发布带有强隐私保护保证的维基百科使用数据

Temilola Adeleye[1], Skye Berghel[2], Damien Desfontaines[2], Michael Hay[2], Isaac Johnson[1], Cléo Lemoisson[1], Ashwin Machanavajjhala[2], Tom Magerlein[2], Gabriele Modena[1], David Pujol[2], Daniel Simmons-Marengo[2], and Hal Triedman[1]

[1]Wikimedia Foundation – htriedman@wikimedia.org

[2]Tumult Labs – science@tmlt.io

Abstract

近 20 年来，维基媒体基金会一直在发布有关每天有多少人访问每个维基百科页面的统计数据。这些数据帮助维基百科编辑确定在哪里集中精力改进在线百科全书，并支持学术研究。2023 年 6 月，在 Tumult Labs 的帮助下，维基媒体基金会回应了来自维基百科编辑和学术研究人员长期以来的要求：开始以更细粒度发布这些统计数据，包括每天页面访问量的来源国家。这项新的数据发布使用差分隐私技术为浏览或编辑维基百科的人提供强大的保护措施。本文描述了这一数据发布的详情：其目标、从启动到部署所遵循的过程、用于生成数据的算法以及数据发布的结果。

1 介绍

维基百科和其他由维基媒体基金会支持的项目是世界上使用最广泛的在线资源之一，每年从世界各地获得数百亿次访问。因此，该基金会可以访问关于维基媒体项目页面访问量的数太字节的数据。这在本文中被称为*页面查看次数*数据。

基金会近 20 年来一直在通过*页面查看 API* [17] 发布这些数据的统计信息。这些数据有助于维基百科编辑者衡量其工作的影响力，并将精力集中在最需要的地方。页面浏览量数据也是学术研究的重要资源：它已被用于更好地理解许多主题，从用户行为 [14] 和浏览模式 [15] 到信息传播 [1]、流行病学 [19]、在线骚扰 [33] 等。随着时间的推移，维基媒体基金会收到了许多请求，要求使这些统计数据更加细化，并发布页面浏览次数*按国家划分*，以便为维基百科编辑者提供更大的帮助，并促进进一步的学术研究。

处理此类对更细化数据的请求与基金会的开放获取政策 [12] 一致，旨在尽可能透明地展示维基媒体项目如何运作。然而，基金会还认为隐私是自由知识运动的关键组成部分：没有强有力的隐私保障就无法创造或消费自由知识。这些保证通过基金会严格的隐私政策 [13] 和数据保留指南 [6] 表达，它们规定了维基百科背后的基础设施如何运作。具体来说，浏览维基百科的人可能期望他们在网站上的行为保持私密：防止有动机的行动者将这些数据与其他外部数据源结合以监视或迫害维基百科用户的行为历史、编辑历史或其他行为至关重要。众所周知，简单地聚合数据本身不足以防止重新识别的风

险 [34, 23, 22, 25] , 因此发布具有更细地理粒度的数据需要对维基百科用户和编辑采取具有坚实隐私保障的方法。

差分隐私 [27](DP) 提供了一种缓解这种紧张局势的方法：它允许组织在降低和更全面理解发布数据的风险的同时进行操作。因此，维基媒体基金会决定调查使用差分隐私来发布按国家划分的每日页面浏览量数据的可能性。经过对现有开源工具 [5] 的深入比较后，维基媒体基金会决定使用 Tumult Analytics[30, 18] 并与 Tumult Labs 开始合作设计并部署用于此数据发布的 DP 管道。该管道现已部署完毕，发布的数据为任何希望更好地了解 Wikipedia 使用情况的人提供了有用的见解。

本文件更详细地描述了此次数据发布。

- 在第 2 节中，我们展示了用于部署差异隐私数据发布的高级工作流程。
- 在第 3 节中，我们概述了问题陈述和此次数据发布成功指标。
- 在第 4 节中，我们描述了用于此次数据发布的技术算法。
- 在第 5 节中，我们总结了此次部署的结果。

