arXiv:2301.10213v1 [cs.CR] 24 Jan 2023

数据库重构并非易事，且不同于重新识别*

Krishnamurty Muralidhar[1] 和 Josep Domingo-Ferrer[2] 1 俄克拉荷马大学普赖斯商学院市场营销与供应链管理系，西布鲁克斯街307号，亚当斯楼10室，诺曼，俄克拉荷马州 73019，美国。krishm@ou.edu 2 罗维拉-维尔吉利大学计算机工程与数学系，CYBERCAT-加泰罗尼亚网络安全研究中心，加泰罗尼亚大道26号，43007 塔拉戈纳，加泰罗尼亚。josep.domingo@urv.cat

2023年1月25日

摘要

近年来，有人声称，发布数据库的准确统计信息很可能会导致数据库被完全重构。有人提出，差分隐私是防止此类攻击的适当方法。美国人口普查局近来对这些说法予以高度重视，并因此采用差分隐私来发布美国人口普查数据。这引发了人口普查数据用户的忧虑，因为受保护输出的准确性不足；也导致美国商务部面临法律诉讼。本文追溯了这一说法的起源：发布数据库信息会自动使其易受重构攻击，从而暴露数据库。我们证明，这一说法实际上并不正确。我们还表明，正确使用传统的统计披露控制（SDC）技术即可避免重构。此外，与所采用的具体统计披露控制方法相比，发布精确计数时所处的地理层级对保护的影响更大。最后，我们提醒读者不要混淆重构与重新识别：以重构质量作为重新识别的衡量指标，会夸大重新识别风险。

关键词： 数据库隐私、数据库重构、统计披露控制、差分隐私。

1 引言

数据库重构似乎是迄今为止人们所熟知的官方统计和统计数据发布的克星。美国人口普查局首席科学家指出：

> 这项 [Dinur 和 Nissim 的数据库重构]定理，宣告了按照传统方式编制供公众使用的详细制表数据和微观数据集的终结。（Abowd 2017年，Abowd等 2019年）

*谨此感谢欧盟委员会（项目 H2020-871042“ SoBigData++”和 H2020-101006879“MobiDataLab”）以及加泰罗尼亚政府（授予 J. Domingo-Ferrer 的 ICREA Acadèmia奖）的支持。

只要数据库包含一组受访者的个人信息，数据保护法就可能要求负责数据库的组织（在欧洲法律术语中称为控制者 ）（2016年《通用数据保护条例》）采取措施保护受访者隐私。统计披露控制（SDC，（Dalenius 1977，Hundepool等 2012年））是一门提供相关方法的学科。统计披露控制方法通过屏蔽即 更改需要保护的数据来实现这一目的；屏蔽可以基于数据扰动、降低数据细节程度，甚至生成能够保留原始数据部分统计特征的合成数据。根据屏蔽的实施时机，统计披露控制文献将保护分为本地 保护（由受访者在数据收集前自行屏蔽数据）、输入 保护（由控制者在收集数据后屏蔽数据，之后的所有查询均基于屏蔽后的数据作答），以及输出 保护（基于受访者真实的原始数据进行查询，并在发布前屏蔽查询输出）。

数据库的形成过程可能在某个特定时间点完成，也可能在数据库的整个生命周期中持续进行。前一种情况下，数据库称为静态数据库；后一种情况下，则称为动态数据库。在静态数据库中，数据先被收集，然后经过整理以形成数据库。由此得到一个“冻结”的数据库，之后可用它来回答各种查询。相比之下，动态数据库包含不断定期变化的数据，随着时间推移，新的记录甚至属性会被添加或删除。组织和企业数据库（例如，包含客户、订单等数据）通常是动态的。静态数据库常见于为研究而收集的数据，以及政府机构收集的某些数据。显然，并非所有政府数据都是静态的，但在许多情况下确实如此。

对于动态数据库而言，输出保护是最方便的选择，因为它无需为底层不断变化的数据创建经过掩蔽的版本。无论采用何种保护方式，所实现的保护程度都取决于数据被修改的程度（Dwork 等，2006年）。一般而言，修改越多，保护程度越高，但准确性也越低，因此受保护数据的效用也越低。尤其是当原始数据包含离群值或唯一记录时，可能需要进行更大程度的修改。关于统计披露控制和隐私—效用权衡的进一步讨论，参见 Traub 等（1984）、Adam 和 Worthmann（1989）、Duncan 等（2001）以及 Hundepool 等（2012年）。

发起重构攻击的可能性已为人所知数十年，Dinur和Nissim于二十年前提出了重构攻击的形式化理论（Dinur和Nissim 2003）。美国人口普查局的方法学家最近指出，此类攻击已不再只是理论上的可能性，而是现实中的危险。因此，他们主张采用差分隐私（Dwork等2006，Dwork和Roth 2014）——下文简称DP——来保护美国2020年十年一次人口普查的统计输出，以阻止对底层微观数据的重构。采用DP的决定促使阿拉巴马州对美国商务部提起诉讼，基本理由是这会导致效用损失（以及数据发布延迟）（阿拉巴马州 2021）。另外16个州支持了这起诉讼（Associated 2021），但法官最近驳回了该诉讼（Percival 2021），理由是只有在DP人口普查数据公布后，才能证明阿拉巴马州遭受了损害。

2021年4月，人口普查局发布了采用其新型基于差分隐私的方法 DAS 编制的 2010年十年一次的人口普查数据版本。研究该版本后，若干用户表达了对 DAS 所致效用损失的担忧（Kenny 等 2021，Ruggles 和 Van Riper（2021），Hotz 等 2022，Dove 2021，Schneider，2022）。

由于采用基于差分隐私的方法来防止重构，在效用方面颇具争议，因此，考察重构攻击的真实威胁，以及差分隐私或其他方法能在多大程度上以合理的效用代价抵御此类攻击，仍然极具现实意义。

本文的贡献与计划

本文重新评估了攻击者根据受保护的查询输出重构原始数据的风险。我们首先介绍重构攻击的背景（第2节），然后考察差分隐私能够为抵御重构提供何种保护（第3节）。接下来（第4节），我们讨论报告精确计数的地理层级对重构至关重要的影响[1]。在第5节中，我们重点阐明重构与重新识别之间的区别：将重构质量用作

