arXiv:2302.02056v1 [cs.DS] 4 Feb 2023

# Sketch-Flip-Merge：用于隐私保护不同元素计数的可合并草图

Jonathan Hehir*[1]、Daniel Ting[2],与Graham Cormode[3]

[1]宾夕法尼亚州立大学统计学系，美国[2]Meta，美国[3]Meta，英国

2023年2月7日

**摘要**

数据摘要是不同元素计数的重要工具，它能够将多重集表示为紧凑的摘要，从而快速估算基数。由于草图可以合并以概括多重集的并集，因此它们是数据仓库中的基础构件。尽管已有许多实用的基数估计草图，但没有一种能在合并时保护隐私。我们提出首批兼具可合并性、差分隐私（DP）和低经验误差的实用基数草图。这些草图采用了一种用于对含噪比特执行逻辑运算的新型随机化算法，并提供严格的隐私分析和可证明最优的估计方法。在模拟和真实世界数据中，我们的草图显著优于现有的理论方案。

## 1 引言

许多用于对海量数据建模的应用，都以跟踪事件或观测值的基数为基础。因此，这些应用广泛使用支持快速近似基数估计的数据草图（Cormode and Yi，2020）。例如，Amazon Redshift、ClickHouse、Google BigQuery、Splunk、Presto 和 Redis 等热门数据管理系统，都通过 HyperLogLog（HLL）草图的各种变体支持近似不同元素计数（Flajolet 等，2007；Heule 等，2013）。这些近似方法只需承受少量估计误差，就能大幅降低不同元素计数的计算成本，使其可在线性时间内完成，且仅需有界内存。不同计数草图的另一项关键特性，是能够合并两个或多个草图，从而得到其并集的基数估计。这不仅支持分布式计算，还能利用先前计算的草图实现丰富多样的聚合。因此，现代数据流水线高度依赖此类基数草图的性能与功能。

隐私问题日益成为数据处理操作的约束。致力于保护用户隐私的组织要求，对收集自个人的数据在

*这项研究是在作者任职于 Meta 期间进行的。

传递给下游处理之前采取适当的缓解措施。具体而言，采用差分隐私等保护措施，在保护敏感数据的同时，仍能提供准确的查询响应。

尽管草图技术似乎能通过减少数据来提供保护，但众所周知，单靠草图并不能自动提供隐私保证（Desfontaines 等，2019）。摘要——甚至根据摘要计算出的估计值——都可能泄露大量信息，让人得知特定项是否属于底层集合。近期研究表明，如果 观察者不知道相关哈希函数，草图内容确实符合隐私标准（Choi 等，2020；Smith 等，2020；Dickens 等，2022）。然而，在大规模系统中由多个实体共享计算时，假设哈希函数是秘密的并不现实。特别是，所有参与者在处理将要合并的草图时都必须知道哈希值，而在多个草图中使用相同的哈希值会产生相关随机性，从而破坏隐私保障。这使得如何让这些高吞吐量系统实现隐私保护成为一个重要的待解问题。以往构建隐私保护草图的尝试（Pagh 和 Stausholm，2021）未能提供实用的可合并草图，因为误差过大（第 6 节）。

在本研究中，我们提出了 Sketch-Flip-Merge（SFM）摘要，这是一种实用、可合并且经证明具备隐私保护能力的不同元素计数草图方法。具体而言，即使哈希函数是公开已知的，我们生成的摘要也满足 $\varepsilon$–差分隐私（DP）的强定义（Dwork 等，2006；Dwork，2008）。由于隐私保证附着于摘要本身，而不仅仅是基数估计，我们可以安全地发布对应敏感多重集的摘要，并使用这些隐私保护摘要代替原始敏感数据，安全地估计这些集合任意并集的基数。

我们方法的关键在于改造 Flajolet 和 Martin（1985）的草图，该草图通常称为 *FM85* 或随机平均概率计数（*PCSA*）。尽管后续的草图（如 HLL）（Durand 和 Flajolet，2003；Flajolet 等，2007；Heule 等，2013）进一步优化了空间占用，但压缩空间会使它们更难实现隐私保护。与 PCSA 不同，这些草图存储极值哈希值；PCSA 简单的分区二进制结构将敏感度限制在一次位翻转，而这些草图中，输入的微小变化可能导致摘要发生巨大变化，因此需要添加更多噪声，结果也就不够准确。此外，我们的方法可推广至任何基于位图的草图。

**相关工作**。 隐私保护基数草图已成为多项早期研究的主题。尽管近期工作为类 HLL 草图提供了差分隐私保证（Smith 等，2020年；Dickens 等，2022），但这些方法依赖随机的秘密哈希函数，因此无法合并草图。Choi 等（2020年）使用固定的公开哈希函数，通过向基数估计量添加噪声，从 LogLog 草图中获得差分隐私基数估计，但草图本身仍然敏感，不适合发布或共享。Stanojevic 等（2017年）设计了一种差分隐私算法，利用经过扰动的布隆过滤器对两个多重集的并集进行基数估计，但其方法无法推广到两个以上多重集的并集，也难以扩展。

有一类研究通过在 PCSA 中引入随机响应并对项进行子采样来实现隐私保护（Tschorsch 和 Scheuermann，2013年；Nu˜nez von Voigt 和 Tschorsch，2019年）。然而，Tschorsch 和 Scheuermann（2013年）未能实现差分隐私保证，而 Nu˜nez von Voigt 和 Tschorsch（2019年）没有探讨如何合并草图。Kreuter 等（2020）设计了两种草图，其中一种基于 PCSA。虽然他们的差分隐私草图无法合并为单个草图，但如果所有草图使用相同的隐私参数，则可以利用多个草图估计并集的基数。Pagh 和 Stausholm（2021年，第 6 节）提出的基于 PCSA 的草图实现了差分隐私并支持合并，但不实用。在我们的实验中，他们的估计器经常无法给出估计值，且返回的误差大得不切实际。最后，Desfontaines 等人（2019年）给出了一个不可能性结果：隐私保护与高准确度无法兼得，但这仅在合并大量

草图时才成立，这与我们的结果一致。

**贡献**。 我们提出了两种实用方法，用于构造可合并的 差分隐私基数草图并获得基数估计。第一种方法采用 Pagh 和 Stausholm（2021）使用的确定性位合并操作。我们证明，即使对先前的隐私分析进行了指数级改进，这种合并仍需要采用次优形式的随机响应（推论 4.5）。我们的主要方法学贡献是一种新颖的随机合并方法，与优化后的确定性合并相比，方差最多可进一步降低 75%。我们将这种随机合并推广到二进制数据上的任意位运算，这一推广本身也可能具有独立价值。我们还开发了一种基于复合似然的基数估计量，并证明，对于基于 PCSA 的隐私保护草图和非隐私草图，该估计量都渐近最优。

**概要**。 我们先在第 2 节简要概述 PCSA 草图方法，然后在第 3 节定义隐私并回顾随机响应。第 4 节通过精心设计随机响应机制以及对随机比特集合的合并操作，实现草图合并。第 5 节中，我们为隐私保护 PCSA 草图提出一种快速基数估计量，并分析其性质。第 6 节中，我们将这些方法与隐私和非隐私替代方案进行比较，并在第 7 节陈述结论。所有证明均收录于附录。

**记号**。 我们记作 $[m] = \{1, \dots, m\}$。 $\otimes$ 表示克罗内克积。逻辑运算分别记作 $\vee$（或）、 $\wedge$（与）、 $\veebar$（异或）和 $\neg$（非）。我们使用自然对数 $\log = \log_e$。分布相等记作 $\overset{D}{=}$。集合 $D$ 的基数记作 $|D|$。

## 2 背景与问题设定

设 $D \in \mathcal{X}^N$ 表示来自某个对象全集 $\mathcal{X}$ 的 $N$ 个项构成的多重集。基数统计问题 是估计 $D$ 中不同元素数量的任务。也就是说，如果 set($D$)表示 $D$中各项的支撑集，那么基数统计问题旨在通过一次遍历数据，使用有界内存的数据草图近似 $n = |\text{set}(D)|$ 。我们研究适用于 可合并 草图的隐私 基数统计问题：草图中的信息满足差分隐私（DP），且草图可以合并，从而得到底层数据集并集的草图。

我们聚焦于基数统计问题的解决方案，其中草图构成二进制向量，并通过逐元素逻辑运算（例如 或）执行合并操作。我们的方法适用的草图包括 PCSA、线性计数（Whang 等人，1990年）、布隆过滤器（Broder 和 Mitzenmacher，2004年）以及 Liquid Legions（Kreuter 等人，2020年）。这些草图尤其适合通过应用随机响应（Warner，1965年）来增强隐私性，但需要谨慎设计针对随机化位的合并操作。尽管这排除了其他常用的草图，例如 HyperLogLog 和 $k$-最小值草图（Bar-Yossef 等人，2002年；Giroire，2009年），但这些草图中存储的值集合更丰富，使其由于具有较高的敏感度而不太适合隐私保护，因此需要引入更多噪声来保障隐私。在本文的其余部分，我们聚焦于 Flajolet 和 Martin（1985年）的 PCSA 草图，并指出第 3 节和第 4 节中关于构建和合并隐私保护草图的结果，可通过直接应用或简单扩展推广到相关草图。

经典 PCSA 草图采用矩阵 $S = \mathcal{S}(D) \in \{0,1\}^{B \times P}$，包含 $B$ 个桶，精度参数为 $P$。给定两个相互独立的通用哈希函数， $h_1(x) \sim$均匀分布($[B]$) $h_2(x) \sim$几何分布($1/2$)，令桶($x$) $= h_1$($x$)、值($x$) $=$最小值$\{P, h_2(x)\}$。那么，当且仅当存在 $x \in D$ 使得桶($x$) $= i$、值($x$) $= j$时，每个位 $S_{ij}$ 都等于 1。 $S$ 的一些理想性质显而易见。首先， $S$ 仅依赖于哈希值集合$\{h_1(x)\}_{x \in D}$ 与 $\{h_2(x)\}_{x \in D}$。因此，它既不受 $D$ 中的重复项影响，也不受处理 $D$ 中元素顺序的影响。此外，两个草图 $\mathcal{S}(D_1)$和 $\mathcal{S}(D_2)$可以合并

通过简单的按位或运算， $\mathcal{S}(D_1) \vee \mathcal{S}(D_2) = \mathcal{S}(D_1 \cup D_2)$，因为合并草图中的每个条目 $S_{ij}$ 当且仅当至少一个 $D_1$， $D_2$ 中存在某个项 $x$，使得桶$(x) = i$，值$(x) = j$时，才等于 1。

重要的是，当对手知道 $h_1$ 与 $h_2$时，草图 $\mathcal{S}(D)$会泄露有关 $D$中元素的信息。例如，任何满足 $S_{\texttt{bucket}(x),\ \texttt{value}(x)} = 0$ 的 $x \in \mathcal{X}$ 都不可能属于 $D$。接下来，我们将扩展 PCSA 草图，以尽量减少此类隐私泄露。

# 3 隐私保护草图

差分隐私（DP）（Dwork 等，2006；Dwork，2008）提供了一种强大且可量化的隐私概念。差分隐私要求作用于数据集 $D$ 的算法（隐私机制）必须采用随机化处理——通常是添加经过精心调整的噪声——使得单条输入记录无法显著影响隐私机制输出的分布。因此，逆向推断单条记录相关信息的能力受到限制，而且仅使用算法输出进行的任何分析也满足差分隐私。差分隐私保证的强度由参数 $\varepsilon > 0$量化，该参数通常称为隐私预算； $\varepsilon$ 越小，隐私保护越强。

**定义 3.1** （Dwork 等人（2006 年））。 若随机化算法 $\mathcal{M}$ 满足 $\varepsilon$**–差分隐私（差分隐私）**，则对于任意两个相邻数据库 $D, D'$ 以及任意输出集合 $E \subseteq \text{Range}(\mathcal{M})$，都有：

$$\Pr(\mathcal{M}(D) \in E) \leq e^{\varepsilon} \Pr(\mathcal{M}(D') \in E).$$

在基数统计问题中，如果多重集 $D, D'$ 可以通过向 $D$添加或移除一个唯一项得到，我们就称它们是相邻的。给定两个相邻多重集 $D, D'$ 及其对应的 PCSA 草图 $\mathcal{S}(D)$ $\mathcal{S}(D')$，根据 PCSA 的定义可知，这些草图至多只有一位不同。因此，要从 $\mathcal{S}(D)$构造差分隐私草图，我们考虑将通用差分隐私机制应用于由 $\{0,1\}$ 位组成的向量，其中两个向量 $x, x' \in \{0,1\}^d$ 若至多有一位不同（即敏感度为 1），则称它们为相邻向量。

当我们将注意力限定在输入和输出均为单个位的机制时，每个差分隐私机制都可视为随机响应（RR）的一种实例。Warner（1965年）提出的随机响应如下所述。令 $\mathcal{F}_{p,\ q}$ 表示一种一般的比特翻转算法，该算法由两个概率 $p$ 与 $q$参数化；当 $q = 1 - p$时，便可得到 Warner 风格的经典随机响应：

$$\mathcal{F}_{p,q}(x) \sim \begin{cases} \text{Bernoulli}(p), & x = 1 \\ \text{Bernoulli}(q), & x = 0 \end{cases}.$$

**定理 3.2**。 假设 $q \leq 1/2 \leq p$。对于敏感度为 *1* 的向量，算法 $\mathcal{M}_{p,\ q}$： $\{0,1\}^d \to \{0,1\}^d$ 对输入中的每个元素独立应用 $\mathcal{F}_{p,\ q}$，当且仅当满足以下条件时，该算法为 $\varepsilon$-差分隐私：

$$p, q \in (0,1) \quad \textit{and} \quad \max\left\{\frac{p}{q}, \frac{1-q}{1-p}\right\} \leq e^{\varepsilon}. \tag{1}$$

定理 3.2 提供了一整族隐私机制，其中任意 $\mathcal{M}_{p,q}$ 满足 $p, q$ 且符合公式 1 的机制，都可应用于 PCSA 草图或任意位向量，使其输出满足 $\varepsilon$-差分隐私。接下来，我们的工作旨在解决几个重要问题：如果草图中的位经过 $\mathcal{F}_{p,q}$ 扰动，要如何合并两个草图？如何根据含噪草图估计基数？以及应如何选择 $p$ and $q$?