2 差分隐私部署的高级工作流程

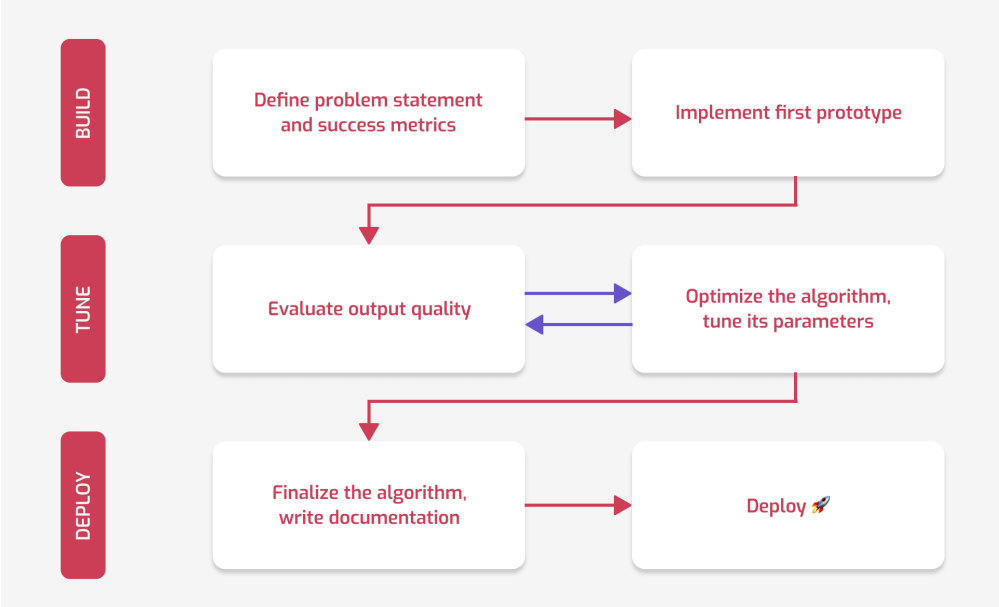

Figure 1: 一种用于差异隐私数据发布的标准化工作流程。

启动 DP 数据产品的过程遵循一个标准的工作流程，包括三个主要阶段：构建, 调整和部署。整个过程如图 1 所示；其三个主要阶段如下。

1. 在初始构建阶段，目标是充分理解问题及其需求，并实现一个初步的算法。这个初期阶段包括两个步骤。首先，我们正确地定义了问题，并确定了该项目成功的标准。这涉及与相关方交谈以了

解数据将用于何处，以及哪些准确性指标能够很好地捕捉到下游用例。其次，我们构建了一个原型机制。这是解决数据发布问题的一个初步的粗略尝试，它揭示了项目固有的“杠杆”。在构建原型时我们需要做出什么选择？这些选择中的哪一些可以在后期进行修改以挑选不同效用或隐私度量之间的权衡？

2. 然后，在调整步骤中，我们使用这些杠杆来尝试不同的设置并优化算法。使用前一步定义的成功指标，我们迭代评估和调整算法，直到其产生的数据符合使用要求并满足隐私需求。

3. 最后，在部署阶段，我们最终确定算法，获得发布数据所需的批准，编写关于数据发布机制的文档以供未来的数据使用者和管道维护者参考，并将其部署到生产环境中。

在第 3 节中，我们概述了第一步的输出：问题陈述及其成功指标的定义。然后，在第 4 节中，我们将描述调优阶段的输出：初始原型经过多轮迭代后，最终算法是什么样的。

3 问题陈述和成功指标

在本节中，我们描述了期望的输出数据（第 3.1 节），输入数据的模式和特征（第 3.2 节），此数据发布所追求的隐私目标（第 3.3 节）以及用于量化成功的准确性指标（第 3.4 节）。

3.1 期望的输出数据

预先存在的 Pageview API 发布关于每个维基媒体页面在给定一天内被访问次数的数据。每个页面通过两个字段进行标识：

- 其*项目*，例如，*法语维基百科*（法语维基百科），*维基教科书中文版*（中文维基教科书，一个开放内容的教材集合），*维基数据*（结构化数据的中央存储）等。；
- 其*页面 ID*，一个数字标识符，唯一标识项目中的每一页。

表 1 是可以通过 Pageview API 获得的虚构数据样本。例如，第一行表示在 2023 年 4 月 2 日，英语维基百科上 ID 为 23110294 的页面有 4217 次访问。

Project	Page ID	Date	Count
en.wikipedia	23110294	2023-04-02	4217
fr.wikipedia	28278	2023-04-02	710
...	...	...	...

Table 1: 一个通过 Pageview API 公开提供的数据中虚构的样本。

该项目的目标是发布更细化的数据，并且每天发布页面浏览次数每国。所需输出数据的虚构样本出现在表 2 中。例如，第一行表示先前提到的访问中有 92 次来自瑞士。

Project	Page ID	Date	Country	Count
en.wikipedia	23110294	2023-04-02	CH	92
fr.wikipedia	28278	2023-04-02	FR	101
...	...	...	...	