[1]事实上，人口普查局已经发布了选区重划所需计数的正式运行结果（https://www.census.gov/newsroom/press-kits/2021/2020-census-redistricting.html）。

重新识别风险的衡量标准，会导致对重新识别风险的夸大估计。结论和未来研究方向总结于第6节。

2 重构攻击

文献中早已有人研究重构攻击。在开创性工作中，Denning和Schlörer（1980年）指出，设计不当的数据库查询应答系统若采用输出扰动，很容易导致部分甚至全部数据库记录泄露。他们使用的工具是跟踪器攻击：精心构造一系列查询，旨在分离并披露特定记录。

Dinur和Nissim（2003年）（下文简称DN）对仅利用查询输出重构数据库内容的能力进行了更为正式的分析。他们最重要的进展在于发现，攻击者在构造查询序列时甚至无需谨慎。作者假设数据库是一个 n位字符串，也就是说，其中包含的记录取值均为0或1。他们还假设所有查询都采用以下形式："这个子集里有多少条记录的值为0？"或"这个子集里有多少条记录的值为1？"在他们的设定中，每个查询的响应都等于真实查询答案加上一个误差 E，该误差被限制在区间 $[-B, B]$ 内，其中 $B > 0$。因此，显然DN假设查询输出通过输出扰动 进行保护，且误差严格有界。

根据 DN 的定义，数据库重构是逐条记录地重构原始值，使得重构值与原始值之间的距离在特定精度界限内。因此，DN 的攻击旨在以高精度推断原始数据库中每条记录的值。他们根据攻击者的计算能力，考虑了两种不同的攻击者：

- 指数时间攻击者。这种攻击者能够发出所有可能的查询。在实践中，这种对手只有在小型数据库的情况下才切实可行，因为例如，对于 $n \geq 100$，即使使用现有最快的计算机，发出所有可能的查询也需要数年甚至数十年。为防范这种攻击者，任何查询的输出都会通过在 $[-B, B]$ 中添加随机噪声来修改。如果针对目标原始数据库获得的查询响应与针对某个特定候选数据库获得的相应查询响应之间的差异在 B 以内，那么该候选数据库就构成原始数据库的一种重构。DN 证明，在这种情况下，候选数据库与目标原始数据库之间的距离在 $4B$ 以内，其中两个数据库都视为二进制 n 向量。因此，除非值 B 相对较大，否则候选数据库就是原始数据库的良好重构。DN 证明，为了阻止这种良好重构，B 相对于 n 必须不可忽略，也就是说，B 必须为 $O(n)$。

- 多项式时间攻击者。该攻击者发起的查询次数关于 n呈多项式增长，对于大型数据库而言是可行的。数据库通过在 $[-B, B]$中加入随机噪声来保护查询输出。利用这些受保护的查询输出，攻击者求解一个线性规划问题，以重构数据库。DN 表明，只要 B 处于 $o(\sqrt{n})$之内，其中"小 o"记号表示随着 n增长，其值远小于 $\sqrt{n}$，那么以高概率，重构数据就会接近原始数据库。因此，为实现保护，对于不可忽略比例的查询，B 必须为 $O(\sqrt{n})$（使用"大 O"记号）。

因此，DN 得出结论：除非添加到查询输出中的噪声与数据库的规模相当（对于指数级对手为$O(n)$，对于多项式级对手为 $O(\sqrt{n})$），否则攻击者就能够重建数据库。至少为 $O(\sqrt{n})$的噪声水平现实吗？考虑一个规模为 n =1,000,000的数据库。要求的是，查询中有不可忽略的比例，其答案与相应的真实答案相差约 1,000。请注意，这种噪声不必应用于所有查询。此外，与数据库的规模以及可能涉及数十万条记录的查询相比，约 1,000 的扰动相对较小。因此，在许多情况下，为防御多项式级对手而要求的噪声水平似乎是可以承受的。

毫无疑问，DN 关于仅利用对查询的响应进行数据库重构，提供了非常相关的见解。作者提出了一个理论框架，说明重构风险如何随对手的计算能力以及施加于查询输出的噪声而变化。然而，提出数据库重构的理论框架，并不意味着每个数据库都能被重构。

首先，*DN* 的研究结果仅适用于输出扰动，而不适用于本地保护或输入保护。DN 提及“CD模型”时，明确承认了这一点:

> **CD模型**。 上面的数据库算法本质上创建了数据库 d' 的一个“私有”版本，然后使用 d' 回答查询。请注意，用户可以通过查询 $q_i = \{i\}$ 来获取 d' 的全部内容，查询条件为 $1 \leq i \leq n$; 此后，她便可以自行回答所有其他查询。这一结果表明，在某些情况下，可以在 *CD*模型 中实现隐私保护：用户获得数据库的一个“私有”版本（刻录在 CD 上），并可对其进行操作（例如，不受只能进行统计查询的限制）。

具体而言，如果采用本地掩蔽或输入掩蔽，所有查询的响应都基于掩蔽后的数据库。因此，对于本地扰动或输入扰动，*DN*框架只能重构掩蔽后的数据库。现在，如果对本地掩蔽或输入掩蔽进行适当配置，以充分保护原始数据库（例如，在受访者层面使用随机响应，或采用 Hundepool 等人（2012年）所述的微观数据统计披露控制方法），那么重构掩蔽后的数据库就不应导致敏感信息泄露。

3 差分隐私抵御重构的性能

在 Dwork（2011年）和 Garfinkel 等人（2019年）看来，解决输出保护数据重构漏洞的所谓方案是差分隐私（DP）。Dwork 等人于2006年提出了差分隐私，将其作为一种框架，用于量化基于机密数据库回答查询所带来的披露风险。假设对手向数据库提交查询并获得查询响应 R。 ϵ-差分隐私要求：对于两个仅在一条记录上有所不同的数据库 D 和 D'，以及查询响应空间中的所有子集 S

$$\Pr(R \in S|D \text{ is used}) \leq e^{\epsilon} \times \Pr(R \in S|D' \text{ is used}). \tag{1}$$