# 4 合并扰动位向量

随机响应机制 $\mathcal{M}_{p,q}$ 可将 PCSA 草图转换为具有隐私保护的等效形式，但这会破坏 PCSA 的合并操作。对于普通 PCSA 和多重集 $D_1, D_2$，逐位或$(D_1) \vee \mathcal{S}(D_2) = \mathcal{S}(D_1 \cup D_2)$ 定义了一种草图合并操作，其结果与先取并集再得到的草图相同。然而，对含噪草图执行相同操作并不具备这一理想性质。我们提出了含噪草图上的合并操作，并确定了这些操作存在的条件。具体而言，定理 4.1 表明，如果合并操作是确定性的，那么 异或 $(\veebar)$ 是含噪草图唯一可能的合并方式，而且它仅适用于机制 $\mathcal{M}_{p,q}$ 的某些选择。我们证明，这些选择意味着，即便隐私预算趋近无穷大，这种含噪草图的基数估计方差也至少会比相同大小的普通 PCSA 草图更大。我们的主要贡献是提出一种新颖的随机化 合并操作，它为草图引入的噪声更少。此外，我们还将该操作推广至含噪比特向量上的任意布尔运算。

## 4.1 确定性合并

对 PCSA 草图应用标准随机响应机制会破坏可合并性。PCSA 使用按位或来合并草图，而在存在随机响应噪声时，或操作会产生非随机响应噪声，使位偏向于 1。Pagh 和 Stausholm（2021年）通过在草图更新或合并时，将按位或（$\vee$）替换为按位异或（$\veebar$）操作来解决这一问题。然而，异或操作会破坏基数信息。具体而言，PCSA 草图与自身进行异或后会得到空草图。他们不再确保草图对重复项保持不变，而是确保（合并后）草图的分布保持不变。为此，他们以 $1/2$ 的概率独立地对子采样项（包括重复项）。这实际上会将 PCSA 中值为 1 的位编码为伯努利分布 $(1/2)$ 值，而值为 0 的位仍保持为 0。不幸的是，这种子采样操作会引入大量噪声。图 4 显示，即使对于较大的 $\varepsilon$，所得基数估计的方差也会达到原来的 4 倍。

我们证明，对于任何确定性合并，这种对准确度的损失都是不可避免的。定理 4.1 表明，异或实际上是唯一可能的确定性合并方式，因此只有随机合并才能改进合并效果。我们的分析还改进了 Pagh 和 Stausholm（2021）的草图，大幅减少了实现 $\varepsilon$-差分隐私保证所需的噪声，并展示了如何合并具有不同隐私预算的草图。

**定理 4.1**。设 $f_1 = \mathcal{F}_{p_1,\ q_1}, f_2 = \mathcal{F}_{p_2,\ q_2}, f_3 = \mathcal{F}_{p_3,\ q_3}$，并令 $\circ\colon\ \{0,1\}^2 \to \{0,1\}$ 表示一个确定性对称运算。以下条件只有在 $\circ = \veebar$ 且 $p_1 = p_2 = 1/2$时才能同时满足：

1. $f_1, f_2$ *are* $\varepsilon_1$*-DP and* $\varepsilon_2$*-DP for* $\varepsilon_1, \varepsilon_2 < \infty$.
2. $f_1(x) \circ f_2(y) \overset{D}{=} f_3(x \vee y)$.
3. $f_i(0) \overset{D}{\neq} f_i(1)$ *for* $i = 1, 2, 3$.

利用我们的一般随机响应机制族，我们定义了一种机制，通过向 PCSA 草图添加噪声来实现隐私保护（引理 4.3），同时保留可合并性（定理 4.4）。推论 4.5 表明，我们的隐私分析比 Pagh 和 Stausholm（2021）的分析严格得多。

**定义 4.2**。对于 $\varepsilon > 0$，令 $\mathcal{M}_\varepsilon^{\mathrm{xor}}\colon\ \{0,1\}^d \to \{0,1\}^d$ 表示这样的机制：对其输入中的每个元素独立应用非对称随机响应 $\mathcal{F}_{p,\ q}$ 其中 $p = 1/2,\ \ q = 1/(2e^\varepsilon)$。

**引理 4.3**。 $\mathcal{M}_{\varepsilon}^{\mathrm{xor}}$是 $\varepsilon$–差分隐私的。

**定理 4.4**。 $\mathcal{M}_{\varepsilon_1}^{\mathrm{xor}}(x) \veebar \mathcal{M}_{\varepsilon_2}^{\mathrm{xor}}(y) \overset{D}{=} \mathcal{M}_{\varepsilon^*}^{\mathrm{xor}}(x \vee y)$，其中 $\varepsilon^* = -\log(e^{-\varepsilon_1} + e^{-\varepsilon_2} - e^{-(\varepsilon_1+\varepsilon_2)})$。

**推论 4.5**。 令 $\mathcal{M}_{\varepsilon}^{\mathrm{PS}}$表示 *Pagh* 和 *Stausholm*（*2021*年，第 *6* 节）的 $\varepsilon$-差分隐私机制。那么 $\mathcal{M}_{\varepsilon}^{\mathrm{xor}} = \mathcal{M}_{2(\exp(\varepsilon)-1)}^{\mathrm{PS}}$。

这项更紧密的隐私分析[1] 大幅减少了为实现隐私保证而添加的噪声，实际上将隐私预算至少提高了 2 倍。从实际角度来看，图 4 显示，当 $\varepsilon \to 0$时，误差呈指数级增长。

## 4.2 随机合并

定理 4.1 表明，只有将 PCSA 草图中的 1 位随机化为伯努利分布$(1/2)$取值，才能进行确定性合并。因此，即使隐私预算近乎无穷大，可合并差分隐私草图仍必须向基础 PCSA 草图添加大量噪声。我们将随机性从基础草图转移到合并过程中，证明这样可以在使用标准随机响应机制（定义 4.6）的同时，降低总体噪声。

**定义 4.6**。 对于 $\varepsilon > 0$，我们将 $\mathcal{M}_{\varepsilon}^{\mathrm{sym}}$ 表示为：$\{0,1\}^d \to \{0,1\}^d$ 一种机制，该机制使用 $\mathcal{F}_{p,1-p}$ 对其输入中的每个元素独立应用标准随机响应机制，并使用 $p = e^{\varepsilon}/(e^{\varepsilon}+1)$。

**引理 4.7**。 $\mathcal{M}_{\varepsilon}^{\mathrm{sym}}$是 $\varepsilon$–差分隐私的。

合并是一种随机化算法 $g_{\varepsilon_1,\varepsilon_2}$： $\{0,1\}^2 \to \{0,1\}$ 它以如下意义与 $\vee$ 可交换：

$$g_{\varepsilon_1,\varepsilon_2}(\mathcal{M}_{\varepsilon_1}^{\mathrm{sym}}(x), \mathcal{M}_{\varepsilon_2}^{\mathrm{sym}}(y)) \overset{D}{=} \mathcal{M}_{\varepsilon^*}^{\mathrm{sym}}(x \vee y).$$

由于 $g_{\varepsilon_1,\varepsilon_2}$ 是一个将位对随机映射为单个位的映射，我们可以将其表示为一个 $4 \times 2$马尔可夫转移矩阵。有效的合并操作是所得矩阵等式的解，其中 $\varepsilon^*$ 是自由参数。我们通过求解能够生成有效解的最大 $\varepsilon^*$，为 $\mathcal{M}^{\mathrm{sym}}$ 获得最优的随机化合并操作，其形式如下：

**定理 4.8**。 假设 $\varepsilon_1, \varepsilon_2 > 0$。令 $q(\varepsilon) = (e^{\varepsilon}+1)^{-1}$,

$$\varepsilon^* = -\log(e^{-\varepsilon_1} + e^{-\varepsilon_2} - e^{-(\varepsilon_1+\varepsilon_2)}),\ q^* = q(\varepsilon^*),$$

$$K_i = \begin{bmatrix} 1-q(\varepsilon_i) & q(\varepsilon_i) \\ q(\varepsilon_i) & 1-q(\varepsilon_i) \end{bmatrix} \textit{ for } i \in \{1,2\}, \textit{ and}$$

$$v^* = (q^*, 1-q^*, 1-q^*, 1-q^*)^T,$$

令 $\otimes$ 表示克罗内克积，定义：

$$(t_{00}, t_{01}, t_{10}, t_{11})^T = (K_1^{-1} \otimes K_2^{-1})\, v^*, \textit{ and}$$

$$g_{\varepsilon_1,\varepsilon_2}(a,b) \sim \mathrm{Bernoulli}(t_{ab}), \quad a,b \in \{0,1\}.$$

然后 $g_{\varepsilon_1,\ \varepsilon_2}(\mathcal{M}_{\varepsilon_1}^{\mathrm{sym}}(x)\ \ \mathcal{M}_{\varepsilon_2}^{\mathrm{sym}}(y)) \overset{D}{=} \mathcal{M}_{\varepsilon^*}^{\mathrm{sym}}(x \vee y)$，其中 $g$ 按位且相互独立地取值 .

[1] Pagh 和 Stausholm（2021）的附录证明了 $q = 1/(e^{\varepsilon}+1)$ 满足 $\varepsilon$-差分隐私，但论文中的建议和主要结果依赖于选择 $q = 1/(2+\varepsilon)$。我们推荐的 $q = 1/(2e^{\varepsilon})$ 在差分隐私约束下是最优的。

当原始向量 $x$ 和 $y$ 与合并向量同时可见时，定理 4.4 和定理 4.8 中的 $\varepsilon^*$ 参数最好理解为衡量合并草图效用的指标，而非隐私预算，因为根据差分隐私的后处理不变性（Dwork 等，2006），发布合并向量不会造成额外的隐私泄露。因此，我们寻求合并时的最大 $\varepsilon^*$。注意到定理 4.4 和定理 4.8 得出的 $\varepsilon^*$ 相同，且 $\mathcal{M}_\varepsilon^{\mathrm{sym}}$的噪声小于 $\mathcal{M}_\varepsilon^{\mathrm{xor}}$，所以合并后 $\mathcal{M}^{\mathrm{sym}}$ 仍是首选机制。

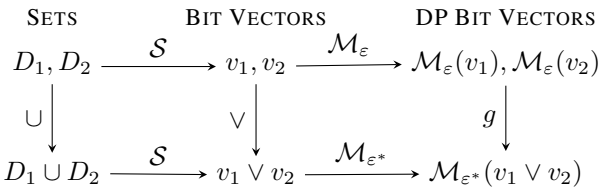
注释 4.9。 根据归纳法，定理 4.4 和定理 4.8 中规定的合并操作允许合并任意长度相同的 $k$ 位向量 $x_1, \ldots, x_k$，这些向量使用 $\varepsilon_1, \ldots, \varepsilon_k$进行了隐私处理，合并后得到向量 $v = (x_1 \vee \cdots \vee x_k)$，该向量经过隐私处理，使用

$$\varepsilon^* = -\log\left(1 - \prod_{i=1}^{k}(1 - e^{-\varepsilon_i})\right).$$

一个自然的问题是，是否存在一种随机化合并算法 $g_{\varepsilon_1,\ldots,\varepsilon_k}\colon \{0,1\}^k \to \{0,1\}$，能够满足类似定理 4.8 的性质，且其 $\varepsilon^*$ 大于通过对成对合并进行归纳所得到的值。在附录 A 中，我们证明了定理 4.8 的一个更一般形式（定理 A.2），并对此问题给出了否定答案。

