-SIGACT-SIGAI 数据库系统原理研讨会论文集》，PODS'21，第 33–41 页，纽约州纽约市，2021年。美国计算机协会。

[21]经济合作与发展组织（OECD）。官方统计基本原则。
http://unstats.un.org/unsd/goodprac/bpabout.asp，2002年。[22]Michael Thieme。人口普查局的技术转型：构建现代化、以数据为中心的生态系统。
h
t
t
p
s://
w
ww.census.gov/newsroom/blogs/research-matters/2022/10/technology-transformation.html，2022年10月。[23] Leon Willenborg 和 Ton de Waal。统计披露控制要素。统计学讲义。施普林格，纽约州纽约市，2000年10月。