从本质上说，差分隐私要求，仅通过观察 R，就无法以超过因子 e^{ϵ} 的程度区分使用的是数据库 D 还是数据库 D'。当 $\epsilon = 0$时，这一要求意味着，观察 R时，所使用的数据库必须完全无法区分。在这种情况下，即使向对手返回了 R，两个数据库 D 和 D' 之间不同的记录值仍完全保密。值 ϵ通常称为“隐私预算”; 要使表达式（1）的隐私条件具有实际意义，该值应当较小：Dwork（2011年）建议 ϵ “例如取 0.01、0.1，或者在某些情况下取 ln 2 或 ln 3”。

差分隐私的一个众所周知的性质是顺序组合：如果分别以隐私级别 ϵ_1、 ϵ_2、.....、 ϵ_k回答 k 个查询，那么回答完所有 k 个查询后的现存隐私级别为 $\epsilon_1 + \epsilon_2 + \cdots + \epsilon_k$。

一种称为 (ϵ, δ)-差分隐私的差分隐私放宽形式也已被提出，其定义如下:

$$\Pr(R \in S|D \text{ is used}) \leq e^{\epsilon} \times \Pr(R \in S|D' \text{ is used}) + \delta, \tag{2}$$

其中 δ 是放宽参数。 δ 的值通常被解释为意味着以概率 $1 - \delta$满足 ϵ-差分隐私。但只需更仔细地比较表达式 (1) 和 (2)，就能发现，对于任何 查询，即使不满足 ϵ-差分隐私，(ϵ, δ)-差分隐私也可能成立。

实现差分隐私的常见做法是返回查询答案 R，其中包含基于原始数据计算出的查询结果，以及拉普拉斯分布噪声。 ϵ 的值越小，查询越敏感（即 单条记录的变化可能导致查询结果发生的变化越大），所需的噪声就越多。

Dwork（2011年）似乎认为，随机响应（Warner 1965）[2]所提供的 $O(\sqrt{n})$ 准确度，可以被差分隐私程序超越。她写道:

> 假设有 n 位受访者各自独立使用随机响应，但所用硬币的偏置已知且固定。那么，给定随机化后的数据，根据二项分布的性质，分析者可以近似得出问题“有多少位受访者的值为 b? ”的真实答案，预期误差的量级为 $O(\sqrt{n})$。正如我们将看到的，误差还可以小得多——达到恒定的预期误差，且不依赖于 n。

[2]尽管随机响应能够满足差分隐私的要求（Dwork 2011年），但随机响应于1965年提出，比统计披露控制学科诞生早十多年，比差分隐私早四十年。事实上，随机响应除了满足差分隐私之外还有其他特性，例如可以根据随机化后的分布估计原始分布。

然而，要实现独立于 n 的恒定误差，就与 Dinur 和 Nissim（2003年）的要求相冲突。根据该要求，为防止（多项式时间的）对手根据查询输出重建数据库，在不可忽略比例的查询中，噪声至少需要达到 $O(\sqrt{n})$。正如 Dwork 上文所指出的，随机响应恰恰会产生 $O(\sqrt{n})$ 级别的噪声。

因此，如果某个差分隐私程序提供的误差是常数且独立于 n，那么根据 DN，它无法防止重构。事实上，如果允许 DN 对手提交多项式数量的查询，例如 $m = O(n)$，则适用顺序组合，因为一般来说，查询可能针对彼此重叠的个人集合。因此，总隐私预算 ϵ 必须拆分为若干份，每次查询使用 ϵ/m 。因此，由于施加于每个查询答案的噪声与其隐私预算成反比，对于拉普拉斯机制，噪声的标准差与 m/ϵ 成正比，因此为 $O(n)$。总而言之，如果正确应用ϵ-差分隐私，它就能通过使用 $O(n)$ 噪声来防止重构，其噪声远多于随机响应。因此，随着输出复杂度不断提高，ϵ-差分隐私机制很可能会过度保护输出，正如 Bach（2022年）所指出的那样。

4 地理区域与政策决策对重构的相关性

美国法律[3] 要求人口普查局不得“发布任何能够识别出由任何特定机构或个人提供的数据 [...] 的内容”。在发布表格所采用的最低地理层级上具有独特特征的任何个人，都面临被重新识别的风险；也就是说，任何计数为 1 的单元格都可能导致重新识别。

在人口普查背景下实施差分隐私的关键原因之一，是有人声称，用于保护此前十年一次的人口普查的数据交换方法无法抵御重构和重新识别。Garfinkel 等人（2019年）通过一个简单的假设示例，运用主要抑制和次要抑制来凸显重构的危险。研究表明（Muralidhar 和 Domingo-Ferrer 2022年，Muralidhar 和 Domingo-Ferrer 2021年），如果以正确的方式实施主要抑制和次要抑制，即使这种重构也无法实现（例如，如人口普查方法文件（Dupre 2020年）及相关统计披露控制文献（Antal 等人2017年，联合国欧洲经济委员会 2015年）所述）。然而，在精细的地理层级上发布统计数据仍可能助长重构。下文将对此问题进行探讨。

4.1 地理区域的影响

如果统计数据按小地理区域发布，就可以通过简单的算术运算进行重构。事实上，无论是采用数据交换还是差分隐私作为统计披露控制方法来保护表格，只要小地理层级的总计数得到精确保留，就可以进行重构（见 Abowd 和 Hawes （ 2022 ），第 8 页）。重构的难易程度很大程度上取决于报告精确计数所针对的地理区域有多小。具体而言，在美国人口普查局对 2010年人口普查数据进行的数据交换与差分隐私比较中：

- 在对 2010年人口普查实施数据交换时，区块级别的总人口和投票年龄人口数保持不变（即精确报告）（ Abowd（2021），第 12 页）。
- 相比之下，在对 2010年人口普查实施差分隐私时，只有州级人口保持不变。请注意，2017年时曾承诺提供区块级别的精确计数：“根据与司法部达成的协议，美国人口普查局将提供人口普查区级别的精确计数[...]”（Dajani等人 2017年）。