Table 2: 一个虚构的样本，我们希望将其作为该项目的一部分发布。

3.2 输入数据

该项目使用两个输入数据集：*当前页面浏览数据集*和*历史页面浏览量数据集*。

当前页面浏览数据集 当用户访问该网站时，他们的单个页面浏览量会被记录并存储在当前的页面浏览数据集中。此数据集包含过去 90 天内所有维基媒体项目的所有页面浏览量。由于维基媒体基金会致力于最小化数据保留，这些数据仅以这种形式保存 90 天。表 3 是当前页面浏览数据集的一个虚构样本，仅显示了该项目感兴趣的列：项目、页面 ID、日期和时间以及国家。

Project	Page ID	Date and Time	Country
en.wikipedia	23110294	2023-04-02 10:32:45	CH
fr.wikipedia	28278	2023-04-02 18:53:11	FR
...	...	...	...

Table 3: 当前页面视图数据集中的感兴趣列的虚构样本。

请注意，与大多数网站类似的日志基础设施不同，这些数据不包含持久的用户标识符。对维基媒体项目的大多数访问来自未登录用户，维基媒体基金会故意没有实现能够提供 Cookie ID 并允许其系统识别两条记录是否来自同一用户的用户跟踪机制。这种做法有利于数据最小化，但使获得用户级别的差分隐私保证变得更加困难，这需要限制来自同一用户的贡献数量。我们在第 4.1.1 节回过头来讨论这个挑战。

历史页面浏览数据集 初始 90 天的保留期过后，页面浏览量会被聚合为每小时总计，并按项目、页面 ID、国家和地区以及一系列用户特征进行细分。这些聚合数据随后被存储在历史页面浏览数据集中。表 4 是历史页面浏览数据集的一个虚构示例样本，同样仅显示感兴趣的列。

Project	Page ID	Date and Time	Country	Count
en.wikipedia	23110294	2023-04-02 10:00	CH	11
fr.wikipedia	28278	2023-04-02 18:00	FR	15
...	...	...	...	...

Table 4: 来自预聚合的历史页面浏览数据集的感兴趣列的虚构样本。

这些预聚合的数据也为执行差分隐私计算带来了挑战：无法确定哪些贡献来自哪些用户，因此也无法限定每个用户的贡献数量。

3.3 隐私目标

在使用差分隐私时，必须决定要保护数据中的哪些内容；或者等效地，邻域数据库的定义应该是什么。对于长时间运行的数据管道，在无界时间范围内定期发布数据，这种选择有两个方面：将哪些时间段视为隐私单元的一部分，以及在这些每个时间段内我们正在保护什么。然后，一个后续问题是隐私参数的选择：ε 和 δ 的数值。

我们的目标是每天发布数据：在隐私单位中使用每日时间段是很自然的。这个区间与几乎所有其他长期运行的 DP 部署保持一致，如苹果公司的遥测收集，或谷歌和 Meta 公司与 COVID-19 危机相关的数据发布。其他发布的周期较短，比如微软在 Windows 中的遥测。天与天之间没有重叠：每个用户每天的隐私参数是固定的，并且不会随着时间增加。

这种隐私单位的选择意味着，如果用户从同一国家使用多个设备（或在每次访问页面时清除 cookies）长时间定期访问同一个页面，这种行为可能在输出数据中被观察到。另一个需要注意的是，这些数据发布显示了群体级别的趋势，比如一个国家内维基媒体项目中的少数语言群体活动。这些见解可能是有用的（例如，可以为该少数语言群体提供专门的支持），但也可能带来风险（例如，可能会导致政府对该少数群体的迫害）。我们通过选择保守的隐私参数来降低这些风险，这在较长的时间段内提供了合理的保护水平，并且通过推迟发布某些国家的数据以及只发布超过一定阈值的汇总数据来实现这一点。

保护每个维基百科用户在每一天的安全是不可能完全实现的，除非有一种方法可以将人们的身份链接到记录和设备上。由于维基媒体基金会没有也不希望具备以这种方式链接记录的能力，我们转而尝试保护每天每个设备的贡献。对于基于当前页面浏览数据集的数据，我们使用*客户端贡献边界*来实现这一目标，如第 4.1.1 节所述。对于基于历史页面浏览数据集的数据，我们无法限定用户的贡献。相反，我们选择保护固定数量的每日页面浏览量，记为 m。这为每天贡献少于 m 次页面浏览的用户提供等同水平的保护。贡献超过 m 次页面浏览的用户将会遭受更大的隐私损失，与他们的贡献超出 m 的数量成正比。这个数值对于 2017 年 2 月 8 日之前的数据设定为 300，而对于从 2017 年 2 月 9 日到 2023 年 2 月 5 日之间的数据则设定为 30。表 5 总结了该项目选择的隐私单元。

Time period of the input data	Unit of privacy
July 1st, 2015 – February 8th, 2017	300 daily pageviews
February 9th, 2017 – February 5th, 2023	30 daily pageviews
February 6th, 2023 – present	one user-day