## 4.3 一般布尔运算

我们暂时将关注点从不同元素计数转移开来，介绍随机响应下布尔运算的一种推广，这可能具有基础性意义，例如可用于设计交集运算。在不同计数草图中，集合并集对应于按位或 运算，而隐私带来的挑战是如何在含噪比特上执行等效运算。PCSA 与其他可合并草图一样，定义了一个从多重集到草图的同态 $\mathcal{S}$。下方的交换图展示了这种可合并性，因为从 $D_1, D_2$ 到 $v_1 \vee v_2$，选择哪条路径并无影响。同样，定理 4.8 中的合并操作 $g$ 确保隐私机制 $\mathcal{M}$ 使该图交换。通过保留并集运算的结构，可以根据合并后的隐私草图推断并集的基数。




在这里，我们将 $\mathcal{M}^{\mathrm{sym}}$ 下的或 ($\vee$) 合并推广到任意逻辑运算。特别地，推论 4.10 表明，只需对目标概率稍作调整 $v^*$，即可得到适用于与 ($\wedge$) 的随机合并；而附录 B 中的引理 B.1 和 B.2 展示了适用于异或 ($\veebar$) 的合并，并说明非 ($\neg$) 与 $\mathcal{M}^{\mathrm{sym}}$可交换。

**推论 4.10**。 假设定理 *4.8* 中的设定成立，但令 $v^* = (q^*, q^*, q^*, 1-q^*)^T$。那么$g_{\varepsilon_1,\varepsilon_2}(\mathcal{M}_{\varepsilon_1}^{\mathrm{sym}}(x)\ \mathcal{M}_{\varepsilon_2}^{\mathrm{sym}}(y)) \overset{D}{=} \mathcal{M}_{\varepsilon^*}^{\mathrm{sym}}(x \wedge y)$，其中 $g$ 按位且相互独立地取值。

形式上，对于每个二元逻辑运算符 $\square$，都有一个函数 $\varepsilon_\square^*(\varepsilon_1, \varepsilon_2)$ 用于合并两个隐私预算，并赋予比特向量与隐私预算对以半群结构 $(v_1, \varepsilon_1) \cdot (v_2, \varepsilon_2) := (v_1 \square v_2, \varepsilon_\square^*(\varepsilon_1, \varepsilon_2))$。隐私机制 $\mathcal{M}^{\mathrm{sym}}$ 随后定义了一个映射

表 1： $\mathcal{M}^{\mathrm{sym}}$下位向量上的布尔运算

| 运算（□） | 差分隐私运算。 | $\varepsilon^*_{\square}(\varepsilon_1,\ \varepsilon_2)$ |
|---|---|---|
| $\neg$ | 引理 B.2 | — |
| $\vee$ | 定理 4.8 | $-\log(e^{-\varepsilon_1}+e^{-\varepsilon_2}-e^{-(\varepsilon_1+\varepsilon_2)})$ |
| $\wedge$ | 推论 4.10 | $-\log(e^{-\varepsilon_1}+e^{-\varepsilon_2}-e^{-(\varepsilon_1+\varepsilon_2)})$ |
| $\veebar$ | 引理 B.1 | $\log\left(\frac{1+e^{\varepsilon_1+\varepsilon_2}}{e^{\varepsilon_1}+e^{\varepsilon_2}}\right)$ |

$\phi(v,\varepsilon) := (\mathcal{M}^{\mathrm{sym}}_{\varepsilon}(v)\ \ \varepsilon)$，它是从该半群到其带噪对应结构的同态。这些运算汇总于表 1。

我们的随机化合并技术对于支持一般布尔运算至关重要。与我们的随机化合并操作不同，推论 4.11 表明，在相同的随机响应机制下，不存在能够同时为或（$\vee$）和与（$\wedge$）定义合并操作的确定性操作 $\circ$ 和 $\bullet$。

**推论 4.11**。 假设处于定理 *4.1* 的设置下。令 $\bullet$： $\{0,1\}^2 \to \{0,1\}$ 表示一种确定性对称运算。除了以下条件外，不可能同时满足定理 *4.1* 的条件 *(1)–(3)*：

4. $f_1(x) \bullet f_2(y) \overset{D}{=} f_3(x \wedge y)$.

# 5 基数估计

迄今开发的 Sketch-Flip-Merge 方法满足隐私与可合并性要求，但仍需说明如何使用 SFM 摘要来估计基数。我们构建了一种基于复合似然的估计器，该估计器具有一致性且渐近最优。我们给出了一种误差的解析估计器，在我们的实验中，它与真实误差非常接近。

基于似然的基数估计方法已在非隐私设置中得到应用（Clifford 与 Cosma，2012 年；Lang，2017年；Ertl，2017年；Ting，2019年），并且相较于 PCSA 的其他估计器展现出更高的准确度。尽管给定草图后，对$n$ 进行真正的最大似然估计由于草图中各位之间不具备独立性而在计算上不可行（Ting，2019年；Ertl，2017年），但任意一位的边际似然都很容易推导。与 Ting（2019）类似，我们针对 $n$推导出一种复合边际似然估计器（Lindsay，1988 年；Varin 等人，2011 年）。

令 $C_{ij}$ 表示草图中映射到桶 $i$、值为 $j$ 的唯一项 $x \in D$ 的数量，$\rho_{ij}$ 表示一个项被映射到该位置的概率。那么，在假设使用通用随机哈希的情况下，以下生成过程描述了 SFM 摘要$T = \mathcal{M}_{p,q}(\mathcal{S}(D))$.

$$(C_{11},\ldots,C_{BP}) \sim \mathrm{Multinomial}(n,\rho_{11},\ldots,\rho_{BP})$$
$$T_{ij} \mid C_{ij} \sim \mathcal{F}_{p,q}(1(C_{ij}>0))$$

虽然 $\{C_{ij}\}_{i,j}$ 的联合分布，以及由此得到的观测值 $\{T_{ij}\}_{i,j}$，涉及对 $n$ 分成至多 $BP$ 个部分的所有整数分拆求和，难以计算，但单个位 $T_{ij}$ 的边缘分布很容易求得。注意，单元格 $C_{ij}$ 的占据概率 $\rho_{ij}$ 仅取决于其层级 $j$，其中 $\rho_{ij} = 2^{-\min\{j,P-1\}}/B$。于是 $\Pr(C_{ij}=0 \mid n) = \gamma_j^n$，其中 $\gamma_j = 1-\rho_{ij}$，并且

$$T_{ij} \sim \mathrm{Bernoulli}\left(p(1-\gamma_j^n) + q\gamma_j^n\right).$$

复合边际对数似然（Lindsay，1988 年；Varin 等人，2011 年）用替代量 $\ell_{p,q}(n;t)$ 取代真实对数似然，该替代量对边际对数概率求和，

$$\ell_{p,q}(n;t)=\sum_{ij}(1-t_{ij})\log\left(1-p+(p-q)\gamma_j^n\right)+\sum_{ij}t_{ij}\log\left(p-(p-q)\gamma_j^n\right),$$

其中，$t=\mathcal{M}_{p,q}(\mathcal{S}(D))$表示一个已实现的 SFM 摘要。相应的复合极大似然估计量为$\hat{n}=\max_n \ell_{p,q}(n;t)$，可通过牛顿法进行优化。$\ell_{p,q}$ 所需的一阶和二阶导数为

$$\ell'_{p,q}(n;t)=\sum_{ij}(1-t_{ij})\frac{(p-q)\gamma_j^n\log(\gamma_j)}{1-p+(p-q)\gamma_j^n}-\sum_{ij}t_{ij}\frac{(p-q)\gamma_j^n\log(\gamma_j)}{p-(p-q)\gamma_j^n},$$
$$\ell''_{p,q}(n;t)=\sum_{ij}(1-t_{ij})\frac{(1-p)(p-q)(\log\gamma_j)^2\gamma_j^n}{(1-p+(p-q)\gamma_j^n)^2}-\sum_{ij}t_{ij}\frac{p(p-q)(\log\gamma_j)^2\gamma_j^n}{(p-(p-q)\gamma_j^n)^2}.$$

在不存在隐私的情况下（即，$p=1,q=0$），$\ell_{p,q}(\cdot;t)$是严格凹函数。虽然在隐私情形下并非如此，但定理 5.1 指出，对于某个 $\Delta>0$，$\ell_{p,q}$ 的期望在区间$(0,n+\Delta)$上仍然是凹函数。

**定理 5.1**。设 $D$ 为一个多重集，使得 $|set(D)|=n$。设$T=\mathcal{M}_{p,q}(\mathcal{S}(D))$。设 $f(\hat{n})=\mathbb{E}[\ell_{p,q}(\hat{n};T)]$，，其中期望是对哈希函数 $h_1$、$h_2$ 以及隐私机制$\mathcal{M}_{p,q}$的随机性取的。则 $f(\hat{n})$ 在$\hat{n}=n$ 处取得全局最大值，并且在一个区间上为凹函数，该区间包含$(0,n]$ 且其内部包含该点。

## 5.1 理论结果

我们选择使用复合边际似然，因为它具有良好的理论性质。具体而言，即使真实似然并不完整，使用真实似然也能确保基数估计渐近一致，而复合似然的海森矩阵可用于估计方差（Ting，2019年）。我们进一步证明，在基数相对于草图大小较大的典型情况下，基数估计渐近最优。

**定理 5.2.** 令 $S_n$ 表示包含 $n$ 个不同项的 *PCSA* 摘要，其中有 $B(n)$ 个桶和 $P=\infty^{\sim}$ 个层级。令 $S_n$ 表示包含泊松$(n)$ 个不同项的修正 *PCSA* 摘要，并令 $S_n$ 表示复合边际似然就是真实似然的修正 *PCSA* 摘要。若 $B\log B=o(\sqrt{n})$，则存在 $^{\sim}$ 个修正 *PCSA* 摘要 $S_n$、$S_n$，使得

$$\Pr(S_n=\tilde{S}_n=\overline{S}_n)\to 1\quad as\ n\to\infty.$$

**推论 5.3.** *SFM* 草图的基数复合似然估计量，在定理 *5.2* 的渐近条件下渐近有效。

我们概述证明，并在附录 A 的 A.5 节中给出详细内容。主要难点在于，SFM 摘要中的条目彼此相关，因为每个项目只能分配到一个单元格。通过构造具有相关条目和独立条目的草图 $S_n$，$S_n$ 之间的耦合，我们证明它们渐近相等。在这些耦合过程中，项目分配差异最大的桶，其差异最多只有一小部分，$O_p(\sqrt{n}/B\log B)$。通过证明一个项目以概率 $O(1/v_i)$ 更新其所在桶的草图值（其中 $v_i$ 是桶$i$中的项目数），我们得出结论：当平均桶分配量 $v_i\approx n/B$ 增长得足够快，以消除项目分配上的微小差异时，耦合后的草图实际上以趋近于 1 的概率完全相同

˜无关。由于¯$S_n$， $S_n$ 具有独立的位，我们通过逆累积分布函数法将它们耦合，并直接界定它们不同的概率。由于这些草图渐近相同，对它们施加完全相同的随机响应噪声，意味着它们隐私化后的版本也相同，并且对它们使用的任何估计器都具有相同的渐近抽样分布。因此，独立条目草图的渐近有效最大似然估计（即复合似然估计量）对于真实的 SFM 摘要 $S_n$也渐近有效。

备注 5.4。 该结果还证明，在假设每个桶包含泊松($n/B$)个项的近似条件下推导出的最大似然估计是渐近有效的。该结论还可推广到 HyperLogLog 和其他草图，从而表明具有良好实证特性的基于伪似然的估计量（Ertl，2017年；Ting，2019年）实际上是渐近最优的。

**误差估计**。 与最大似然估计的费歇尔信息矩阵一样，戈丹贝信息（或夹心信息）的逆矩阵可为估计器的方差提供一致估计。戈丹贝信息为 $G(n) = H(n)J^{-1}(n)H(n)$，其中 $-J(n)$ 是在 $n$ 处期望复合似然对数的海森矩阵，而 $H(n) = \mathrm{Var}(\ell'_{p,\,q}(n,\ T))$ 是复合得分函数的方差。在非隐私情形下，Ting（2019年）证明，基于费歇尔信息和戈丹贝信息得到的 HyperLogLog 复合边际似然方差估计，在大基数时几乎相同；而由于桶之间存在负相关，费歇尔信息会在小基数时高估方差。图 2 表明，加入独立的随机响应噪声后，这种高估现象会显著减弱。因此，我们仅使用海森矩阵来定义估计标准误，即

$$\begin{aligned}\widehat{\mathrm{SE}}_{p,q}(B,P,n) &= \sqrt{1/\mathbb{E}[-\ell''_{p,q}(n;T)]} \\ &= \left[B(p-q)\sum_{j=1}^{P}(\log\gamma_j)^2\gamma_j^n\left(\tfrac{p}{p-(p-q)\gamma_j^n}-\tfrac{1-p}{1-p+(p-q)\gamma_j^n}\right)\right]^{-1/2},\end{aligned} \tag{2}$$