如今，区块超过 600 万个，而州只有 50 个，另加哥伦比亚特区。因此，与差分隐私相比，数据交换受到的限制要多得多，因而也更容易泄露信息。取消数据交换的区块层级限制（保留总人口和投票年龄人口计数），或许能使数据交换所提供的隐私保护与差分隐私处于同等水平。要公平比较数据交换与差分隐私，就需要人口普查局报告针对置换后的数据和

[3]https://www.c ensus.gov/about/adrm/链接/about/authority .html

2010年十年一次的人口普查的差分隐私保护数据所实施的重构攻击结果，并确保在相同地理层级保留精确计数。这样就能比较两种方法提供的隐私保护和效用；尤其值得探究的是，能否按照对数据交换的描述，以同样的方式对差分隐私保护数据进行重构（Garfinkel等 （2019年），第 34 页）。

尽管人口普查局声称已对差分隐私与数据交换和数据抑制进行了比较分析，但并未公布具体的比较结果。仅提供了以下陈述：“为了达到必要的隐私保护水平，增强型数据交换和数据抑制都会对数据质量和可用性造成极其严重的负面影响”（Abowd（2021），第25页）。

另一个值得关注的问题是，在差分隐私下，即使将要求从精确计数放宽为在多个地理层级上保持计数一致，也存在困难。根据 Garfinkel（2019）（第59页），只要保持一致性，就可以在2020年人口普查的所有地理层级上添加噪声。为确保这种一致性，人口普查局基于差分隐私开发的 DAS方法涉及多个后处理步骤（Kenny等人2021）。

4.2 隐私预算对差分隐私的影响

如果主张以差分隐私取代传统统计披露控制方法，就应明确规定隐私预算 ϵ，正如 Dwork 等人（2019）所要求的那样。取值非常小的 ϵ 会带来难以承受的效用损失，而取值 ϵ非常大则几乎不添加噪声，对数据库重构几乎没有保护作用，更不用说防止重新识别了（Domingo-Ferrer 等人2021）。

事实上，近期美国人口普查文件提到，2021年的 ϵ 值高达19.61（U.S. Census 2021），2022年的值高达39.907（U.S. Census 2022）。我们以 2021 ϵ 值为例，说明它能提供的隐私保护有多么有限（2022年的值所能提供的保护更少）。我们从两个不同的角度来分析，得出相似的结论：

- 首先，我们利用差分隐私与随机响应之间的联系（Wang 等人 2016）。考虑二元属性的随机响应，即报告的随机化答案以 $p \geq 0.5$ 的概率与真实答案相同，以 $1-p$的概率不同。那么，对于任意 ϵ， ϵ-差分隐私所导致的披露风险，与 $p = \exp(\epsilon)/(1+\exp(\epsilon))$时随机响应所导致的披露风险相同。具体而言，$\epsilon = 19.61$ 相当于参数为 $p = 0.99999999696$的二元随机应答，也就是说，以实际上接近1的概率报告原始值，这基本上等于没有提供任何披露保护。
- 或者，我们采用 Dinur 和 Nissim 的视角。为便于说明，假设添加的噪声取自拉普拉斯分布。对于 $\epsilon = 19.61$，噪声以高于 0.999999997 的概率处于 $[-1,1]$范围内，以高于 0.999944825 的概率处于 $[-0.5, 0.5]$ 范围内。添加如此水平的噪声将违反 DN 要求，即噪声应为 $O(\sqrt{n})$，并会使数据能够被准确重构。

更糟的是，即使 ϵ 的值很高，在某些情况下，发布的差分隐私数据效用仍可能很低，Van Riper等（2020年）、Ruggles 和 Van Riper（2021年）以及 Kenny等（2021年）都指出了这一点。这是因为：(a) 顺序组合要求在所有彼此不独立的发布输出之间分配隐私预算（例如，分配给表格中的所有单元格、不同地理层级之间、彼此相关的查询之间等）；以及 (b) 需要通过人口普查的自顶向下算法（TDA）进行后处理，以发布一致、整数且非负的数据。

所提出的 ϵ 高值所获得的实际保护和效用损失，应与在相同不变性约束下使用传统统计披露控制方法（例如，2010年十年一次的人口普查所采用的方法）能够达到的保护和效用损失进行比较。

4.3 透明度

使用差分隐私时的一个关键主张是：“这反过来使人口普查局这样的机构能够量化保护隐私所需的精确统计噪声量。这种精确性使人口普查局能够校准并分配精确数量的统计噪声，从而在保护隐私的同时维持数据整体的统计有效性”（Abowd（2021），第22页）。

事实上，任何保护方法都是如此，包括数据交换。可以选择置换参数来：(1) 决定纳入数据交换的记录数量（多或少），(2) 决定要交换的属性，以及 (3) 决定是否对每个属性分别独立进行数据交换。

像人口普查局所做的那样，公开 DP 中使用的 ϵ 参数，无疑是朝正确方向迈出的一步。然而，仅此一项并不能使保护方法变得透明。所采用的后处理对用户而言仍然是不透明的。直到 2010年人口普查之前，数据交换方法受到的一项主要批评是，置换参数（置换记录的比例）没有向公众公开。但人口普查局确实公布了置换记录比例的上限。鉴于数据交换方法简单，这使整个流程相当透明。此外，数据交换还确保某些计数即使在区块级别也能得到保留，从而进一步提高了透明度。

在我们看来，透明度不仅关乎参数发布，也与方法的复杂程度有关。方法越复杂，对用户而言就越不透明。从这个角度看，当前基于差分隐私的方法可以被认为比简单的数据交换更不透明。

5 重构与重新识别是不同的

重构和重新识别是两个不同的概念:

- 重构只是披露过程的第一步。请注意，根据统计查询的输出或发布的制表数据进行重构，得到的重构数据不包含任何标识符。
- 重新识别是完成披露攻击的第二个必要步骤。在这一步中，重构数据会与某个特定个体关联 起来。为此，攻击者需要一个包含标识符以及可用于与重构数据关联的某些属性的外部数据源。在最坏的情况下（对攻击者最有利的情况），外部数据源可能包含带有身份识别信息的完整 原始数据。例如，如果攻击是由负责保护数据的同一组织发起的（以测试重构质量），那么这种最坏情况就说得通。

Abowd（2021）（附录B）和 Garfinkel（2019）对重构和重新识别的描述如下:

- 308,745,538人的微观数据记录被“重构”了。
- 研究使用了四个关于2010年美国人口的外部商业数据库，其中记录了人们的姓名、地址、年龄 和性别 。
- 重构后的记录与商业数据库关联，得到一个包含姓名、地址、年龄、性别、族裔与种族的关联数据库。45%的记录可以关联起来。
- 研究将关联数据库与美国人口普查局机密数据进行了比较。据称，这次攻击获取了38%的关联记录中的所有属性，换言之，相当于获取了美国人口17%的相关信息。
- 因此，作者声称重新识别了美国人口的17%，尽管他承认，外部攻击者并不知道哪些重新识别是正确的。

之后，作者进一步批评2000年和2010年人口普查所采用的保护系统存在缺陷，该系统依赖于传统的统计披露控制技术。这被用作转向形式化隐私（即差分隐私）的理由。Garfinkel（2019）解释说，选择隐私预算 ϵ 是一项公共政策选择。

5.1 关于重新识别声明的问题

Abowd（2021）（附录B）中的声明存在几个问题:

1. 对 308,745,538 人的微观数据进行“重构”究竟意味着什么，并不清楚。根据 Van Riper 等人 （2020年）的说法，这指的是根据已发布的人口普查区和普查小区汇总表重新生成微观数据记录，也就是说，依据包含人口普查区、年龄和性别属性的频数表）。这并非 DN 所描述的真正重构。需要注意的是，通常无法从频数表唯一地重新生成微观数据，因为频数表所包含的信息少于据以计算它的微观数据

。因此，仅仅重新生成一个与某个频数表相兼容的微观数据集，并不构成对原始数据的重构：在 DN 的重构概念中，精度界限至关重要，而人口普查所谓的重构并未给出此类界限（Muralidhar 2022年）。

2. 重新识别是指将匿名微观数据中的记录与外部数据集中相应的记录进行关联；该外部数据集包含标识符，且覆盖相近的人群。这并非重构。由于此次攻击是基于由频数制表重新生成的微观数据，因此只有计数为 1 的单元格才能进行适当的重新识别。在其他所有情况下，都不可能明确无误地完成重新识别。

3. 至少从 Sweeney（2000）起，人们就已知道，将包含人口统计属性（例如居住市镇_的_所在地、出生日期 和性别 ）的数据库，与包含同一人群的这些相同属性及标识符的外部数据库进行匹配，很可能会得到很高比例的重新识别结果。事实上，Ruggles 和 Van Riper（2021）通过模拟表明，人口普查局实验报告的区块级别匹配中，大多数都可能是随机产生的，因此无法证明其对保密性构成可信威胁[4]。因此，可能没有必要使用差分隐私[5]。

上述问题清楚地表明，与其关注重新识别，不如说人口普查实验关注的是寻找（非唯一的）候选重构。接下来我们将说明，在某些情况下，将重构（错误地）解释为重新识别可能会高估真实的重新识别风险，而在另一些情况下则可能低估这一风险。

5.2 将重构误解为重新识别可能会高估或低估重新识别风险

回顾一下，在 Abowd（2021）和 Garfinkel（2019）所述、并在第 5 节开头概述的人口普查局“重新识别”程序中，报告性别、年龄、种族和族裔 的重构微观数据，会与报告姓名、地址、年龄 和性别的外部商业数据库进行链接。因此，链接使用年龄 和性别 属性。链接完成后，将得到一个关联数据库，其中报告姓名、地址、年龄、性别、种族 和族裔。

考虑区块级别的三种情形：

- 场景*1*。 该区块的重构数据包含10名个人，其年龄=44、性别 = 为男性，种族 = 为白人，且族裔 = 为非_西班牙裔。商业数据库中，该区块所有个人的姓名、地址、年龄=44和性别=均为男性。

- 场景*2*。 数据与场景1相同，但增加了一个关系属性。根据 Garfinkel（2019）的说法，该属性也会针对区块中的每个人收集，并且有17种不同取值。假设在此场景中，区块内10个人的关系 取值各不相同。与场景1一样，商业数据库中该区块所有个人的姓名、地址、年龄=44 和性别=均为男性。

- 场景*3*。 一个区块，其重构数据由10名个人组成，这些人的年龄=44、性别 = 均为男性，并且这10名个人分别属于不同的（种族、族裔）组合。商业数据库包含该区块中所有个人的姓名、地址、种族 和族裔 ，但不包含年龄 或性别。

[4]Ruggles（2021年）还提出了另一种研究重构对重新识别影响的方法。其思路是先将外部数据库与重构后的人口普查数据进行匹配，由此得到一定的匹配率 r。然后，取出外部数据库中未匹配的记录，并根据区块 ID 和 PIK（人口普查局为每条原始记录创建的受保护识别密钥），将这些记录与人口普查编辑文件（原始机密数据）进行比较。令 r' 为比较得出的重新识别率。如果 $r \approx r'$，那么数据库重构对重新识别几乎没有影响或毫无影响；要证明重构会增加重新识别风险，r 就应显著大于 r'。人口普查局尚未进行这项比较。[5]事实上，即使保密性受到威胁的说法可信，目前人口普查局基于差分隐私的自顶向下算法是否能提供最佳保护也尚不明确。Francis（2022年）表明，与人口普查局先前的保护算法的输出相比，从自顶向下算法的输出中可以用更少的先验知识、更准确地推断种族和族裔。

在场景1中，人口普查局的程序会得出100%的重构，因为攻击者总是能够将正确的种族 和族裔 与区块中的10个姓名和地址关联起来。然而，声称这种100%的重构相当于100%的重新识别，显然是错误的，因为攻击者无法确认这10名彼此无法区分的个人的（姓名、地址）身份；在这种情况下，“重新识别”可归因于区块内的同质性，并不是真正的重新识别。美国人口普查局的文件 McKenna 和 Haubach（2019年）指出：“有必要通过将微观数据中被隐去的身份与外部数据集中的身份进行比较，来验证拟议的匹配结果，以确定这些匹配是真匹配还是误匹配。”

场景2清楚地表明了上述观点：正确重构族裔 和种族 并不等同于重新识别。当为所有10位个人添加具有不同取值的关系 属性后，便可清楚看出，针对任意特定个人的重新识别概率实际上为 1/10。