Table 5: 本项目中使用的隐私单元的概述。

这种在保护多少贡献方面的差异是由于 2017 年 2 月发生了一种变化，影响了输入数据的生成方式。在 2017 年 2 月 8 日之前，编辑维基媒体页面并使用 Web UI 预览更改的用户每次刷新预览都会记录为一个页面访问量。这意味着，在长时间的编辑会话中，一名编辑者可能在同一页面上累积很多页面访问量。当结合我们无法限制用户贡献的能力时，这在该日期前后造成了显著不同的风险水平，我们的历史页面访问算法需要解决这一问题。从 2017 年 2 月 9 日起，预览不再被记录为页面访问量。

对于隐私参数，我们对较新的数据使用零集中差分隐私 [20] 与 $\rho = 0.015$[1] ，而对于历史数据则使

[1]这是一个比 (ε, δ)-DP [26] 与 $\varepsilon = 1$ 和 $\delta = 10^{-7}$ 更为严格的优势保证。

用纯差分隐私 $\varepsilon = 1$。这些值通常被认为在差分隐私的研究者和从业者中是保守的 [31]，并且低于大多数实际部署中的差分隐私 [24]。

3.4 准确度指标

我们从三个维度测量效用：*相对误差分布*、*丢弃率*和*虚假率*。这些指标中的每一个都是使用*真实数据*作为基线进行计算的：该数据对应于仅仅运行一个分组查询（对于当前页面浏览数据集是计数行的数量，对于历史页面浏览数据集则是求和计数），不包含任何贡献边界设定、噪声添加或抑制。

相对误差分布 我们正在发布页面浏览次数，并且 DP 过程会向这些计数中注入统计噪声。因此，自然想要测量添加到这些计数中的噪声有多少。我们根据*相对误差*来衡量准确性：每个带噪声的计数 $\hat{c}$ 的相对误差是 $|\hat{c}/c|$，其中 c 是真实计数。当然，我们会发布很多计数，所以我们需要查看相对误差的*分布*。更具体地说，我们观察相对误差小于 10%、25% 和 50% 的释放计数的百分比。

掉线率 动态规划算法使用*抑制*：如果一个带噪声的计数低于给定阈值，我们将它从输出数据中移除。为了量化由于这一抑制步骤导致的数据损失，计算*掉线率*是很自然的：即那些在真实数据中非零但在输出中未出现的计数的百分比。然而，在真实数据中，许多计数非常低；抑制这样的计数并不像抑制一个流行页面那样糟糕。因此，我们计算那些真实计数大于固定阈值 t 的页面中被抑制的比例（即*掉线率高于* t），以及在真实数据的前 1000 行页面中被抑制的比例（即前 *1000 位淘汰率*）。

虚假率 许多页面、项目和国家组合在任何特定的一天都会收到零次访问。当向这些零计数添加噪声时，它们最终可能会变成正的（尽管相对较小）计数。我们把这些称为*虚假*的计数。虚假计数可能通过错误地表明某些组合具有活动来误导数据用户。它们还会增加输出数据集的大小，这可能会造成使用上的挑战。因此，我们计算了一个额外的指标：*虚假率*，它捕捉了输出中所有计数中虚假计数的比例。

4 算法的技术描述

在本节中，我们描述了用于生成差异隐私数据的算法。为了简化起见，我们将一个 < 页面 ID 和项目 > 对称为一个*页*。

4.1 当前页面浏览量

对于使用当前页面浏览量数据集的数据，我们希望提供隐私保护，以在每一天中保护每个用户。这需要限定每个用户在一个单一日期内可以贡献的最大页面浏览量数量。执行这种贡献限制的典型方法是使用用户标识符来对每个用户的贡献数量进行子采样，选取前 k 条记录 [29]，或者使用蓄水池抽样 [32]。然而，在没有用户标识符的情况下，我们不得不采用一种新颖且替代的方法来解决这个问题：*客户端过滤*。

4.1.1 客户端过滤

没有用户 ID，服务器无法知道多个贡献是否来自同一用户，并执行贡献边界以获得用户级别的隐私保证。相反，我们在客户端添加了一些逻辑。每个终端用户的设备计算他们在每天记录的贡献次数，并将每次贡献连同布尔标志一起发送，该标志指示此贡献是否应用于服务器端的 DP 计算。输入到 DP 算法中的标准如下：每天我们包括前 10 个唯一页面浏览量。这意味着如果用户在一天内多次访问同一页面，只有第一次访问会被计入。这也意味着如果用户在一天内访问了超过 10 个不同的页面，则第 10 次之后的所有页面浏览将不会被纳入。

该客户端过滤步骤的伪代码可以在算法 1 中找到。请注意，此算法不会在客户端 Cookie 中跟踪原始页面 ID。相反，它使用带盐哈希函数 [16] 来记住哪些页面 ID 已经被访问过。这为防止攻击者获取该 Cookie 提供了额外一层保护。

