arXiv:2305.02263v2 [cs.DS] 27 Sep 2023

局部边差分隐私下的三角形计数

Talya Eden*，Quanquan C. Liu†，Sofya Raskhodnikova‡，Adam Smith‡

摘要

业界许多差分隐私部署采用本地模型，在该模型中，每个参与方都通过差分隐私随机化器发布其私人信息。我们研究了边差分隐私下非交互式和交互式本地模型中的三角形计数（直观而言，该隐私定义要求：对于仅相差一条边的图，算法在这些图上的输出无法区分）。在此模型中，每个参与方的本地视图由一个顶点的邻接表构成。

在非交互模型中，我们证明加性误差 $\Omega(n^2)$ 是必要的，其中 n 表示节点数。这一下界是我们的主要技术贡献。证明采用了一种针对新类线性查询的重构攻击，并提出一种新颖的混搭策略：使用不同的邻接表补全结果运行局部随机化器。该下界与基于随机响应的算法所达到的加性误差相匹配。该算法由 Imola、Murakami 和 Chaudhuri 提出（USENIX2021），并由 Imola、Murakami 和 Chaudhuri 针对常数 ε进行了分析（CCS2022）。我们对随机响应采用了不同的后处理，并给出了所得算法方差的紧确界。

在交互式设置中，我们证明加性误差的下界为 $\Omega(n^{3/2})$。此前，除了可从中心化模型的结果直接推导出的结论外，尚无关于本地模型中交互式边隐私算法的困难性结果。我们的工作显著提升了本地模型中差分隐私图分析的现有水平。

*巴伊兰大学，talyaa01@gmail.com
†西北大学，quanquan@northwestern.edu
‡波士顿大学，{sofya, ads22}@bu.edu

1 引言

三角形计数是图分析中的一项基础操作，应用广泛，并在不同的计算模型中得到深入研究 [ELRS17, KP17, MVV16, PSKP14, PT12, AKK19, BEL+22, MV20, CEI+21]。基于三角形计数的统计量揭示了网络的重要结构信息（例如，参见 [MSOI+02, FDBV01, PDBL16]）。这些统计量用于在社交网络上执行许多计算任务，包括社区检测 [PDFV05], 链接预测 [EM02], 和垃圾信息过滤 [BBCG08]。关于三角形计数算法及其应用的综述，参见 [AHD18]。

在图（例如社交网络）包含敏感信息的应用中，对图进行计算的算法必须保护个人信息，例如特定个体之间的友谊关系。差分隐私 [DMNS06] 已成为严格隐私保证的标准。差分隐私图分析的综述，参见 [RS16] 。研究最为广泛的差分隐私设置称为中心化模型。它隐含地假设有一名数据管理者收集所有数据、对其进行计算，并发布数据。在某些情况下，将所有信息集中到一处可能并不可取，例如出于信任或责任方面的考虑。为此，人们提出了差分隐私的本地模型 [EGS03，DMNS06、KLN+11] ，如今已在许多行业部署中得到应用 [EPK14、BEM+17、CJK+18、App17、DKY17]。

在该模型中，每个参与方通过差分隐私随机化器发布其私人信息。随后，算法处理这些信息；在本地非交互式模型中，算法会输出答案。在本地交互式模型中，算法可能进行多轮交互，要求所有参与方对其私人数据运行不同的随机化器。这些随机化器可以任意依赖于先前的消息。本地模型中的差分隐私是针对参与方与算法之间的完整交互记录来定义的。在应用于图数据的本地模型中，每个顶点代表一个参与方。它接收其邻居列表作为输入，并对其应用局部随机化器。与信息属于各个参与方的典型数据集不同，在图设置中，每一对参与方（顶点）共享一项信息，即它们之间是否存在一条边。