在场景3中，正确重构的概率和正确重新识别的概率均为1，但原因各不相同：

- 由于所有个人的（年龄，性别）组合都相同，因此重构这10位个人的这些属性值轻而易举，重构率达到100%。请注意，如果并非所有个人的组合都相同，那么正确重构的概率就会低于1。

- 由于所有个人的（种族，族裔）组合都不同，因此很容易将重构数据中的10条记录分别明确地链接到商业数据库中对应的记录，从而实现100%的重新识别。

以上表明，重构和重新识别是不同的概念。归根结底：缺失的机密属性越同质，越有利于重构；而用于执行链接的准标识符越异质，越有利于重新识别。

因此，Abowd（2021）和 Garfinkel（2019）所述的重构程序并不能恰当地衡量重新识别风险。事实上，它很可能会高估重新识别风险（如场景1所示），因为在区块级别，种族和族裔预计会相当同质，因此场景1比场景3更有可能发生（Ruggles 和 Van Riper（2021））。

有趣的是，美国人口普查局的研究人员过去曾对重新识别风险开展过广泛研究（例如，Winkler（1999））。要评估真实的重新识别风险，必须作如下假设。在区块级别，重构数据包含（区块_标识符、性别、年龄、种族、族裔、关系），而攻击者拥有以下属性（姓名、地址、区块_标识符、性别、年龄、种族、族裔、关系）。重新识别的目标是，将一条来自（未识别的）重构数据的记录唯一地链接到攻击者（已识别）数据中的一条记录，从而为重构数据添加（姓名、地址）。这样的程序能够正确评估上述场景中的重新识别风险；如上所述，也如 McKenna 和 Haubach（2019）所述，建立链接后，必须通过检查链接是否唯一，以及攻击者记录中的身份信息（此处为姓名和地址）是否与未识别的重构记录所对应的原始记录相符，来验证重新识别。单独重构未识别的记录并不会构成真正的披露威胁。按照 DN 的含义，“重构”还必须辅以正确的重新识别。只有这样才会构成真正的披露。

6 结论与未来工作

本文重新评估了根据数据集上计算的统计查询输出重构该数据集的可行性。重构的风险一直被用作在官方统计中采用差分隐私的理由，尤其是在美国2020年人口普查中。然而，采用差分隐私很可能会降低该次人口普查统计输出的效用。本文探讨了重构在多大程度上构成真正的风险。

我们首先考察了重构理论的现状——Dinur和Nissim的框架——并得出结论：局部保护或输入保护似乎是抵御重构的有效方式。如果美国人口普查局坚持采用所谓的CD模型，并发布经过局部保护或输入保护的数据（例如，使用Hundepool等人（2012年）讨论的随机响应或微观数据屏蔽，或采用2010年

年人口普查中使用的方法），那么重构就不会构成真正的威胁：攻击者最多只能重构经过局部保护或输入保护的数据，而非原始数据。差分隐私也是一种选择，但它可能会加入超出抵御重构所严格必需的更多噪声，从而造成不必要的效用损失（Dove 2021、Hotz等人2022、Bach 2022），也可能比以往的方法提供更少的保护（Francis 2022）。

接着，我们强调了保护保留精确人口计数的地理层级的重要性。无论采用何种统计披露控制方法，在小地理区域保留人口计数都会助长重构；而不在小地理区域保留人口计数，则能在很大程度上避免重构，但也会降低效用。

最后，我们警告，不应将重构程度作为重新识别风险的衡量标准，因为这会夸大重新识别风险。重构只需要查询输出或表格化未识别输出，且缺失属性值的同质性有利于重构；而重新识别还需要外部已识别数据源，并且用于链接这些数据源的属性值越异质，就越有利于重新识别。

另一个令人担忧的问题是，在此过程中，ϵ 的值连续增加。增加 ϵ 意味着隐私损失。人口普查局最初采用拉普拉斯噪声添加，设置为 $\epsilon = 4.5$，随后提高到 $\epsilon = 10.2$。为了进一步提升发布数据的效用，人口普查局采用零集中差分隐私（噪声来自离散高斯分布，Bun 和 Steinke，2016）取代了拉普拉斯噪声。零集中差分隐私的参数 ρ 可用于计算（ϵ，δ）-差分隐私的等效值 ϵ 和 δ。对于给定的 ρ，存在许多等效组合（ϵ，δ）。然而，当 δ 固定时（在人口普查中，该值为 $\delta = 10^{-10}$），每个 ρ 都只有一个等效的 ϵ。2021年，这一等效的全局 ϵ 为19.61（美国人口普查局，2021）。2022年，这一数值又被修订为全局 $\epsilon = 39.907$（美国人口普查局，2022）。这对隐私产生了重大影响。与 $\epsilon = 4.5$ 所提供的隐私水平相比，$\epsilon = 39.907$ 对应的隐私水平差了 $e^{39.907}/e^{4.5} = 2.382 \times 10^{15}$ 倍。谈及苹果公司使用 $\epsilon = 14$一事，差分隐私的发明者之一 Frank McSherry 评论称，就隐私而言，这种做法“毫无意义”（Greenberg 2017）。实际上，$\epsilon = 39.907$ 比 $\epsilon = 14$差了超过 1.78×10^{11} 倍。在 Abowd（2021b）中，时任人口普查局首席科学家谈到 ϵ 时表示：“具体而言，它限制了所有可能的检验在统计功效上的表现，这些检验用于判断某个特定个体的数据记录（或其中一部分）是否被用于生成某一统计数据集合，而不是使用另一名任意个体的记录。”采用2022年的 $\epsilon = 39.907$，这意味着某个特定个体的数据记录被使用的概率，可能比另一名任意个体的记录高出 2.14×10^{17} 倍以上。鉴于当前美国人口仅为 3.31×10^8，在 $\epsilon = 39.907$ 的情况下，任何目标美国居民都可能被重新识别。事实上，这一点对于2021年采用的 $\epsilon = 19.61$ 也成立。