其中，对于大小为 $B \times P$、基数为 $n$ 的随机 PCSA 草图 $S$，$T = \mathcal{M}_{p,q}(S)$。第 6 节中的图 4 通过实验证明，我们的误差估计 $\mathrm{SE}_{p,q}(B,P,n)$ 能够很好地近似误差。

# 6 评估

我们在真实世界数据集和合成数据集上评估了所提出的方法。实验结果表明，SFM 摘要是首个具有实用性能、可合并的 $\varepsilon$-差分隐私不同元素计数草图，因为 Pagh 和 Stausholm（2021）提出的草图误差大得不切实际。在隐私保护草图中，我们新颖的随机合并草图构造优于确定性合并草图。因此，我们在构造、估计和隐私分析方面的改进带来了实际收益。此外，我们的理论误差与经验误差非常接近。

## 6.1 实验设置

我们在实验中考虑四种不同的私有不同元素计数草图。在我们的方法中，*SFM*（对称）将 $\mathcal{M}_\varepsilon^{\mathrm{sym}}$ 与我们的随机合并过程配对，而 *SFM*（异或）将 $\mathcal{M}_\varepsilon^{\mathrm{xor}}$ 与确定性的 异或 合并配对。两种 SFM 方法都使用第 5 节中的估计器。我们将这些方法与 Pagh 和 Stausholm（2021）的草图和估计器进行比较，该方法有两种实现方式：*PS*（宽松）按照 Pagh 和 Stausholm（2021）的规定，使用 $\mathcal{M}_\varepsilon^{\mathrm{PS}} = \mathcal{M}_{p,q}$ 和 $p = 1/2, q = 1/(2+\varepsilon)$ 构造草图。

而 *PS*（严格）使用收紧后的 $\mathcal{M}_{\varepsilon}^{\mathrm{xor}}$（定义 4.2）。根据推论 4.5，*PS*（严格）在 $\varepsilon$ 处的草图构造等价于 *PS*（宽松）在 $\varepsilon' = 2(e^{\varepsilon} - 1)$ 处的草图构造。

在最后一组实验中，我们还将两种非隐私草图作为基准进行比较，并指出，我们不应期待隐私草图的准确度能与非隐私草图相当。下文中，*FM85* 表示使用 Flajolet 和 Martin（1985）估计器的非隐私 PCSA，而*HLL* 表示 HyperLogLog（Flajolet 等，2007）。我们在桶数相同的条件下比较 HLL 草图与 PCSA 草图，并指出每个桶所占用的空间，HLL 草图比对应的 PCSA 草图更小。

我们主要以相对均方根误差（*RRMSE*）的形式衡量估计误差，其定义为

$$\mathrm{RRMSE}(\hat{n}_1, \ldots, \hat{n}_m; n) = \frac{1}{n}\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{n}_i - n)^2}.$$

我们还通过两种方法的均方误差之比来衡量它们的相对效率，

$$\mathrm{RE}(\hat{n}^{(1)}, \hat{n}^{(2)}) = \frac{\mathrm{MSE}(\hat{n}^{(2)})}{\mathrm{MSE}(\hat{n}^{(1)})}.$$

如果两个草图的估计器都是无偏的，那么相对效率为 $r$ 表示效率较低的草图渐近而言必须使用 $r$ 倍的桶数，才能达到相同的准确度。这是因为对于这些估计器，渐近均方误差（方差）与 $1/B$ 成比例地下降。

模拟默认使用维度为 $B = 4096$、$P = 24$ 的草图，并采用 xxHash6 4 (Co哈希函数（llet，2022年），对 $m = 1000$ 次试验取平均。

**对 Pagh 和 Stausholm（2021）的修改。** 在我们的实验中，他们原始的估计器经常无法给出估计值。对于目标误差容限 $\beta$，该方法会计算草图所有 $P$ 个层级的区间，并取这些区间的交集以得出估计值。当 $\beta$ 较小时，这些区间的交集经常为空。为解决这一问题，我们搜索能产生非空交集的最小 $\beta$。我们使用该区间的中点来估计 $n$。

**数据源。** 我们的实验使用合成数据和真实数据。合成数据由基数固定的随机整数集合构成。真实数据取自 BitcoinHeist 论文（Akcora 等人，2020年），该论文提供了一个包含 $n = 2{,}631{,}095$ 个唯一地址的比特币交易数据库。该数据集可在 UCI 仓库获取，采用 CC BY 4.0 许可证。值得注意的是，不同计数草图对输入的值分布并不敏感，因为处理过程中会对值进行哈希。

## 6.2 结果

图 1 比较了四种隐私方法在合成数据上的准确度，其中基数 $n$ 从 $10^2$ 到 $10^6$ 变化，隐私预算固定为 $\varepsilon = 1$。对于大基数，各方法的相对均方根误差趋于一个固定常数。即使我们收紧了 Pagh 和 Stausholm（2021年）方法的隐私分析，我们的方法的误差仍比其低一个数量级。对于小基数，相对误差会随着基数减小而增大，这符合差分隐私方法的预期。图 2 比较了 $\varepsilon$ 取多个值时的准确度，但仅针对表现最好的草图 *SFM*（对称）。图中还显示，无论 $\varepsilon$ 如何选择，相对均方根误差都会随着 $n \to \infty$ 而稳定下来。与差分隐私方法不同，具有无限隐私预算的 SFM 摘要（等同于采用 Lang（2017年）MDL 估计量的 PCSA）在小 $n$ 时能给出特别准确的估计。图 2 还 $\widehat{\ }$显示，估计的相对误差 SE/$n$ （公式 2 ）是经验误差的上界，并且能很好地估计相对均方根误差，尤其是在基数较大或 $\varepsilon$ 较小时。

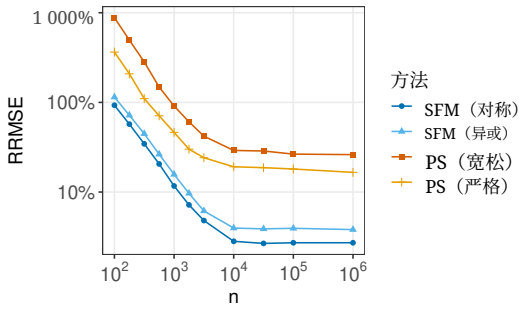



图 1：在双对数坐标轴上比较四种方法在 $n$ 处、 $\varepsilon = 1$ 时的相对均方根误差。对于较大的 $n$，相对均方根误差趋于稳定。

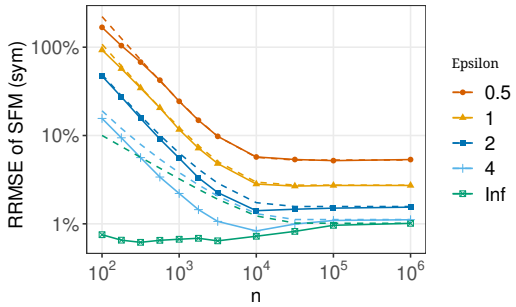



图 2：虚线表示估计的相对误差 SE/$n$，当 $n$ 较大且 $\varepsilon$较小时，估计非常准确。

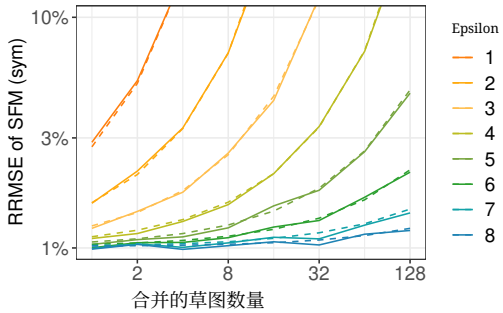



图 3：合并给定数量的*SFM*（对称）摘要后，在 $n = 10^6$ 处的相对均方根误差；每个摘要的隐私预算均为 $\varepsilon$。虚线表示估计的相对误差。

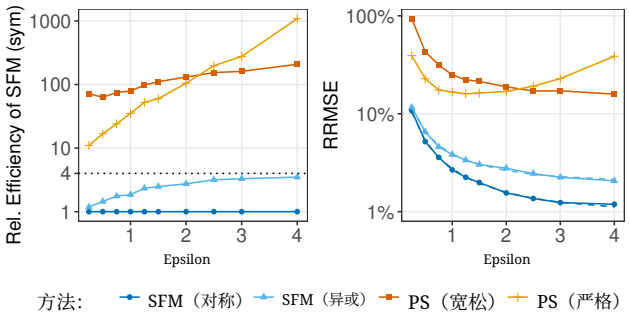



图 4: （左）BitcoinHeist（$n \approx 2.6$M）数据集上，*SFM*（对称）与其他方法的相对效率。SFM 的效率远高于 PS。对于较大的隐私预算，*SFM*（对称）的效率是*SFM*（异或）的 4 倍。（右）相对于 $\varepsilon$ 的相对均方根误差。对于 SFM，虚线表示估计的相对误差，与真实误差相匹配。

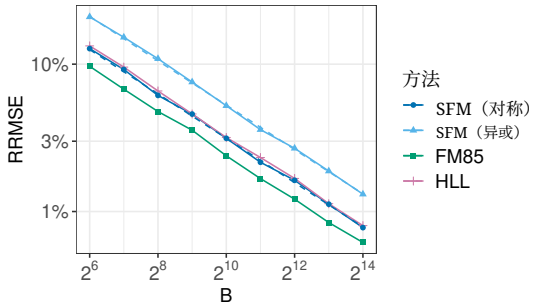



图 5：隐私预算为 $n = 10^6$ 时，隐私保护草图在 $\varepsilon = 2$ 下的相对均方根误差与桶数 $B$ 的关系，并与常见的非隐私替代方案进行比较（采用双对数坐标轴）。对于 SFM，虚线表示估计的相对误差。对于所有方法，相对均方根误差均随 $B^{-0.5}$ 变化。

图 3 展示了在基数较大时（$n = 10^6$）SFM 摘要中合并与隐私之间的权衡。合并操作会导致噪声累积，因此需要使用更大的隐私预算来应对更多次合并。此处的估计相对误差根据备注 4.9 计算，再次与经验误差高度吻合。

我们还使用不同隐私预算 $\varepsilon$（范围从 0.25 到 4），将隐私方法与真实世界的 BitcoinHeist 数据进行比较。图 4 左侧面板展示了 *SFM*（对称） 相较于其他隐私方法的相对效率。*SFM*（对称） 的效率始终比 PS估计器至少高一个数量级。与*PS*（宽松）相比，$100\times$ 更高的效率意味着*SFM*（对称）只需使用 1% 的空间即可达到相同的误差。此外，*SFM*（对称）优于*SFM*（异或），且在较大的 $\varepsilon$ 下，相对效率趋近于 4。这表明，对于较大的隐私预算，随机合并*SFM*（对称）只需四分之一的空间，就能达到与*SFM*（异或）相当的准确度。右侧面板以绝对值展示了本次实验中的估计误差；我们再次看到，SFM 的估计相对误差能够很好地近似相对均方根误差（RRMSE）。

最后，我们将 SFM 与流行的非隐私替代方案进行比较，并展示误差同样会随桶数量增加而降低。我们使用基数为 $n = 10^6$ 的合成数据，构建桶数量各异的草图 $B$，并为 SFM 使用隐私预算 $\varepsilon = 2$。图 5 展示了

各方法的 RRMSE 随 $B$ 变化的情况。与常见的非隐私不同元素计数草图一样，我们的 RRMSE 会随着 $B^{-0.5}$ 增加而降低。因此，与非隐私草图类似，在大基数情况下，我们的差分隐私摘要的 RRMSE 可以通过一个简单公式 $c_\varepsilon/\sqrt{B}$ 轻松表征，其中 $c_\varepsilon$ 是特定于某种方法和隐私预算的常数。

# 7 讨论与结论

Sketch-Flip-Merge 摘要相较于当前最先进的可合并差分隐私基数计数草图，展现出显著的改进。这得益于新颖的合并算法（定理 4.8 和第 4.3 节）、渐近最优估计（第 5 节）以及改进的隐私分析（推论 4.5）。

可合并的隐私摘要面临一个重要局限：隐私与可合并性之间存在内在的张力。虽然两者都可以实现，但在隐私场景中反复合并会降低准确度。Desfontaines 等人（2019年）在一般基数计数场景中论述了这种权衡，而备注 4.9 和图 3 则对其在 SFM 中的具体情况进行了量化。

最后，我们指出，部分研究发现具有普遍性。具体而言，我们用于聚合含噪二进制数据的方法提供了基础工具，并量化了随机响应下位运算所产生的噪声累积效应；这些方法和量化结果适用于广泛的问题，尤其是隐私保护领域。