直观地说，差分隐私保证对于任意两个相邻数据集，算法的输出分布大致相同。关于相邻图有两种自然的定义：边相邻和节点相邻。如果两个图相差一条边，则称它们边相邻；如果两个图相差一个节点及其相邻的边，则称它们节点相邻。一般而言，边差分隐私更容易实现，但节点差分隐私能够提供更强的保证。边差分隐私由 [NRS07]提出，并首次应用于三角形计数。来自 [NRS07] 的边差分隐私算法在 [KRSY14]中得到推广和实现。首批节点差分隐私算法出现在 [BBDS13、KNRS13、CZ13], 中，这三篇文章都研究了三角形计数问题。 [QYY+17、GLCZ18、SXK+19、YHA+20、IMC21、IMC22a、DLR+22] 研究了本地模型中的边差分隐私，其中大多数文献都聚焦于三角形计数。

在本研究中，我们研究了本地模型下用于估计图中三角形数量的边差分隐私算法。我们的目标是了解此类算法在非交互模型和交互模型中所能达到的加性误差。对于非交互模型，我们给出了加性误差的上界和下界。就输入图中的节点数量而言，我们的界是紧的 n, 。对于交互模型，我们给出了首个针对本地边差分隐私（LEDP）算法的专门下界。中心化模型存在基于全局敏感度的平凡下界，这些下界也适用于本地模型；但此前对于任何图问题，甚至对于 2 轮算法，都没有已知的专门针对本地模型的下界。总体而言，我们的结果加深了对非交互式和交互式 LEDP 算法的理解。

1.1 结果

我们的结果以及与先前工作的比较总结于表 1。

1.1.1 非交互本地模型的下界

我们的主要技术贡献是针对非交互式设定的一个下界。该下界采用了重构攻击（用于中心化模型），结合一类新的线性查询，以及一种新颖的混搭策略，即运行

模型		先前结果		我们的结果	
Non-交互式	下界	$\Omega(n^{3/2})$	[IMC21]	$\Omega(n^2)$	定理 1.1
	上界	$O(n^2)$（常数 ε）	[IMC22b]	$O\left(\frac{\sqrt{C_4(G)}}{\varepsilon}+\frac{n^{3/2}}{\varepsilon^3}\right)$	定理 1.2
交互式	下界	$\Omega(n)$	easy	$\Omega(\frac{n^{3/2}}{\varepsilon})$	定理 1.3
	上界	$O\left(\frac{\sqrt{C_4(G)}}{\varepsilon}+\frac{n^{3/2}}{\varepsilon^2}\right)$	<escape>[IMC22a]</escape>		

表1：非交互式和交互模型中三角形计数加性误差的下界和上界汇总。注意，$C_4(G)$ 的最大值为 $\binom{n}{4}=\Theta(n^4)$。为便于比较，[IMC21] 和 [IMC22a] 的结果均针对具有 $d_{max}=\Theta(n)$ 的图。

使用不同邻接表补全方式的局部随机化器。重构攻击是证明中心化差分隐私模型下界的有力工具，但尚未被用于推导本地模型下的界。本地模型以往的下界基于截然不同的技术，通常采用信息论论证（例如，参见 [KLN+11、BNO08、DJW13] 以及此后的许多研究）。

定理 1.1 (非交互式下界，非正式版本)。设 $\varepsilon\in(0,1/20)$ 和 $\delta\geq 0$ 为充分小的常数。存在一个图族，使得对于任何获取该图族中一个具有 n 个节点的图作为输入，并以加性误差至多为 α （以充分高的常数概率）近似图中三角形数量的非交互式(ε,δ)-局部边差分隐私算法，都必须满足 $\alpha=\Omega(n^2)$。

我们的下界适用于所有较小的 $\delta\geq 0$ （这种情况称为“近似”差分隐私）。注意，这类下界比针对 $\delta=0$ （这种情况称为“纯”差分隐私）的下界更强，因为它们将 $\delta=0$ 作为一种特殊情况包含在内。此前唯一已知的下界由 Imola 等人提出 [IMC21],，证明了非交互式算法的误差必须为 $\Omega(\sqrt{n\cdot d_{max}})$。

为了证明下界定理 1.1，我们为非交互式 e
本地模型提出了一种新颖的混搭技术。关于证明的技术概述，请参见 定理 1.1的第 1.2 节。

我们的下界与 $O(n^2)$ 给出的上界相匹配，该上界由 [IMC22b的定理 G 证明。3] （对于常数 ε），该算法基于随机响应。在这项工作中，我们给出了该算法的一个更简单的变体，并进行了更精细的分析，该分析适用于所有 ε.