即使放宽了对 ϵ 值的要求，数据效用方面仍然存在非常严重的问题。请看《纽约时报》的一篇报道（Wines 2022）：“根据2020年人口普查，那里有14人居住[在芝加哥市中心的人口普查区块1002] ——13名成年人和1名儿童。根据2020年人口普查，他们还住在水下。因为该区块完全由芝加哥河一段700英尺长的弯道组成。”再看康奈尔大学（Cornell 2021）对纽约州2021年人口普查DAS发布数据的分析：6.1%的区块中，家庭人口大于0，但有人居住的房屋数为0；2.5%的区块中，家庭人口少于有人居住的房屋数（这意味着每户不到1人）；还有0.8%的区块中，家庭人口为0，但有人居住的房屋数大于0。这些结果不可能出现，在2010年人口普查中也不会出现。因此，即使 ϵ 很大，差分隐私噪声添加程序也无法提供准确且一致的结果。事实上，美国人口普查局最近宣布，“目前，美国社区调查公共使用微观数据样本 [American Community Survey Public Use Microdata Sample] 数据产品仍将采用传统的披露规避方法加以保护”，因为“目前尚不清楚差分隐私最终是否会是最佳选择”。（Daily 2022）

在本研究中，我们得出结论：认为数据库重构会导致大规模披露，这种担忧并无根据。我们认为，这些说法基于不完整且不透明的比较——只有人口普查局能够评估或核实真正的重新识别结果。其他研究人员（其中一些已在上文提及）对输出的准确性和一致性提出了严重质疑。因此，目前尚不清楚差分隐私是否是处理2020年十年一次人口普查数据的最佳选择。我们建议对替代方法的有效性开展全面、独立、完整记录并经过同行评审的评估。

参考文献

J. M. Abowd. 2017. 研究数据中心、可重复科学与机密性保护：21世纪统计机构的作用。向 *SummerDemSem*、威斯康星联邦统计研究数据中心作报告。6月5日。 https://www2.census.gov/cac/sac/meetings/2017-09/ role-statistical-agency.pdf J. M. Abowd、I. M. Schmutte、W. N. Sexton 和 L. Vilhuber. 2019. 为什么经济学界不能将隐私保护的讨论拱手让给计算机科学家。向信息披露风险与成本上升背景下的经济研究未来，联合社会科学协会年会作报告。1月5日。J. M. Abowd. 2021. John M. Abowd 的声明。案件编号 3:21-CV-211-RAH-ECM-KCN，美国阿拉巴马州中区联邦地区法院。4月13日。J. M. Abowd. 2021. John M. Abowd 的补充声明。案件编号 3:21-CV-211-RAH-ECM-KCN，美国阿拉巴马州中区联邦地区法院。4月26日。J. M. Abowd 和 M. B. Hawes. 2022. 2020年美国人口普查人口与住房数据的机密性保护。https://arxiv.org/pdf/2206.03524.pdf N. R. Adam 和 J. C. Worthmann. 1989. 统计数据库的安全控制方法：比较研究。《*ACM*计算调查》，21(4):515-556. 阿拉巴马州诉美国商务部案。2021. 布伦南司法中心。访问日期：2023年1月25日。https: // www.brennancenter.org/our-work/court-cases/alabama-v-us-dept-commerce L. Antal、T. Enderle 和 S. Giessing. 2017. ESS人口普查数据协调保护：人口普查数据协调保护的统计披露控制方法。欧盟统计局统计披露控制卓越中心，海牙。美联社。2021. 16个州支持阿拉巴马州对人口普查隐私工具提出的质疑。《美国新闻》。4月13日。访问日期：2023年1月25日。 https://www.usnews.com/news/us/articles/2021-04-13/ 16-states-back-alabamas-challenge-to-census-privacy-tool F. Bach. 2022. 欧洲人口统计中的差分隐私与噪声化机密性概念。《调查统计与方法学杂志》，10:642-687. M. Bun 和 T. Steinke. 2016. 集中差分隐私：简化、扩展与下界。载于：密码学理论会议-*TCC 2016*，施普林格，第635-658页。*2020*年人口普查结果：美国人口普查局逐步发布数据产品后，纽约州的数据与分析。康奈尔大学应用人口学项目，9月2日。 https://pad. human.cornell.edu/census2020/index.cfm#das T. Dalenius. 1977. 迈向统计披露控制方法论。《统计学杂志》，15:429–444. D. Daily. 2022. 美国社区调查的披露规避保护措施。美国人口普查局，12月14日。https://www.census.gov/newsroom/blogs/random-samplings/2022/12/ disclosure-avoidance-protections-acs.html A. N. Dajani、A. D. Lauger、P. E. Singer、D. Kifer、J. P. Reiter、A. Machanavajjhala、S. L. Garfinkel、S. A. Dahl、M. Graham、V. Karwa、H. Kim、P. Leclerc、I. M. Schmutte、W. N. Sexton、L. Vilhuber 和 J. M. Abowd. 2017. 美国人口普查局统计披露限制的现代化。载于：人口普查科学咨询委员会会议，9月14日至15日。https://www.census. gov/library/video/2017/2017-09-sac.html