# 参考文献

Cuneyt G. Akcora、Yitao Li、Yulia R. Gel 与 Murat Kantarcioglu。Bitcoinheist：用于比特币区块链勒索软件预测的拓扑数据分析。载于 第二十九届国际人工智能联合会议论文集，*IJCAI-20*，第 4439–4445 页，2020年7月。金融科技人工智能特别专题。Ziv Bar-Yossef、TS Jayram、Ravi Kumar、D Sivakumar 与 Luca Trevisan。数据流中不同元素的计数。载于 国际计算机科学随机化与近似技术研讨会，第 1–10 页。施普林格，2002年。Andrei Broder 与 Michael Mitzenmacher。布隆过滤器的网络应用：综述。*Internet Mathematics*，1(4)：485–509，2004年。Seung Geol Choi、Dana Dachman-Soled、Mukul Kulkarni 与 Arkady Yerukhimovich。用于大规模统计的差分隐私多方草图。密码学电子印刷档案，2020年。Peter Clifford 与 Ioana A Cosma。概率计数算法的统计分析。斯堪的纳维亚统计学杂志，39(1)：1–14，2012年。Yann Collet。xxHash——极速非密码学哈希算法。https://cyan4973. github.io/xxHash/，2022年。访问日期：2022年11月16日。Graham Cormode 与 Ke Yi。大数据小型摘要。剑桥大学出版社，2020年。Damien Desfontaines、Andreas Lochbihler 与 David Basin。基数估计器无法保护隐私。隐私增强技术论文集，2019(2)：26–46，2019年4月。doi：10.2478/popets-2019-0018。Charlie Dickens、Justin Thaler 与 Daniel Ting。（几乎）所有基数估计器都是差分隐私的。*arXiv*预印本 *arXiv:2203.15400*，2022年。Marianne Durand 与 Philippe Flajolet。大基数的 LogLog 计数。载于 欧洲算法研讨会，第 605–617 页。施普林格，2003年。Cynthia Dwork。差分隐私：研究成果综述。载于 国际计算模型理论与应用会议，第 1–19 页。施普林格，2008年。Cynthia Dwork、Frank McSherry、Kobbi Nissim 与 Adam Smith。为隐私数据分析校准噪声与敏感度。载于 密码学理论会议，第 265–284 页。施普林格，2006 年。Ú lfar Erlingsson、Vasyl Pihur 与 Aleksandra Korolova。RAPPOR：可聚合的随机响应隐私保护有序响应。载于 *ACM SIGSAC*计算机与通信安全会议，第 1054–1067 页。ACM，2014年。doi：10.1145/2660267.2660348。URL https://doi. org/10.1145/2660267.2660348。Otmar Ertl。HyperLogLog草图的新基数估计算法。*arXiv*预印本 *arXiv:1702.01284*，2017年。Philippe Flajolet 与 G Nigel Martin。用于数据库应用的概率计数算法。计算机与系统科学杂志，31(2)：182–209，1985年。

´Philippe Flajolet、Eric Fusy、Olivier Gandouet 和 Fr´ed´eric Meunier。《Hyperloglog：近乎最优基数估计算法的分析》。载于离散数学与理论计算机科学，第137–156页。离散数学与理论计算机科学，2007年。Fr´ed´eric Giroire。次序统计量与海量数据集基数估计。离散应用数学，157(2):406–427，2009年。Stefan Heule、Marc Nunkesser 和 Alexander Hall。《Hyperloglog 实践：最先进基数估计算法的算法工程》。载于第*16*届国际数据库技术拓展会议论文集，第683–692页，2013年。˘ V.F. Kolchin、B.A. Sevast'janov、V.P. Christ'Yakov、V.P. Cistjakov 和 A.V. Balakrishnan。《随机分配》。Halsted Press 出版的一本书。V. H. Winston，1978。国际标准书号9780470993941。网址 https: //books.google.com/books?id=jxjvAAAAMAAJ。Benjamin Kreuter、Craig William Wright、Evgeny Sergeevich Skvortsov、Raimundo Mirisola 和 Yao Wang。隐私保护的安全基数与频率估计。2020年。Kevin J Lang。回到未来：一种更接近最优的基数估计算法。*arXiv* 预印本 *arXiv:1708.06839*，2017年。Bruce G Lindsay。复合似然方法。当代数学，80(1):221–239，1988 年。Saskia Nu˜nez von Voigt 和 Florian Tschorsch。RRTxFM：用于差分隐私统计的概率计数。载于电子商务、电子服务与电子社会会议，第86–98页。施普林格，2019年。Rasmus Pagh 和 Nina Mesing Stausholm。高效的差分隐私 $F_0$ 线性草图。载于第*24*届国际数据库理论会议（*ICDT 2021*）。Schloss Dagstuhl-Leibniz-Zentrum f¨ur Informatik，2021年。Adam Smith、Shuang Song 和 Abhradeep Guha Thakurta。Flajolet-Martin 草图本身即可保护差分隐私：以最小空间实现隐私计数。神经信息处理系统进展，33:19561–19572，2020年。Rade Stanojevic、Mohamed Nabeel 和 Ting Yu。通过差分隐私分布式估计集合运算的基数。载于*2017*年 *IEEE* 隐私感知计算研讨会（*PAC*），第37–48页。IEEE，2017年。Daniel Ting。面向数十亿数据集的近似去重计数。载于*2019*年国际数据管理会议论文集，第69–86页，2019年。Florian Tschorsch 和 Bj¨orn Scheuermann。一种隐私保护的分布式用户统计算法。计算机网络，57(14):2775–2787，2013年。Cristiano Varin、Nancy Reid 和 David Firth。复合似然方法概述。《中华统计学志》，第5–42页，2011 年。Stanley L Warner。随机响应：一种消除回避作答偏差的调查技术。美国统计学会会刊，60(309):63–69，1965年。Kyu-Young Whang、Brad T Vander-Zanden 和 Howard M Taylor。一种用于数据库应用的线性时间概率计数算法。*ACM*数据库系统汇刊（*TODS*），15 (2):208–229，1990年。

# A 主要结果的证明

## A.1 补充记号

在接下来的证明中，我们使用以下记号。 $\mathbf{1}_k$ 是由 $k$ 个 1 组成的列向量， $I_k$ 是 $k \times k$ 单位矩阵（维度明确时省略下标）。我们用 $e_i$ 表示第$i$个标准基向量，即除位置 $i$处的元素为 1 外，其余元素均为 0 的向量。我们用 $\langle \cdot, \cdot \rangle$ 表示两个向量的内积。

## A.2 第 3 节中结果的证明

定理 *3.2* 的证明。证明思路与 Erlingsson 等人（2014年，定理 2）类似。我们首先证明，当且仅当 $\mathcal{F}_{p,\ q}$ 是 $\varepsilon$-差分隐私时， $p \in (0,\ 1]$ 且 $\max\{p/q, (1-q)/(1-p)\} \le e^{\varepsilon}$。为满足这一条件，我们必须满足

$$\Pr(\mathcal{F}_{p,q}(x) = y) \le e^{\varepsilon}\ \Pr(\mathcal{F}_{p,q}(1-x) = y)$$

对所有 $x,\ y \in \{0,1\}$。若 $p = q$, $\mathcal{F}_{p,\ p}$ 是一种忽略输入并随机输出一个比特值的机制，则该条件显然成立。假设 $p \neq q$，当

$$\Pr(\mathcal{F}_{p,q}(1-x) = y) > 0 \quad \text{and} \quad \frac{\Pr(\mathcal{F}_{p,q}(x) = y)}{\Pr(\mathcal{F}_{p,q}(1-x) = y)} \le e^{\varepsilon}.$$

第一个条件等价于要求 $p,\ q \in (0,\ 1)$，第二个条件等价于 $\max\{p/q, (1-q)/(1-p)\} \le e^{\varepsilon}$。

对于 $\mathcal{M}_{p,\ q}$，注意到 $\mathcal{M}_{p,\ q}$ 的各项根据构造相互独立，因此，对于 $x,\ x' \in \{0,1\}^d$ 仅在一位 $x_j = 1 - x'_j$ 与某个 $y \in \{0,1\}^d$ 上不同的情况，我们有：

$$\frac{\Pr(\mathcal{M}_{p,q}(x) = y)}{\Pr(\mathcal{M}_{p,q}(x') = y)} = \frac{\prod_i \Pr(\mathcal{F}_{p,q}(x_i) = y_i)}{\prod_i \Pr(\mathcal{F}_{p,q}(x'_i) = y_i)} = \frac{\Pr(\mathcal{F}_{p,q}(x_j) = y_j)}{\Pr(\mathcal{F}_{p,q}(1-x_j) = y_j)}.$$

如上所示，在满足关于 $p$ 和 $q$ 的所述条件时，该量以 $e^{\varepsilon}$ 为上界。 □

下文将用定理 3.2 通过推论 4.3 与 4.7 证明 $\mathcal{M}_{\varepsilon}^{\mathrm{xor}}$ 与 $\mathcal{M}_{\varepsilon}^{\mathrm{sym}}$ 均满足 $\varepsilon$-差分隐私。

## A.3 第 4.1 节结果的证明

在证明定理 4.1 之前，我们先给出一个有用的结果：

**事实 A.1**。 设 $X \sim$ 伯努利分布$(p)$ $Y \sim$ 伯努利分布$(q)$ 相互独立。则：

*1.* $X \wedge Y \sim$ 伯努利分布$(pq)$.*2.* $X \vee Y \sim$ 伯努利分布$(p(1-q) + q(1-p) + pq)$.*3.* $X \veebar Y \sim$伯努利分布$(p(1-q)\ + q(1-p))$. 此外，如果 $p = \frac{1}{2}$或 $q = \frac{1}{2}$，则 $X \veebar Y \sim$ 伯努利分布$(\frac{1}{2})$.

证明。

$$\begin{aligned}\Pr(X \wedge Y = 1) &= \Pr(X = 1, Y = 1)\\ &= pq,\\ \Pr(X \veebar Y = 1) &= \Pr(X = 1, Y = 0) + \Pr(X = 0, Y = 1)\\ &= p(1-q) + q(1-p),\\ \Pr(X \vee Y = 1) &= \Pr(X \wedge Y = 1) + \Pr(X \veebar Y = 1).\end{aligned}$$

□

定理 *4.1* 的证明。 我们先给出条件 (1)–(3) 成立的一些简单必要条件。根据定理 3.2，我们知道 $p_1$，$p_2$，$q_1$，$q_2$ 必须属于 (0，1)。另一个必要条件是 $p_1 \neq q_1$ 与 $p_2 \neq q_2$，否则条件 (3) 不成立。因此，我们可以假设 $p_1 \neq q_1$ 与 $p_2 \neq q_2$。

除去取反的情况，存在四种对称运算 $\{0,1\}^2 \to \{0,1\}$：与 (∧)、或 (∨)、异或 (⊻)，以及将所有输入映射为 0 的平凡算子。现在我们将排除 ⊻ 以外的运算。（平凡算子） 令 $\circ$ 表示算子 $x \circ y = 0$。那么，对于任意 $x$，$y$，都有 $f_1(x) \circ f_2(y) = 0$。如果条件 (2) 成立，则 $f_3(0) = f_3(1) = 0$，这与条件 (3) 矛盾。

*(*与*)*设 $\circ = \wedge$，并假设 (2) 成立。那么我们必有：

$$f_1(0) \wedge f_2(1) \overset{D}{=} f_3(1) \overset{D}{=} f_1(1) \wedge f_2(1).$$

根据事实 A.1，这意味着：

$$q_1 p_2 = p_1 p_2.$$

由于我们假设 $p_2 \neq 0$，这意味着 $p_1 = q_1$，与我们的假设矛盾。

*(*或*)*设 $\circ = \vee$，并假设 (2) 成立。那么我们必有：

$$f_1(0) \vee f_2(1) \overset{D}{=} f_3(1) \overset{D}{=} f_1(1) \vee f_2(1).$$

根据事实 A.1，这意味着：

$$p_1(1-q_2) + q_2(1-p_1) + p_1 q_2 = p_1(1-p_2) + p_2(1-p_1) + p_1 p_2.$$

重新排列各项后，我们得到：

$$0 = (1-p_1)(p_2 - q_2).$$

由于我们假设了 $p_1 \neq 1$，这意味着 $p_2 = q_2$，与我们的假设矛盾。

*(*异或*)* 现在我们将证明，当 $\circ = \veebar$ 时，必有 $p_1 = p_2 = \frac{1}{2}$。假设条件 (2) 成立，则有：

$$f_1(0) \veebar f_2(1) \overset{D}{=} f_3(1) \overset{D}{=} f_1(1) \veebar f_2(1),$$