Algorithm 1 客户端过滤算法

Require: $P = p_1, p_2, \ldots$ ： 一个页面浏览量的流。

Require: H ： 一个加盐哈希函数

Require: k ： 要包含的独特页面浏览量数量。

Ensure: 输出是一个相同页面浏览量的流，每个都被标注了一个布尔值，表示是否应将其纳入 DP 计算中。此布尔值为 `true` 当且仅当页面视图来自之前未输出的页面，并且

$S \leftarrow \{\}$
for p in P **do**
    if $|S| \geq k$ or $H(p) \in S$ **then**
        输出 $\langle p, \texttt{false} \rangle$
    else
        $S \leftarrow S \cup H(p)$
        输出 $\langle p, \texttt{true} \rangle$
    end if
end for

客户端过滤遵守维基媒体基金会的数据最小化原则：仅将执行贡献边界所需绝对最少的信息——与每个页面视图关联的布尔值，以指示是否应将其计入——添加到日志基础设施中。使用标识符或每次贡献时递增的计数器等替代方案则需要向服务器发送更多数据，并增加指纹识别风险。

4.1.2 服务器端算法

一旦每个页面视图被客户端过滤算法标注后，它将作为输入用于服务器端的差分隐私算法。该算法每天运行一次，使用前一日的数据，并分为三个阶段。

1. 首先，我们收集要汇总的 < 页面、国家 > 元组列表。
2. 第二，我们计算每个组的页面浏览量，并对每个计数添加噪声。
3. 最后，我们抑制低计数，并发布数据。

所有可能的元组列表理论上都是提前已知的：维基媒体页面和国家的列表都是公开信息。然而，大多数 < 页面、国家 > 组合在输入数据中并未出现：如果包含所有这些组合将是低效且会导致虚假数据增加。相反，我们使用现有的公共数据仅包括这些可能计数中的小部分。每天，我们会列出根据现有页面视图 API 拥有超过 t 全球页面浏览量的所有维基媒体页面，其中 t 是一个任意的摄入阈值。然后，我们将这些页面与国家列表 [2] 进行交叉组合以创建分组。

第二步使用高斯机制 [28] 向计数中添加噪声。这提供了两个优势。首先，因为每个用户最多可以贡献 10 个不同的 < 页和国家 > 元组，但对每个只贡献一次，我们得到了一个比使用 L_1 敏感度（k）更紧密的 L_2 敏感度界限（$\sqrt{k}$）：这使我们可以添加较少的噪声。其次，因为高斯噪声分布的尾部衰减非常快，这使得阈值步骤在防止输出中出现零计数方面更为高效，将虚假率保持在可接受的较低水平。我们使用零集中差分隐私（zCDP）[20] 来量化高斯机制的隐私保证。

第三步很简单：所有低于阈值 τ 的计数都被从输出中移除。这一步是必要的，因为第一步会产生许多非噪声用户计数非常低甚至为 0 的 < 页面、国家 > 元组。这样的计数会导致相对误差过高和虚假率增加。与数据用户的交流表明，这些使得输出数据集难以使用，并且用户最感兴趣的是最受关注的页面，而不是那些浏览量很少的长尾页面。抑制低于固定且可配置阈值 τ 的计数可以解决这个问题，但代价是非零的删除率。

该机制在算法 2 中给出；在此算法中，$\mathcal{N}(0, \sigma^2)$ 表示从均值为 0，方差为 σ^2 的正态分布中抽取的一个随机样本。第一步仅使用公共数据，第二步提供 ρ-zCDP [20]，第三步是后处理步骤：整个算法满足 ρ-zCDP。

我们使用 $k = 10$ 作为每个用户的每日贡献上限，$t = 150$ 作为摄入阈值，并且使用 $\tau = 90$ 作为抑制阈值。这些值是在广泛的实验之后选定的，目的是为了输入数据集的完整性并优化在第 3.4 节中描述的效用指标。

为了选择这些算法参数，我们使用真实数据计算了指标。这样的指标在原则上是敏感的，而这些参数本身并不是差异性私有的。为减轻此调参过程中的隐私风险，我们在整个调参过程中将细粒度效用指标保密，以最小化数据泄露。除了这一考虑之外，我们仅公开通信全局效用指标的大致值以及从这个调参过程中获得的算法参数。

尽管如此，这仍然是一个有效的批评，我们希望进一步研究通过在敏感指标上进行保密调整所导致的隐私损失。

4.2 历史页面浏览量

为了使用历史页面浏览数据集作为输入数据计算差分隐私计数，我们遵循一个类似的过程，但有一个关键的区别：由于数据已经被预先聚合，因此不可能执行每个用户的贡献限制。因此，我们不使用客户端过滤步骤，而是使用如第 3.3 节所述的不同隐私单位。我们也必须对预聚合数据的 Count 列进行求和，而不仅仅是计算每组中的行数。另一个不同之处在于使用拉普拉斯噪声而不是高斯噪声，这是由于我们只对聚合的 L_1 敏感性有界，而不像当前页面浏览数据那样是 L_2。整个过程在其他方面与之前的类似。