D. E. Denning 和 J. Schlörer。1980年。快速查找统计数据库中追踪器的程序。*ACM*《数据库系统汇刊》，5(1):88-102。I. Dinur 和 K. Nissim。2003年。在保护隐私的同时揭示信息。载于：第 *22* 届 *ACM SIGMOD-SIGACT-SIGART* 数据库系统原理研讨会论文集，第 202-210 页。J. Domingo-Ferrer、D. Sánchez 和 A. Blanco-Justicia。2021年。差分隐私的局限（及其在数据发布和机器学习中的误用）。《美国计算机协会通讯》，64(7):33-35。I. Dove。2021年。将差分隐私保护应用于 *ONS* 死亡率数据：试点研究。英国国家统计局。G. T. Duncan、S. A. Keller-McNulty 和 S. Lynne。2001年。披露风险与数据效用：*RU* 保密性地图。国家统计科学研究所，技术报告第 121 号。https: // www.niss.org/sites/default/files/technicalreports/tr121.pdf S. Dupre。2020年。披露规避与人口普查。《国际人口普查专题选编》，美国人口普查局。10月。https://www.census.gov/content/dam/Census/library/working-papers/ 2020/demo/disclosure_avoidance_and_the_census_brief.pdf C. Dwork。2011年。隐私数据分析的坚实基础。《美国计算机协会通讯》，54(1):86-95。C. Dwork、N. Kohli 和 D. Mulligan。2019年。实践中的差分隐私：公开你的 epsilon！《隐私与保密期刊》，9(2)。https://journalprivacyconfidentiality.org/index.php/ jpc/article/view/689 C. Dwork、F. McSherry、K. Nissim 和 A. Smith。2006年。在隐私数据分析中根据敏感度校准噪声。载于：密码学理论会议 – *TCC 2006*，施普林格，第 265-284 页。C. Dwork 和 A. Roth。2014年。《差分隐私的算法基础》。Now 出版社。P. Francis。2022年。关于误解美国人口普查再识别攻击的一则短评。载于：统计数据库隐私 – *PSD 2022*，施普林格，即将出版。预印本见：https://arxiv.org/ pdf/2202.04872.pdf S. Garfinkel。2019年。为2020年人口与住房普查部署差分隐私。在隐私增强技术研讨会–*PETS 2019*上的主题演讲。访问日期：2023年1月25日。https://simson.net/page/Main_Page S. Garfinkel、J. M. Abowd 和 C. Martindale。2019年。理解针对公开数据的数据库重构攻击。《美国计算机协会通讯》，62(3):46-53。《通用数据保护条例》。《欧盟条例 2016/679》。https://gdpr-info.eu A. Greenberg。2017年。苹果公司的一项关键隐私保护措施为何未能奏效。《连线》，9月15日。https://www. wired.com/story/apple-differential-privacy-shortcomings/ V. J. Hotz、C. R. Bollinger、T. Komarova、C. F. Manski、R. A. Moffitt、D. Nekipelov、A. Sojourner 和 B. D. Spencer。2022年。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。A. Hundepool、J. Domingo-Ferrer、L. Franconi、K. Spicer、P.-P. De Wolf、S. Giessing 和 E. Schulte Nordholt。2012年。《统计披露控制》。威立。C. T. Kenny、S. Kuriwaki、C. McCartan、E. T. R. Rosenman、T. Simko 和 K. Imai。2021年。差分隐私在人口普查数据中的应用及其对选区重划的影响：以2020年美国人口普查为例。《科学进展》，7(41)。https://www.science.org/doi/10.1126/sciadv.abk3283 L. McKenna 和 M. Haubach。2019年。美国人口普查局的传统技术与披露规避领域的当前研究。研究与方法局，美国人口普查局。4月。

K. Muralidhar。2022年。重新审视人口普查局的重构与重新识别攻击。载于：统计数据库隐私 – *PSD 2022*，施普林格，即将出版。预印本见：https: //arxiv.org/pdf/2205.03939.pdf

K. Muralidhar 和 J. Domingo-Ferrer。2021年。数据库重构在实践中非常困难！载于：*2021*年联合国欧洲经济委员会/欧盟统计局统计数据保密专家会议，波兹南，波兰，2021年12月1日至3日。

K. Muralidhar 和 J. Domingo-Ferrer。2022年。人口普查再思考（致编辑的信）。《美国计算机协会通讯》，65(6):11。

K. Muralidhar 和 R. Sarathy。2009年。州教育机构向公众发布的问责数据中的隐私侵犯。载于：联邦统计方法研究委员会研究会议，美国华盛顿特区。

K. Percival。2021年。法院驳回阿拉巴马州对人口普查选区重划及隐私保护方案的质疑。布伦南司法中心。访问于2023年1月25日。https://www.brennancenter.org/our-work/analysis-opinion/court-rejects-alabama-challenge-census-plans-redistricting-and-privacy S. Ruggles。2021年。个人通信。11月9日。S. Ruggles 和 D. Van Riper。2021年。随机性在人口普查局数据库重建实验中的作用。《人口研究与政策评论》，自8月22日起提前发布。https: //doi.org/10.1007/s11113-021-09674-3 M. Schneider。2022年。研究人员要求人口普查局停止有争议的隐私保护方法。美联社新闻，8月8日。https://apnews.com/article/census-2020-us-bureau-government-and-politics-20e683c71eeb62ee4b7792d7d8530419 L. Sweeney。2000年。简单的人口统计特征往往足以唯一识别个人。卡内基梅隆大学，《数据隐私工作论文》第3号。访问于2023年1月25日。http://ggs685.pbworks.com/w/file/fetch/ 94376315/Latanya.pdf J. F. Traub、Y. Yemini 和 H. Woźniaknowski。1984年。统计数据库的统计安全性。*ACM*《数据库系统汇刊》，9(4):672-679。联合国欧洲经济委员会—欧洲统计学家会议—UNECE-CES。2015年。2020年人口和住房普查建议，联合国出版物，纽约州纽约市。美国人口普查局。2021年。人口普查局设定关键参数，以保护2020年人口普查结果的隐私。新闻稿编号 CB21-CN.42，6月9日。https://www.census.gov/newsroom/press-releases/ 2021/2020-census-key-parameters.html 美国人口普查局。2022年。隐私损失预算分配，3月16日。https://www2.census.gov/programs-surveys/decennial/2020/program-management/data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/ 2022-03-16_Summary_File/2022-03-16_Privacy-Loss_Budget_Allocations.pdf D. Van Riper、T. Kugler 和 S. Ruggles。2020年。人口普查局2010年演示数据产品中的披露规避。载于：统计数据库隐私 – *PSD 2020*，施普林格，第353-368页。Y. Wang、X. Wu 和 D. Hu。2016年。使用随机响应实现差分隐私保护的数据收集。载于：*EDBT/ICDT 2016*联合会议论文集，法国波尔多。S. L. Warner。1965年。随机响应：一种消除回避作答偏差的调查技术。《美国统计学会杂志》，60(309):63-69。

M. Wines。2022年。2020年人口普查显示，人们似乎生活在水下。这是有原因的。《纽约时报》，4月21日。https://www.nytimes.com/2022/04/21/us/census-data-privacy-concerns. html

W. Winkler。1999年。记录链接的现状与当前研究问题。技术报告，统计研究部，美国人口普查局。