由此可得（根据事实 A.1）：

$$q_1(1-p_2) + p_2(1-q_1) = p_1(1-p_2) + p_2(1-p_1).$$

整理各项得：

$$p_2(p_1 - q_1) = (1-p_2)(p_1 - q_1).$$

由于我们假设了 $p_1 \neq q_1$，因此必有 $p_2 = \frac{1}{2}$。类似的论证表明 $p_1 = \frac{1}{2}$。 – □

引理 *4.3* 的证明。 我们需要证明，对于 $p = 1/2$ 和 $q = 1/(2e^{\varepsilon})$， $p/q \le e^{\varepsilon}$ 和 $(1-q)/(1-p) \le e^{\varepsilon}$。显然， $p/q = e^{\varepsilon}$，满足第一个分量。接下来，考虑表达式

$$\begin{aligned} e^{\varepsilon} - \frac{1-q}{1-p} &= e^{\varepsilon} - \frac{1 - \frac{1}{2}e^{-\varepsilon}}{\frac{1}{2}} \\ &= e^{\varepsilon} - (2 - e^{-\varepsilon}) \\ &= e^{\varepsilon} + e^{-\varepsilon} - 2 \\ &= (e^{\varepsilon/2} - e^{-\varepsilon/2})^2 \\ &\ge 0. \end{aligned}$$

因此，$(1-q)/(1-p) \le e^{\varepsilon}$，如所需。 □

定理 *4.4* 的证明。 由于 $\mathcal{M}^{\mathrm{xor}}(\cdot)$ 的各项相互独立，只需证明对于任意 $x_i$， $y_i \in \{0,1\}$，将 $\mathcal{M}^{\mathrm{xor}}$ 应用于其时结论成立。注意，若 $x_i = 1$ 或 $y_i = 1$，则 $x_i \vee y_i = 1$，因此 $\mathcal{M}^{\mathrm{xor}}_{\varepsilon^*}(x_i \vee y_i)$ ∼服从伯努利分布( 1 2)。另一方面，由于我们知道 $x_i = 1$ 或 $y_i = 1$，因此 $\mathcal{M}^{\mathrm{xor}}_{\varepsilon_1} x_i \sim \frac{1}{2} \mathcal{M}^{\mathrm{xor}}_{\varepsilon_2} y_i \sim \frac{1}{2}$ 1 $\mathcal{M}^{\mathrm{xor}}_{\varepsilon_1} x_i \veebar ( )$ $\mathcal{M}^{\mathrm{xor}}_{\varepsilon_2} y_i \sim \frac{1}{2}$ 服从伯努利分布( ) 或 ( ) 服从伯努利分布( )，由事实 A. 可得，( ) ( ) 服从伯努利分布( )。

因此，只需证明当 $x_i = y_i = 0$ 时， $\mathcal{M}^{\mathrm{xor}}_{\varepsilon^*}(x \vee y) \overset{D}{=} \mathcal{M}^{\mathrm{xor}}_{\varepsilon_1}(x_i) \veebar \mathcal{M}^{\mathrm{xor}}_{\varepsilon_2}(y_i)$。在这种情况下，$\mathcal{M}^{\mathrm{xor}}_{\varepsilon_1}(x_i)$ 服从伯努利分布($\frac{1}{2}e^{-\varepsilon_1}$)，且 $\mathcal{M}^{\mathrm{xor}}_{\varepsilon_2}(y_i)$ 服从伯努利分布($\frac{1}{2}e^{-\varepsilon_2}$)。根据事实 A.1，我们有：

$$\mathcal{M}^{\mathrm{xor}}_{\varepsilon_1}(x_i) \veebar \mathcal{M}^{\mathrm{xor}}_{\varepsilon_2}(y_i) \sim \mathrm{Bernoulli}(q^*),$$

其中

$$\begin{aligned} q^* &= \frac{1}{2}e^{-\varepsilon_1}\left(1 - \frac{1}{2}e^{-\varepsilon_2}\right) + \frac{1}{2}e^{-\varepsilon_2}\left(1 - \frac{1}{2}e^{-\varepsilon_1}\right) \\ &= \frac{1}{2}\left(e^{-\varepsilon_1} + e^{-\varepsilon_2} - e^{-(\varepsilon_1+\varepsilon_2)}\right) \\ &= \frac{1}{2}\exp\left(-\left[\underbrace{-\log\left(e^{-\varepsilon_1} + e^{-\varepsilon_2} - e^{-(\varepsilon_1+\varepsilon_2)}\right)}_{\varepsilon^*}\right]\right) \\ &= \frac{1}{2}e^{-\varepsilon^*}. \end{aligned}$$

最后，由于 $x_i \vee y_i = 0$，我们有 $\mathcal{M}^{\mathrm{xor}}_{\varepsilon^*}(x_i \vee y_i)$ 服从伯努利分布($\frac{1}{2}e^{-\varepsilon^*}$)。 □

## A.4 第 4.2 节结果的证明

引理 *4.7* 的证明。由于 $p = e^{\varepsilon}/(e^{\varepsilon}+1)$ 且 $q = 1 - p = 1/(e^{\varepsilon}+1)$，我们有 $p/q = (1-q)/(1-p) = e^{\varepsilon}$，因此可立即由定理 3.2 得出结果。 □

现在我们陈述定理 4.8 的一个更一般的形式以供证明。定理 4.8 给出了针对 2 位的随机合并，可重复调用以合并 $k > 2$ 位；而定理 A.2 考虑的是同时合并 $k \ge 2$ 位。除了用于证明原有的成对定理外，这一推广还表明，同时合并 $k$ 位与重复进行成对合并相比，在 $\varepsilon^*$（即最终草图的噪声水平）方面并无收益。

**定理 A.2.** 固定一个整数 $k \geq 2$。对于 $i \in [k]$，假设 $\varepsilon_i > 0$。令

$$q(\varepsilon) = \frac{1}{e^{\varepsilon}+1},\ K_{\varepsilon} = \begin{bmatrix} 1-q(\varepsilon) & q(\varepsilon) \\ q(\varepsilon) & 1-q(\varepsilon) \end{bmatrix},$$

$$\varepsilon^* = -\log\left(1 - \prod_{i=1}^{k}(1-e^{-\varepsilon_i})\right),\ q^* = q(\varepsilon^*),$$

并令 $v^* \in \mathbb{R}^{2^k}$ 为一个向量，其第一个元素为 $q^*$，其余元素均为 $1-q^*$。令

$$t = (t_{\ldots 01}, t_{\ldots 10}, t_{\ldots 11}, \ldots)^T = (K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1})v^*,$$

$$g(x_1, \ldots, x_k) \sim \text{Bernoulli}(t_{x_1 \ldots x_k}).$$

则 $g(\mathcal{M}_{\varepsilon_1}^{\text{sym}}(x_1) \ldots,\ \mathcal{M}_{\varepsilon_k}^{\text{sym}}(x_k)) \overset{D}{=} \mathcal{M}_{\varepsilon^*}^{\text{sym}}(x_1 \vee \cdots \vee x_k)$。

在证明定理 A.2 之前，我们先给出一个将在证明中用到的事实。

**事实 A.3**。设 $A_{m_A \times n_A}$、$B_{m_B \times n_B}$ 为满足 $A\mathbf{1} = \mathbf{1}$ 和 $B\mathbf{1} = \mathbf{1}$ 的矩阵。则 $(AB)\mathbf{1} = \mathbf{1}$ 且 $(A \otimes B)\mathbf{1} = \mathbf{1}$。此外，如果 $A^{-1}$ 存在，则 $A^{-1}\mathbf{1} = \mathbf{1}$。

证明。由于 $A\mathbf{1} = \mathbf{1}$ 与 $B\mathbf{1} = \mathbf{1}$，必有 $(AB)\mathbf{1} = A(B\mathbf{1}) = A\mathbf{1} = \mathbf{1}$。

接下来，写作 $\mathbf{1}_{m_A m_B} = \mathbf{1}_{m_A} \otimes \mathbf{1}_{m_B}$。则

$$\begin{aligned}(A \otimes B)\mathbf{1}_{m_A m_B} &= (A \otimes B)(\mathbf{1}_{m_A} \otimes \mathbf{1}_{m_B}) \\ &= (A\mathbf{1}_{m_A}) \otimes (B\mathbf{1}_{m_B}) \\ &= \mathbf{1}_{m_A} \otimes \mathbf{1}_{m_B} \\ &= \mathbf{1}_{m_A m_B}.\end{aligned}$$

最后，当 $A$ 可逆时，$A^{-1}A = I$，因此我们必须有 $A^{-1}\mathbf{1} = A^{-1}A\mathbf{1} = I\mathbf{1} = \mathbf{1}$。

□

定理 *A.2* 的证明。由于所有操作都是按位且彼此独立地执行的，不失一般性，假设位向量 $x_i$ 是标量，即 $x_1, \ldots, x_k \in \{0,1\}$。

基本思路是将对比特 $x_i$ 执行的一系列操作建模为马尔可夫链。这涉及三种不同类型的转移概率矩阵。定理陈述中定义的矩阵 $K_{\varepsilon}$，通过应用 $\mathcal{M}_{\varepsilon}^{\text{sym}}$，将 $\{0,1\}$ 中单个位的状态空间映射到另一个比特 $\{0,1\}$。接下来，我们定义 $K^{\text{or}}$ 为将 $\{0,1\}^k$ 中的 $k$ 个比特通过 或 操作映射为单个位 $\{0,1\}$ 的矩阵。最后，我们定义 $K^{\text{merge}}$ 为对应于所需合并操作的 $2^k \times 2$ 矩阵，它将 $\{0,1\}^k$ 映射到 $\{0,1\}$。

由于 $K_{\varepsilon}$ 的位翻转操作是独立执行的，矩阵 $K_{\varepsilon_1} \otimes \cdots \otimes K_{\varepsilon_k}$ 表示将原始位的状态空间 $\{x_i\}_{i=1}^k$ 联合映射到 $\{\mathcal{M}_{\varepsilon_i}^{\text{sym}}(x_i)\}_{i=1}^k$ 的 $2^k \times 2^k$ 矩阵。因此，我们希望求解：

$$(K_{\varepsilon_1} \otimes \cdots \otimes K_{\varepsilon_k})K_{\varepsilon'}^{\text{merge}} = K^{\text{or}}K_{\varepsilon'}, \tag{3}$$

其中，$\varepsilon'$ 是一个我们将确定的自由参数，$K_{\varepsilon'}^{\text{merge}}$ 是未知量。我们通过求解上述矩阵方程，找出使 $K_{\varepsilon'}^{\text{merge}}$ 表示有效转移概率矩阵的最大 $\varepsilon'$。

令 $q_i = q(\varepsilon_i)$ $q' = q(\varepsilon')$。注意到 $K_{\varepsilon_i}$ 可逆，并将 $K_{\varepsilon_i}^{-1}$写成如下形式：

$$K_{\varepsilon_i}^{-1} = \frac{1}{1-2q_i}\begin{bmatrix} 1-q_i & -q_i \\ -q_i & 1-q_i \end{bmatrix}.$$

因此，我们可以通过左乘 $(K_{\varepsilon_1} \otimes \cdots \otimes K_{\varepsilon_k})^{-1}$ 来求解矩阵方程 (3)：

$$K_{\varepsilon'}^{\text{merge}} = (K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1}) K^{\text{or}} K_{\varepsilon'}.$$

$K^{\text{or}} K_{\varepsilon'}$ 的第一列等于

$$w' = (1-q', q', \ldots, q')^T.$$

由事实 A.3 可知， $K_{\varepsilon'}^{\text{merge}} \mathbf{1} = \mathbf{1}$ ，因此 $K_{\varepsilon'}^{\text{merge}}$ 是随机的，当且仅当 ：

$$u' = (K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1})\, w' \in [0,1]^{2^k}.$$

我们可以将 $u'$ 的第 $i$ 个元素 $u_i'$ 写成 $w'$ 与 $(K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1})$ 的第 $i$ 行的内积。我们将该行向量记为 $r_i$。将 $w' = q'\mathbf{1} + (1-2q')e_1$ 写作 $u_i'$，可得：

$$\begin{aligned} u_i' &= \langle r_i,\ w' \rangle \\ &= \langle r_i,\ q'\mathbf{1} + (1-2q')e_1 \rangle \\ &= q'\langle r_i,\ \mathbf{1} \rangle + (1-2q')\langle r_i, e_1 \rangle \\ &= q' + (1-2q')(K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1})_{i1}, \end{aligned}$$

其中，最后一个等号是因为 $\langle r_i, \mathbf{1} \rangle = 1$ （事实 A.3），且 $\langle r_i, e_1 \rangle$ 等于 $K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1}$ 的第 $(i,1)$ 个元素。

由于每个 $u_i'$ 条目都必须位于 $[0,1]$, 中，因此每个条目都会对 $q'$定义一个约束。特别地，由于这些值是 $q'$的仿射函数，每个约束都对应一个区间。我们可以看到，$q' = \frac{1}{2}$ 对这些约束中的每一个都有效，因为当 $q' = \frac{1}{2}$ 时，对于所有 $i$都有 $u_i' = \frac{1}{2}$。因此，只需利用这些约束为 $q'$ 寻找一个下界。