1.1.2 随机响应的紧确分析

非交互式模型中最自然的算法是随机响应，其历史可追溯至 Warner [War65]。在该算法中，每个比特以 $\frac{1}{e^{\varepsilon}+1}$ 的概率翻转，其中 ε 是隐私参数。对于图而言，每个比特表示一条边是否存在。基于随机响应的三角形计数算法最初由 [IMC21] 针对 Erd˝os-R´enyi 图这一特殊情形进行了分析，随后 [IMC22b] 证明，该算法在一般图上的加性误差为 $O(n^2)$，其中 ε 为常数。这些研究首先将噪声边视为真实边，计算三角形数量以及其他由三个顶点诱导的子图数量，然后利用这些计数对估计值进行适当调整，使其无偏。

我们对随机响应采用了不同的后处理方式。我们立即对带噪边进行重新缩放，因此无需计算三角形以外的图的计数，这使得分析简单得多。我们得到了所得算法方差的紧确上界和下界，对所有 ε. 均成立。我们的界更为精细，因为它们以 $C_4(G)$（即图中的四边形数量）表示。

定理 1.2 （随机响应分析）。对于所有 $\varepsilon>0$，存在一种基于随机响应的非交互式 ε-*LEDP* 算法，该算法以 n 个节点的图作为输入，并返回图中三角形数量的无偏估计 $\hat{T}$，其方差为 $\Theta\left(C_4(G)+\frac{n^3}{\varepsilon^6}\right)$。

In 具体而言，以较高的常数概率，$\hat{T}$ 的加性误差为 $\alpha=O\left(\sqrt{C_4(G)}+\frac{n^{3/2}}{\varepsilon^3}\right)$.

注意，对于常数 ε，定理 1.2给出的上界为 $O(n^2)$，这是该算法估计值的加性误差上界。因此，根据定理 1.1中的下界，对于具有 $C_4 = \Theta(n^4)$ 的图，随机响应是最优的。此外，随机响应实现了纯差分隐私（其中$\delta = 0$），而定理 1.1 中的下界即使在近似差分隐私下也成立。尽管允许 $\delta > 0$ 能提高许多问题的准确性，但对于非交互式三角形计数，它并不能带来额外的效用。

1.1.3 交互式本地模型的下界

接下来，我们研究交互式设置下的三角形计数。Imola 等人 [IMC21] 提出了一种 ε-LEDP，用于交互模型中的三角形计数，其加性误差为 $O\big(\sqrt{C_4(G)}/\varepsilon + \sqrt{n \cdot d_{max}}/\varepsilon^2\big)$，其中 d_{max}是最大度的一个上界。

我们给出了交互模型中用于三角形计数的 LEDP 算法的加性误差下界。注意，即使在中心化模型中，三角形计数也不可避免地存在 $\Omega(n)$ 的加性误差，因为三角形数量的（边）全局敏感度为 $n-2$（并且该下界在中心化模型中是紧的）。此前，针对交互式 LEDP 模型的这一问题（或图上的任何其他问题），尚无已知的、适用于至少 2 轮算法的专门下界。我们的下界适用于具有任意 轮数的交互式算法。