1. 首先，我们收集要汇总的 < 页面、国家 > 元组列表。

[2] 此列表基于 [7]；排除了维基媒体基金会认为对记者或网络自由可能构成危险的国家 [3]。

Algorithm 2 服务器端算法用于当前页面访问量统计

Require: t ： 一个摄取阈值。

Require: τ ： 一个抑制阈值。

Require: ρ ： zCDP 的隐私参数。

Require: $P = \langle p_1, b_1 \rangle, \langle p_2, b_2 \rangle, \ldots$ ： 一个带有标注的页面浏览私有数据集，其中每个用户最多与 k 个独特的页面浏览 $\langle p_i, b_i \rangle$ 相关联，且 $b_i = \texttt{true}$，所有这些都具有不同的 p_i。

Require: $P_{daily} = \langle p_1, n_1 \rangle, \langle p_2, n_2 \rangle, \ldots$ ： 一个列出每一页全球页面浏览数量的公共数据集。

Require: C ： 一组预定义的国家列表。

步骤 *1*: 收集聚合组

$G \leftarrow \{\}$
for $\langle p, n \rangle$ in P_{daily} **do**
  if $n \geq t$ **then**
    for c in C **do**
      $G \leftarrow G \cup \langle p, c \rangle$
    end for
  end if
end for

步骤 *2*: 计算噪声计数

$\sigma \leftarrow \sqrt{\frac{k}{2\rho}}$
$O \leftarrow \{\}$
for g in G **do**
  $c \leftarrow |\{p \in P \mid p = g\}|$
  $\hat{c} \leftarrow c + \mathcal{N}(0, \sigma^2)$
  $O \leftarrow O \cup \langle g, \hat{c} \rangle$
end for

步骤 *3*: 抑制低计数

for $\langle g, \hat{c} \rangle$ in G **do**
  if $\hat{c} < \tau$ **then**
    $G \leftarrow \{\}0$
  end if
end for
$G \leftarrow \{\}1$

2. 其次，我们对每个组的页面浏览次数进行求和，并向每个总和添加拉普拉斯噪声。

3. 最后，我们抑制低和数，并发布数据。

完整算法如 Algorithm 3 所示；其中，$\mathrm{Lap}(0, \lambda)$ 表示从均值为 0、尺度为 λ 的拉普拉斯分布中抽取的一个随机样本。其隐私分析很简单：第 1 步仅使用公共数据，第 2 步提供 ε-DP 保证 [27]，而第 3 步是后处理步骤，因此整个算法满足 ε-DP。

如第 3.3 节所述，我们使用 $m = 300$ 处理 2015–2017 年的数据，并使用 $m = 30$ 处理 2017–2023 年的数据。对于 2015–2017 年的数据，我们采用 $t = 150$ 作为摄入阈值和 $\tau = 3500$ 作为抑制阈值。对于 2017–2023 年的数据，我们使用 $t = 150$ 作为摄入阈值和 $\tau = 450$ 作为抑制阈值。这些值被选择以优化第 3.4 节中描述的全局效用指标。

4.3 实现

算法使用了 Tumult Analytics [30, 18] 进行实现和部署，该框架因其稳健性、生产就绪性、与 Wikimedia 的计算基础设施的兼容性以及对基于 zCDP 的隐私核算等高级功能的支持而被选中 [5]。这在使用机制上产生了非常细微的区别：对于整数值数据，Tumult Analytics 使用的是双侧几何分布而不是拉普拉斯噪声，并且采用高斯机制的离散版本 [21]。基于当前输入数据的数据发布需要在一个框架中实现新的邻域关系概念：不是保护固定数量的行，或者与单个用户标识关联的任意数量的行，而是保护固定数量的行与不同的聚合组相关联。这得益于底层框架 Tumult Core 的可扩展性。

5 结果

该项目现在允许维基媒体基金会发布有关用户访问维基媒体项目的大得多且内容丰富得多的数据集。发布的页面浏览数据量的增加程度总结在表 6 中。

	Before this project	After this project	Percentage change
每天释放的数据点中位数	9,000	360,000	+4,000%
每日发布的中位数页面浏览量	50 million	120 million	+240%
自 2021 年以来发布的数据点总数	8 million	120 million	+1,500%
自 2021 年以来发布的总页面浏览量	47 billion	116 billion	+250%