接下来，我们考察 $u_i'$ 的哪些值最为极端。对于 $q' \in [0, \frac{1}{2}]$的任意固定选择，当 $(K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1})_{i1}$ 取最大值时，我们得到最大的条目 $u_i'$；当同一数值取最小值时，则得到最小的条目。特别地：

$$\max_i (K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1})_{i1} = \left(\prod_i (1-2q_i)^{-1}\right) \prod_i (1-q_i),$$

因为对所有 $i$， $1-q_i > q_i > 0$ 均成立。（常数 $\prod_i (1-2q_i)^{-1}$ 出现在所有条目中。）同样，我们可以看到：

$$\min_i (K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1})_{i1} = \left(\prod_i (1-2q_i)^{-1}\right) (-q_j) \prod_{i \neq j} (1-q_i),$$

其中 $j =$ 最大值自变量$_j\, q_j$，因为这会得到最极端的负项。

只需约束 $u'$ 到 $[0,1]$ 中最极端的两个条目。这两个条目定义的约束为：

$$q' + (1-2q') \underbrace{\frac{\prod_i (1-q_i)}{\prod_i (1-2q_i)}}_{=:c_1} \leq 1 \iff q' \overset{(a)}{\geq} \frac{c_1 - 1}{2c_1 - 1},$$

and

$$q' + (1-2q') \underbrace{(-q_j)\frac{\prod_{i\neq j}(1-q_i)}{\prod_i(1-2q_i)}}_{=:-c_2} \geq 0 \iff q' \overset{(b)}{\geq} \frac{c_2}{1+2c_2}.$$

我们注意到 $c_1 > 1$ 与 $c_2 > 0$。此外，比较 $c_1$ 与 $c_2$，我们发现 $c_1 > c_2 + 1$，如下：

$$\begin{aligned}
c_1 - (c_2+1) &= \frac{\prod_i(1-q_i)}{\prod_i(1-2q_i)} - \frac{q_j\prod_{i\neq j}(1-q_i) + \prod_i(1-2q_i)}{\prod_i(1-2q_i)} \\
&= \frac{(1-q_j)\prod_{i\neq j}(1-q_i)}{\prod_i(1-2q_i)} - \frac{q_j\prod_{i\neq j}(1-q_i) + \prod_i(1-2q_i)}{\prod_i(1-2q_i)} \\
&= \frac{(1-2q_j)\prod_{i\neq j}(1-q_i) - \prod_i(1-2q_i)}{\prod_i(1-2q_i)} \\
&> 0.
\end{aligned}$$

由此可得：

$$\begin{aligned}
\frac{c_1-1}{2c_1-1} - \frac{c_2}{1+2c_2} &= \frac{(c_1-1)(1+2c_2) - c_2(2c_1-1)}{(2c_1-1)(1+2c_2)} \\
&= \frac{c_2-(c_1+1)}{(2c_1-1)(1+2c_2)} \\
&> 0,
\end{aligned}$$

这表明约束 $(a)$ 蕴含约束 $(b)$。

我们将约束 $(a)$ 下允许的最小 $q'$ 记为 $q^*$：

$$q^* = \frac{c_1-1}{2c_1-1} = \frac{\prod_i \frac{1-q_i}{1-2q_i} - 1}{2\prod_i \frac{1-q_i}{1-2q_i} - 1}.$$

用 $\varepsilon$ 表示这一点，注意 $1-q_i = \frac{e^{\varepsilon_i}}{e^{\varepsilon_i}+1}$，~~$1-2q_i = \frac{e^{\varepsilon_i}-1}{e^{\varepsilon_i}+1}$，因此~~：

$$\frac{1-q_i}{1-2q_1} = \frac{e^{\varepsilon_i}}{e^{\varepsilon_i}-1} = (1-e^{-\varepsilon_i})^{-1},$$

因此：

$$q^* = \frac{\prod_i(1-e^{-\varepsilon_i})^{-1} - 1}{2\prod_i(1-e^{-\varepsilon_i})^{-1} - 1} = \frac{1-\prod_i(1-e^{-\varepsilon_i})}{2-\prod_i(1-e^{-\varepsilon_i})},$$

最终得到 $\varepsilon^*$：

$$\begin{aligned}
\varepsilon^* &= q^{-1}(q^*) \\
&= \log\left((q^*)^{-1} - 1\right) \\
&= \log\left(\frac{2-\prod_i(1-e^{-\varepsilon_i})}{1-\prod_i(1-e^{-\varepsilon_i})} - 1\right) \\
&= \log\left(\frac{1}{1-\prod_i(1-e^{-\varepsilon_i})}\right) \\
&= -\log\left(1 - \prod_{i=1}^{k}(1-e^{-\varepsilon_i})\right).
\end{aligned}$$

最后，要从转移概率矩阵转换回定理陈述，请注意，$K_{\varepsilon^*}^{\text{merge}} = (K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1}) K^{\text{or}} K_{\varepsilon^*}$ 将 $\{0,1\}^k$ 状态空间映射到 $\{0,1\}$——也就是说，$2^k$ 可能的输入会映射为伯努利随机变量，其概率取自 $K_{\varepsilon^*}^{\text{merge}}$ 的第二列。根据前面的讨论可知，这个向量恰好是（$K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1}$）$v^*$。□

## A.5 第 5 节结果的证明

**定理 *5.1* 的证明。** 令 $f(\hat{n}) = \mathbb{E}[\ell(\hat{n}; T)]$。我们将用记号 $\mathbb{E}_n[\cdot]$ 表示基数为 $n$ 时的期望，同时滥用记号，用 $\mathbb{E}_{\hat{n}}[\cdot]$ 表示将 $\hat{n}$替换 $n$后对应的量，如下所示：

$$\begin{aligned}
\mathbb{E}_n[T_{ij}] &= \Pr(T_{ij} = 1) \\
&= p(1 - \gamma_j^n) + q\gamma_j^n \\
&= p - (p - q)\gamma_j^n \\
\mathbb{E}_{\hat{n}}[T_{ij}] &= p - (p - q)\gamma_j^{\hat{n}}.
\end{aligned}$$

（注意，当 $\hat{n}$ 不是整数时，$T$ 的分布并未定义，因此 $\mathbb{E}_{\hat{n}}[\cdot]$ 并不是真正的期望。）注意：

$$\begin{aligned}
\mathbb{E}_n[f'(\hat{n})] &= B \sum_{j=1}^{P} (1 - \mathbb{E}_n[T_{ij}])(p - q)\gamma_j^{\hat{n}} \log(\gamma_j)(1 - \mathbb{E}_{\hat{n}}[T_{ij}])^{-1} \\
&\quad - B \sum_{j=1}^{P} \mathbb{E}_n[T_{ij}](p - q)\gamma_j^{\hat{n}} \log(\gamma_j)(\mathbb{E}_{\hat{n}}[T_{ij}])^{-1} \\
&= B(p - q) \sum_{j=1}^{P} \phi_j'(\hat{n}),
\end{aligned}$$

其中

$$\phi_j'(\hat{n}) = \gamma_j^{\hat{n}} \log(\gamma_j) \left( \frac{1 - \mathbb{E}_n[T_{ij}]}{1 - \mathbb{E}_{\hat{n}}[T_{ij}]} - \frac{\mathbb{E}_n[T_{ij}]}{\mathbb{E}_{\hat{n}}[T_{ij}]} \right).$$

如预期，当 $\hat{n} = n$时，该值为零。此外，当 $\hat{n} < n$ 时，该值严格为正；当 $\hat{n} > n$时，该值严格为负。因此，$f'(\hat{n})$也具有相同的性质，所以 $n$ 是 $f$的全局最大值点。

依据类似的推理，我们可以写出

$$\mathbb{E}_n[f''(\hat{n})] = B(p - q) \sum_{j=1}^{P} \phi_j''(\hat{n}),$$

其中

$$\phi_j''(\hat{n}) = (\log \gamma_j)^2 \gamma_j^{\hat{n}} \left( (1 - p) \frac{1 - \mathbb{E}_n[T_{ij}]}{(1 - \mathbb{E}_{\hat{n}}[T_{ij}])^2} - p \frac{\mathbb{E}_n[T_{ij}]}{(\mathbb{E}_{\hat{n}}[T_{ij}])^2} \right).$$

尽管对于足够大的 $\hat{n}$，$\phi_j''(\hat{n}) > 0$，但请注意，括号中的量随 $\hat{n}$ 单调递增。此外，我们知道 $\phi_j''(n) < 0$，因为 $\phi_j'(n)$ 对应于最大值。因此，对于 $\hat{n} \le n$，括号中的量（实际上，$\phi_j''(\hat{n})$ 的所有部分）必定为负。由于这对所有 $j$ 都成立，因此对于 $\hat{n} \le n$，$f''(\hat{n}) < 0$。□

**引理 A.4.** *考虑 PCSA 摘要中的一个桶，其中有 $v$ 项，且该桶有 $P = \infty$ 位。对于分配到该桶的新项，它修改该桶的概率对所有 $v > v_0$ 都有上界 $c/v$，其中某些常数为 $c$、$v_0$。*

证明。含有 $v$ 项的桶被分配到该桶的新项修改的概率为 $\sum_{\ell=1}^{\infty} 2^{-\ell}\left(1-2^{-\ell}\right)^{v}$。将此和拆分为以下区间：$\ell \in \mathcal{I}_1 := (0,\ \log_2 v - \log_2 \log_2 v)$、$\ell \in \mathcal{I}_2 := [\log_2 v - \log_2 \log_2 v,\ \log_2 v)$、$\ell \in \mathcal{I}_3 := [\log_2 v,\ \infty)$。由于 $2^{-\ell} \le 1$ 且 $(1-2^{-\ell})^v < \exp(-2^{-\ell} v) \le 1$，

$$
\begin{aligned}
\sum_{\ell \in \mathcal{I}_1} 2^{-\ell}\left(1-2^{-\ell}\right)^{v} &\le \log_2 v \exp(-\log_2 v) = o(1),\\
\sum_{\ell \in \mathcal{I}_2} 2^{-\ell}\left(1-2^{-\ell}\right)^{v} &\le \int_{\log_2 v - \log_2 \log_2 v}^{\log_2 v} 2^{-x}\exp(-2^{-x} v)\,dx\\
&= \frac{1}{v \log 2}\exp(-v 2^{-x})\Big|_{\log_2 v - \log_2 \log_2 v}^{\log_2 v}\\
&= \frac{e^{-1}}{v\log(2)} - O(\exp(-\log_2 v)/v),\\
\sum_{\ell \in \mathcal{I}_3} 2^{-\ell}\left(1-2^{-\ell}\right)^{v} &\le \sum_{\ell \in \mathcal{I}_3} 2^{-\ell} \le \frac{1}{v}.
\end{aligned}
$$

将这些部分相加即可得到所需结果。 □

定理 5.2 的证明。</style>修正 PCSA 摘要 $S_n$ 可通过以下方式生成。抽取一个新的基数 $N \sim$ 泊松$(n)$。前 $\min\{n, N\}$ 个项由常规 SFM 摘要 $S_n$ 与修正摘要 $S_n$共享。这些项在两个摘要中都分配到同一个桶。将剩余的项记为 $R = |N - n|$。泊松$(n)$ 的方差为$R = O_p(\sqrt{n})$。

利用 Kolchin 等人（1978）的定理 3 关于多项式向量中最大值的渐近分布，分配到剩余项数最多的桶中有$O_p(R/B\log B) = O_p(\sqrt{n/B}\log B)$ 个项。同样，项数最少的桶中有 $n/B + O_p(\sqrt{n/B}\log B)$ 个项。根据引理 A.4，给定 $V_i$（桶中已有的项数），一个桶中的新项更新该桶值的概率为 $O(1/V_i)$。因此，剩余项中没有任何一项更新任何桶的概率为

$$
\begin{aligned}
\left(1 - \frac{O(\sqrt{n}/B\log B)}{n/B + O(\sqrt{n}/B\log B)}\right)^{B} + o(1) &= \left(1 - \frac{O(\log B)}{\sqrt{n}}\right)^{B} + o(1)\\
&= (1 - o(1/B))^{B} + o(1) \to 1
\end{aligned}
$$

为 $n \to \infty$。由此可得 $\Pr(S_n = \widetilde{S_n}) \to 1$ 为 $n \to \infty$，真实的 PCSA 草图与泊松化后的草图渐近相等。

我们还可以将泊松化的 PCSA 草图与其真实似然为复合边际似然的草图联系起来。根据泊松拆分，$S_n$ 的各项彼此独立。因此，我们可以使用相同的底层 Uniform(0，1) 随机变量，通过逆累积分布函数法将 $S_n$ 的各项与 $S_n$ 的各项耦合起来，从而耦合~。在级别 $j$ 中，两者草图对应项不同的概率为