定理 1.3（交互式下界）。存在一个图族和常数 $c > 0$，使得对于每个 $\varepsilon \in (0, 1)$，$n \in \mathbb{N}$、$\alpha \in (0, n^2]$ 以及 $\delta \in \left[0, \frac{1}{10^5} \cdot n^{-5}\, e^{-\varepsilon^3 \alpha_0^2 \ln(n^3/\varepsilon\alpha_0)}\right]$，任何一个能够从该图族中获取一个 n 节点图作为输入，并以至多 α 的加性误差（成功概率至少为 *2/3*）近似图中三角形数量的（可能是交互式的）（ε、δ）-本地边差分隐私算法，都必须具有$\alpha \geq c \cdot \frac{n^{3/2}}{\varepsilon}$。

我们的下界通过以下归约得到：将 n 在 $\{0,1\}$ LDP 模型中对随机采样比特求和的问题进行归约。该问题曾在一系列工作 [BNO08、CSS12、DJW+18、JMNR19]中进行研究。我们的下界与 [IMC21] 在常数 ε 以及满足 $d_{max} = \Theta(n)$ 和 $C_4(G) = O(n^3)$ 的图上的上界相匹配。对于一般图，是否能够达到 $o(n^2)$ 的加性误差仍是一个开放问题。

1.2 非交互式下界的技术概述

在本地模型中证明下界的常见技术，很大程度上依赖于对较简单数据集成立的两个事实：首先，其他参与方看不到任一参与方的信息；其次，必须保护对某一参与方信息的任意更改。在 LEDP 模型的图中，这两个条件都不成立：每条边由两个参与方共享，而且只有单条边的更改才受到差分隐私所施加的相邻数据集强意义上的保护。

为克服这些困难，我们提出了一种新的下界方法，该方法基于中心化模型中的重构攻击。这类攻击利用大量查询的准确答案，重构秘密数据集中的几乎所有条目 [DN03，DMT07、DY08、KRSU10、KRS13、De12]。这些攻击通常应用于能够发布许多不同数值的算法。然而，三角形估计算法只返回一个数字。考虑一种朴素的攻击尝试：将该算法视为黑盒，即通过分别调用三角形计数算法来模拟每个查询。这样一来，我们就需要多次运行局部随机化器，从而降低其隐私参数，使得隐私泄露的论证变得毫无意义。

为克服这一困难，在我们的攻击中，我们将非交互式三角形估计算法用作灰盒。由于该算法是非交互式的，因此它由针对所有顶点的局部随机化器以及一个对随机化器输出执行的后处理算法指定。我们使用秘密数据集 X 创建秘密子图，仅对秘密子图中的顶点运行两次随机化器，并发布结果。根据随机化器的性质以及组合，所得过程是差分隐私的。在下一阶段，我们对已发布的信息进行后处理，将秘密子图补全为与攻击所需查询相对应的不同图。然后，我们将这些图输入三角形近似算法，但对于秘密子图中的顶点，仅使用已发布的输出。如果三角形计数算法是准确的，我们就能得到查询的准确答案。尽管用于回答不同查询的随机性彼此相关，我们证明了一个良好的近似算法

对于三角形计数能够使我们正确回答大多数查询。最后，我们使用一种新颖的反集中界（下文的引理 1.4）证明，我们的攻击以高概率成功重构秘密数据集中的大部分内容。这表明，我们在此过程中运行的整体算法并非差分隐私的，从而得出结论：在非交互式 LEDP 设置中，不可能存在非常准确的三角形计数算法。

我们将攻击中使用的查询称为外积查询。这些查询是线性的，但其各个条目之间存在依赖关系。为了定义这类查询，我们将包含 n^2 比特的秘密数据集 X 表示为一个 $n \times n$ 矩阵。针对图 G 的外积查询指定两个长度为 n 、其条目取自 $\{-1, 1\}$ 的向量 A 和 B ，并返回$A^T XB$, 即 $\sum_{i,j\in[n]} A_i X_{ij} B_{j}$。