Table 6: 对比该项目前后发布数据的数量，截至 2023 年 6 月 29 日。

从 2015 年到 2021 年的超过 2,000 天的历史数据之前未发表。在此项目中使用差分隐私使得维基媒体基金会能够发布关于这些数据的超过 1.35 亿条统计信息，涵盖 3,250 亿次页面浏览。

输出数据根据我们的成功指标具有可接受的质量。

- 对于基于当前页面浏览数据集的数据，超过 95%的计数其相对误差低于 50%，高于 150 的下降率低于 0.1%，全局虚假率低于 0.01%，除 3 个国家外的所有国家均低于 3%。

Algorithm 3 历史页面浏览量算法

Require: m ： 每天受保护的页面访问次数。

Require: t ： 一个摄入阈值。

Require: τ ： 一个抑制阈值。

Require: ε ： 用于差分隐私的隐私参数。

Require: $P_{hourly} = \langle p_1, c_1 \rangle, \langle p_2, c_2 \rangle, \ldots$ ： 一个预先聚合每小时页面浏览量的私有数据集列表。

Require: $P_{daily} = \langle p_1, n_1 \rangle, \langle p_2, n_2 \rangle, \ldots$ ： 一个列出每个页面全球页面浏览量的公共数据集列表。

Require: C ： 预定义的国家列表。

Require: t ： 将页面包含在输出中的最小浏览量阈值。

步骤 *1*: 收集聚合组

$G \leftarrow \{\}$

for $\langle p, n \rangle$ in P_{daily} **do**

  if $n \geq t$ **then**

    for c in C **do**

      $G \leftarrow G \cup \langle p, c \rangle$

    end for

  end if

end for

步骤 *2*: 计算噪声和

$\lambda \leftarrow \frac{m}{\varepsilon}$

$O \leftarrow \{\}$

for g in G **do**

  $s \leftarrow \sum_{\langle p,c \rangle \in P_{hourly} \text{where} p=g} c$

  $\hat{s} \leftarrow s + \text{Lap}(0, \lambda)$

  $O \leftarrow O \cup \langle g, \hat{s} \rangle$

end for

步骤 *3*: 抑制低计数

for $\langle g, \hat{s} \rangle$ in G **do**

  if $\hat{s} < \tau$ **then**

    $G \leftarrow \{\}0$

  end if

end for

$G \leftarrow \{\}1$

- 对于 2017–2023 年的数据，前 1000 名的下降率中位数低于 8%，超过 450 名的下降率低于 3%，全球虚假率低于 0.1%。
- 对于 2015 年至 2017 年的数据，前 1000 位的下降率低于 40%，超过 3500 的下降率低于 3%，全局虚假率低于 20%。

这些指标表明，近期数据的隐私与准确性权衡比历史数据好得多：这归因于客户端过滤带来的更严格的敏感度界限，使得可以充分利用高斯机制及其快速衰减的尾部。

6 结论

在这篇论文中，我们描述了维基媒体基金会发布有关维基百科及其他维基媒体项目用户行为的大型数据集的过程和机制。多个关键因素使得这一发布成为可能。

- Tumult Labs 的差分隐私出版物的系统工作流程，在第 2 节中描述，提供了从项目启动到部署所需的基本结构。
- 将客户端过滤与服务器端聚合相结合，如第 4.1 节所述，是我们获得当前页面浏览数据的用户级差分隐私保证而不追踪用户标识符的关键创新。
- 混乱核心，作为混乱分析背后的数据隐私框架，旨在具有可扩展性。这使我们能够在此框架中添加一种新颖的邻域定义，以捕捉客户端过滤特性，同时仍能使用紧缩的隐私计算技术。
- 最后，Tumult Analytics 提供的可扩展性在处理本项目中用作输入的海量数据集方面是至关重要的。

数据现已在线发布 [9, 10, 11]，沿同客户端过滤基础设施的源代码 [2] 和服务器端算法 [4, 8]。我们期待看到这些数据将启用哪些用例！

7 致谢

我们感谢 Luke Hartman、Tomoko Kitazawa、Nuria Ruiz 和 Xabriel J. Collazo Mojica 对本项目提供的帮助，以及 Leila Zia 和匿名审稿人对此论文提出的宝贵意见和建议。

References

[1] Academic studies about Wikipedia – Wikipedia. https://en.wikipedia.org/wiki/Academic_studies_about_Wikipedia.

[2] Clien-side filtering code – Wikimedia Gerrit. https://gerrit.wikimedia.org/r/plugins/gitiles/operations/puppet/+/refs/heads/production/modules/varnish/templates/analytics.inc.vcl.erb#171.

[3] Country protection list – Wikitech. https://wikitech.wikimedia.org/wiki/Country_protection_list.

[4] Differential Privacy – Wikimedia GitLab. https://gitlab.wikimedia.org/repos/security/differential-privacy/.