耦合后的草图为

$$\begin{aligned}\Pr(\tilde{S}_n(i,j) \neq \overline{S}_n(i,j)) &= \exp\left(-\frac{n}{B\,2^j}\right) - \left(1-\frac{1}{B\,2^j}\right)^n \\ &= \exp\left(-\frac{n}{B\,2^j}\right)\left(1-\exp\left[n\log\left(1-\frac{1}{B2^j}\right)+\frac{n}{B2^j}\right]\right) \\ &< \exp\left(-\frac{n}{B\,2^j}\right)\left(1-\exp\left[-\frac{n}{B^2 2^{2j+1}}\right]\right) \\ &< \exp\left(-\frac{n}{B\,2^j}\right)\frac{n}{B^2 2^{2j+1}}.\end{aligned}$$

应用并集界可得，并在某个正整数 $k$ 处拆分求和可得

$$\begin{aligned}\Pr(\tilde{S}_n \neq \overline{S}_n) &\le B\sum_{j=1}^{\infty}\exp\left(-\frac{n}{B\,2^j}\right)\frac{n}{B^2 2^{2j+1}} \\ &\le \sum_{j=1}^{k}\exp\left(-\frac{n}{B\,2^j}\right)\frac{n}{B} + \sum_{j=k+1}^{\infty}\frac{n}{B2^{2j+1}} \\ &\le \exp\left(-\frac{n}{B\,2^k}\right)\frac{nk}{B} + \frac{n}{B2^{2k}}.\end{aligned}$$

取 $k = \log_2\left(2^{\underline{n/B}_{\log(n/B)}}\right) + \delta$，其中 $\delta \in [-1/2, 1/2)$。那么第一部分

$$\begin{aligned}\exp\left(-\frac{n}{B\,2^k}\right)\frac{nk}{B} &= \frac{nk}{B}\exp\left(-2^{1-\delta}\log(n/B)\right) \\ &\le \frac{nk}{B}(n/B)^{-3/2} \to 0.\end{aligned}$$

当 $n/B \to \infty$时。同样，第二部分

$$\frac{n}{B2^{2k}} = \frac{n}{B}\left(\frac{n/B}{2\log(n/B)}\right)^{-2} 2^{-2\delta} = \frac{4\log^2(n/B)}{n/B}2^{-2\delta} \to 0$$

˜当 $n/B \to \infty$时。因此 $\Pr(S_n \equiv S_n) \to 1$ 当 $n \to \infty$时也成立。

□

推论 *5.3* 的证明。 由于 PCSA 草图 $S_n$ 与具有独立桶的修改后的草图 ˜$S_n$ 在 $n \to \infty$时相等的概率趋于 1，因此，对它们施加相同的随机响应噪声后得到的私有 SFM 草图 $T_n$、 $T_n$ 相等的概率也趋于 1。设 $\theta(T_n)$ 是某个基数估计量， $V(\theta(T_n))$ 表示其渐近方差。那么 ˆˆ ˜$\min_{\hat{\theta}\in\Theta} V(\theta(T_n)) = \min_{\hat{\theta}\in\Theta} V(\theta(T_n))$，并且，针对 $T_n$ 的基数估计量˜渐近有效，当且仅当它针对 $T_n$渐近有效。由于针对 $T_n$ 的复合似然估计量˜是针对 $T_n$的真实最大似然估计量，因此它渐近有效。

□

# B 一般布尔运算的证明与结果

在正文中，我们主要关注随机响应下模拟逻辑运算符 或 ($\vee$) 的合并操作。在这里，我们讨论将这些合并操作推广到其他布尔运算。

## B.1 对称随机响应下的布尔运算

回顾一下，推论 4.10 中给出了在 $\mathcal{M}^{\mathrm{sym}}$ 下针对 与 (∧) 的合并操作。

推论 *4.10* 的证明。我们在合并 $k$ 位这一更一般的情形下证明该结论，如定理 A.2 所示。事实上，证明过程与定理 A.2 基本等价，只需将 $K^{\mathrm{or}}$ 替换为转移矩阵 $K^{\mathrm{and}}$，该矩阵仅在 $x_1 = \cdots = x_k = 1$ 时才会将 $(x_1, \ldots, x_k)$ 映射为 $\mathbf{1}$。于是，对于固定的 $\varepsilon'$，我们有一个可能的解:

$$K_{\varepsilon'}^{\mathrm{merge}} = (K_{\varepsilon_1}^{-1} \otimes \cdots \otimes K_{\varepsilon_k}^{-1}) K^{\mathrm{and}} K_{\varepsilon'}.$$

同样，我们要寻找使 $K_{\varepsilon'}^{\mathrm{merge}}$ 成为有效转移概率矩阵的最大 $\varepsilon'$。这一次，我们将利用 $K^{\mathrm{and}} K_{\varepsilon'}$ 的第二列来确定对 $\varepsilon'$ 的约束。（这是允许的，因为 $K^{\mathrm{and}} K_{\varepsilon'} \mathbf{1} = \mathbf{1}$。）我们将第二列写作:

$$w' = (q', \ldots, q', 1 - q')^T = q'\mathbf{1} + (1 - 2q')e_{2k}.$$

按照定理 A.2 的证明中对 $w'$ 的处理方式，我们得到关于 $\varepsilon'$ 的相同约束。接着沿用定理 A.2 证明的其余部分，即可得到最终结果。 □

接下来，我们展示在 $\mathcal{M}^{\mathrm{sym}}$下合并异或 (⩛)。

**引理 B.1.** $\mathcal{M}_{\varepsilon_1}^{\mathrm{sym}}(x) \veebar \mathcal{M}_{\varepsilon_2}^{\mathrm{sym}}(y) \overset{D}{=} \mathcal{M}_{\varepsilon^*}^{\mathrm{sym}}(x \veebar y)$ 对于 $\varepsilon^* = \log(1 + e^{\varepsilon_1 + \varepsilon_2}) - \log(e^{\varepsilon_1} + e^{\varepsilon_2})$。

证明。 令 $p_1 = e^{\varepsilon_1}/(e^{\varepsilon_1} + 1) p_2 = e^{\varepsilon_2}/(e^{\varepsilon_2} + 1)$。根据事实 A.1，我们有:

$$\mathcal{M}_{\varepsilon_1}^{\mathrm{sym}}(x_i) \veebar \mathcal{M}_{\varepsilon_2}^{\mathrm{sym}}(y_i) \sim \mathrm{Bernoulli}(\theta_{x_i, y_i}),$$

其中

$$\theta_{0,0} = \theta_{1,1} = p_1(1 - p_2) + p_2(1 - p_1)$$
$$\theta_{0,1} = \theta_{1,0} = p_1 p_2 + (1 - p_1)(1 - p_2).$$

经过一些代数运算，我们得到

$$\begin{aligned}
\theta_{0,0} = \theta_{1,1} &= \frac{e^{\varepsilon_1}}{(e^{\varepsilon_1} + 1)(e^{\varepsilon_2} + 1)} + \frac{e^{\varepsilon_2}}{(e^{\varepsilon_1} + 1)(e^{\varepsilon_2} + 1)} \\
&= \frac{e^{\varepsilon_1} + e^{\varepsilon_2}}{e^{\varepsilon_1 + \varepsilon_2} + e^{\varepsilon_1} + e^{\varepsilon_2} + 1} \\
&= \frac{1}{\frac{1 + e^{\varepsilon_1 + \varepsilon_2}}{e^{\varepsilon_1} + e^{\varepsilon_2}} + 1},
\end{aligned}$$

而

$$\theta_{1,0} = \theta_{0,1} = 1 - \theta_{0,0}.$$

因此我们有

$$\mathcal{M}_{\varepsilon_1}^{\mathrm{sym}}(x_i) \veebar \mathcal{M}_{\varepsilon_2}^{\mathrm{sym}}(y_i) \overset{D}{=} \mathcal{M}_{\theta_{1,0}, 1 - \theta_{1,0}}(x_i \veebar y_i).$$

最后，为了得到 $\varepsilon^*$，注意到 $\mathcal{M}_{\theta_{1,0},\ 1-\theta_{1,0}} = \mathcal{M}_{\varepsilon^*}^{\mathrm{sym}}$ 对于

$$e^{\varepsilon^*} = \frac{1 + e^{\varepsilon_1 + \varepsilon_2}}{e^{\varepsilon_1} + e^{\varepsilon_2}}.$$

□

作为支持在 $\mathcal{M}^{\mathrm{sym}}$下进行一般布尔运算的最后一步，我们证明一元运算 非(¬) 与 $\mathcal{M}^{\mathrm{sym}}$可交换。

**引理 B.2**。 对于任意位向量 $x$， $\neg(\mathcal{M}_{p,q}(x)) \overset{D}{=} \mathcal{M}_{p,q}(\neg x)$ 当且仅当 $q = 1 - p$。特别地， $\neg(\mathcal{M}_{\varepsilon}^{\mathrm{sym}}(x)) \overset{D}{=} \mathcal{M}_{\varepsilon}^{\mathrm{sym}}(\neg x)$。

证明。 注意，

$$\neg(\mathcal{M}_{p,q}(x_i)) \overset{D}{=} \mathcal{M}_{1-p,1-p}(x_i) \overset{D}{=} \mathcal{M}_{1-q,1-p}(\neg x_i).$$

因此，

$$\neg(\mathcal{M}_{p,q}(x)) \overset{D}{=} \mathcal{M}_{p,q}(\neg x) \iff p = 1 - q \iff q = 1 - p.$$

□

## B.2 布尔运算与确定性合并

与采用随机合并的结果不同，我们证明，对于给定的布尔运算，确定性合并需要选择特定的随机响应机制，因此无法在单一随机响应机制下使用一般的布尔运算。具体而言，我们证明，确定性与 (∧) 合并所需的隐私机制不同于定理 4.1 中描述的或(∨) 合并所需的隐私机制。

**推论 B.3**。 设 $f_1 = \mathcal{F}_{p_1,q_1}$、 $f_2 = \mathcal{F}_{p_2,q_2}$、 $f_3 = \mathcal{F}_{p_3,q_3}$，并令 $\bullet\colon \{0,1\}^2 \to \{0,1\}$ 表示一个确定性对称运算。以下条件只有在 $\circ = \veebar$ 且 $q_1 = q_2 = 1/2$ 时才能同时满足：

1. $f_1$、 $f_2$ 分别对于 $\varepsilon_1$、 $\varepsilon_2 < \infty$满足 $\varepsilon_1$-差分隐私和 $\varepsilon_2$-差分隐私。
2. $f_1(x) \bullet f_2(y) \overset{D}{=} f_3(x \wedge y)$。
3. $f_i(0) \neq f_i(1)$ 对于 $i = 1$，

2，3。

证明。 注意，该定理陈述与定理 4.1 的定理陈述完全相同，区别在于条件 (2) 中将 或 运算 $\vee$ 改为与 运算 $\wedge$，且必要条件不再针对 $p_1$， $p_2$ 而是针对 $q_1$， $q_2$。

考虑定理 4.1 的原始条件 (2)。由于该条件必须对所有 $x$， $y \in \{0,1\}$成立，因此我们也可以改用 $\neg x$ 和 $\neg y$ 来表述该条件：

$$\begin{aligned}
\mathcal{F}_{q_1,p_1}(x) \circ \mathcal{F}_{q_2,p_2}(y) &\overset{D}{=} \mathcal{F}_{p_1,q_1}(\neg x) \circ \mathcal{F}_{p_2,q_2}(\neg y) \\
&\overset{D}{=} f_1(\neg x) \circ f_2(\neg y) \\
&\overset{D}{=} f_3((\neg x) \vee (\neg y)) \\
&\overset{D}{=} f_3(\neg(x \wedge y)) \\
&\overset{D}{=} \mathcal{F}_{q_3,p_3}(x \wedge y).
\end{aligned}$$

根据定理 4.1，我们知道，只有在 $p_1 = p_2 = 1/2$ 的情况下，才能同时满足该条件与条件 (1) 和 (3)。认识到这一陈述等价于推论的条件 (2)，只是将 $p_i$ 和 $q_i$ 的角色互换后，我们显然可以得出，要满足推论中的条件 (1)–(3)，必须有 $\bullet = \veebar$ 和 $q_1 = q_2 = 1/2$。

□

由此可知，没有任何隐私机制 $\mathcal{M}_{p,\ q}$ 能同时满足定理 4.1 和推论 B.3 的条件。

推论 *4.11* 的证明。 假设条件 (1)–(4) 均成立。根据定理 4.1，我们必须有$p_1 = p_2 = 1/2$，而推论 B.3 表明 $q_1 = q_2 = 1/2$。这就产生了矛盾：$f_1(0) \overset{D}{=}$ 伯努利分布$(1/2) \overset{D}{=} f_1(1)$，违反了条件 (3)。 □