[5] Differential privacy/Docs/Infrastructure and framework decision-making process – Wikimedia Meta-Wiki. https://meta.wikimedia.org/wiki/Differential_privacy/Docs/Infrastructure_and_framework_decision-making_process.

[6] Legal:Data retention guidelines – Wikimedia Foundation. https://foundation.wikimedia.org/wiki/Legal:Data_retention_guidelines.

[7] List of countries by ther United Nations geoscheme – Wikipedia. https://en.wikipedia.org/wiki/List_of_countries_by_the_United_Nations_geoscheme.

[8] Pageview historical notebooks – htriedman GitLab. https://gitlab.wikimedia.org/htriedman/stat-spark3/-/tree/main/pageview_historical/notebooks.

[9] Pageviews Differential Privacy – Current – README. https://analytics.wikimedia.org/published/datasets/country_project_page/00_README.html.

[10] Pageviews Differential Privacy – Historical – README. https://analytics.wikimedia.org/published/datasets/country_project_page_historical/00_README.html.

[11] Pageviews Differential Privacy – Historical (pre-2017) – README. https://analytics.wikimedia.org/published/datasets/country_project_page_historical_pre_2017/00_README.html.

[12] Policy:Open access policy – Wikimedia Foundation. https://foundation.wikimedia.org/wiki/Policy:Open_access_policy.

[13] Policy:Privacy policy – Wikimedia Foundation. https://foundation.wikimedia.org/wiki/Policy:Privacy_policy.

[14] Research:Projects – Wikimedia Meta-Wiki. https://meta.wikimedia.org/wiki/Research:Projects.

[15] Research:Wikipedia clickstream – Wikimedia Meta-Wiki. https://meta.wikimedia.org/wiki/Research:Wikipedia_clickstream.

[16] Salt (cryptograph) – Wikipedia. https://en.wikipedia.org/wiki/Salt_(cryptography).

[17] Wikimedia Downloads: Analytics Datasets. https://dumps.wikimedia.org/other/analytics/.

[18] Skye Berghel, Philip Bohannon, Damien Desfontaines, Charles Estes, Sam Haney, Luke Hartman, Michael Hay, Ashwin Machanavajjhala, Tom Magerlein, Gerome Miklau, Amritha Pai, William Sexton, and Ruchit Shrestha. Tumult Analytics: a robust, easy-to-use, scalable, and expressive framework for differential privacy. *arXiv preprint arXiv:2212.04133*, December 2022.

[19] Nicola Luigi Bragazzi, Cristiano Alicino, Cecilia Trucchi, Chiara Paganino, Ilaria Barberis, Mariano Martini, Laura Sticchi, Eugen Trinka, Francesco Brigo, Filippo Ansaldi, et al. Global reaction to the recent outbreaks of zika virus: Insights from a big data analysis. *PloS one*, 12(9):e0185263, 2017.

[20] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

[21] Clément L Canonne, Gautam Kamath, and Thomas Steinke. The Discrete Gaussian for Differential Privacy. In *Advances in Neural Information Processing Systems*, volume 33, pages 15676–15688. Curran Associates, Inc., 2020.

[22] Aloni Cohen. Attacks on deidentification's defenses. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1469–1486, 2022.

[23] Damien Desfontaines. Demystifying the US Census Bureau's reconstruction attack. https://desfontain.es/privacy/us-census-reconstruction-attack.html, 05 2021. Ted is writing things (personal blog).

[24] Damien Desfontaines. A list of real-world uses of differential privacy. https://desfontain.es/privacy/real-world-differential-privacy.html, 10 2021. Ted is writing things (personal blog).

[25] Travis Dick, Cynthia Dwork, Michael Kearns, Terrance Liu, Aaron Roth, Giuseppe Vietri, and Zhiwei Steven Wu. Confidence-ranked reconstruction of census microdata from published statistics. *Proceedings of the National Academy of Sciences*, 120(8):e2218605120, 2023.

[26] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 486–503. Springer, 2006.

[27] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[28] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

[29] Aleksandra Korolova, Krishnaram Kenthapadi, Nina Mishra, and Alexandros Ntoulas. Releasing search queries and clicks privately. In *Proceedings of the 18th international conference on World wide web*, pages 171–180, 2009.

[30] Tumult Labs. Tumult Analytics. https://tmlt.dev, December 2022.

[31] Joseph Near and David Darais. Differential privacy: Future work & open challenges. *Cybersecurity insights*, 2022.

[32] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. Differentially private SQL with bounded user contribution. *Proceedings on Privacy Enhancing Technologies*, 2:230–250, 2020.

[33] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399, 2017.

[34] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data. In *Proceedings of the 26th international conference on world wide web*, pages 1241–1250, 2017.