为了分析我们的重构攻击，我们证明了以下关于随机外积查询的反集中界，该界可能具有独立的研究价值。

引理 1.4（随机外积查询的反集中）。设 M 为一个 $n \times n$ 矩阵，其元素满足$M_{ij} \in \{-1, 0, 1\}$，对所有 i、$j \in [n]$ 均成立，并且 m 是 M. 中非零元素的数量。设 A 和 B 从 $\{-1, 1\}^n$. 中均匀且独立地抽取。若对于某个常数 γ，有 $m \geq \gamma n^2$，则

$$\Pr\left[|A^T MB| > \frac{\sqrt{m}}{2}\right] \geq \frac{\gamma^2}{16}.$$

关于重构攻击的文献描述了其他类型的相关查询 [KRSU10]；而此处涉及的外积查询则需要一种全新且在性质上截然不同的分析。

1.3 补充相关工作

在本地模型中证明下界的一大难点是，随机响应尽管能够提供强有力的隐私保证，却能提供足够的信息来计算相当复杂的统计量。例如，Gupta、Roth 和 Ullman [GRU12] 展示了如何利用随机响应的输出估计图中所有割的密度。Karwa 等人 [KSK14] 展示了如何根据随机响应输出拟合指数随机图模型。对于某些模型族，这需要估计三角形数量；然而，他们没有提供理论误差分析，只提供了关于收敛性的实验性证据。统计学文献也研究了随机响应，重点关注边翻转概率较小的情况。Balachandran 等人 [BKV17] 分析了在翻转概率极低时，对随机响应结果中的三角形数量进行计数的朴素估计量的分布。Chang 等人 [CKY22] 针对翻转率未知、但有多个独立噪声重复样本可用的情形，提出了估计策略。据我们所知，这些工作并未阐明最具隐私相关性的情形，即边翻转概率接近 $1/2$。

一些研究在经验设置下考察了三角形计数及其他图问题，采用的是“去中心化”隐私模型 [SXK+19, QYY+17, GLCZ18, YHA+20] 。除了 [SXK+19]之外，本地视图均由邻接表构成。Sun 等人 [SXK+19] 的本地视图由两跳邻域构成。这种模型的误差较小，因为节点可以看到所有与其相邻的三角形，并可使用几何机制报告这些相邻三角形。

1.4 组织结构

差分隐私的各种模型（包括 LEDP）在 第 2 节中定义。我们针对非交互模型的下界证明（定理 1.1）见第 3 节。外积查询的反集中引理（引理 1.4）在第 3.2 节中证明。我们对随机响应的分析及定理 1.2 的证明见第 4 节。定理 1.3 针对交互式 LEDP 模型的证明见第 5 节。在附录A中，我们陈述了若干隐私工具（附录 A.1）和集中界（附录 A.2）。

2 差分隐私背景

我们首先介绍适用于以向量表示的数据集和图数据集的差分隐私定义。

定义 2.1 (差分隐私 [DMNS06、DKM+06])。 设 $\varepsilon > 0$ 和 $\delta \in [0,1)$。随机化算法 $\mathcal{A}$ 是 (ε, δ)-差分隐私（差分隐私）（相对于数据集全集上的邻接关系），当且仅当对于所有事件 S，在 $\mathcal{A}$ 的输出空间中，以及所有相邻数据集 X 和 X'，

$$\Pr[\mathcal{A}(x) \in S] \leq \exp(\varepsilon) \cdot \Pr[\mathcal{A}(X') \in S] + \delta.$$

当 $\delta = 0$ 时，该算 法是 ε-差分隐私的（有时也称为"纯差分隐私"）.

差分隐私可以针对任意邻接数据集的概念来定义。当数据集表示为向量时，如果数据集 X 和 Y 仅有一个条目不同，则认为它们相邻。在图的语境中，有两种自然的图邻接概念可用于定义：边相邻和节点相邻。我们主要采用前者，但为清楚地讨论先前的工作，也对两者都进行定义。

定义 2.2。 如果两个图 $G = (V, E)$ 和 $G' = (V', E')$ 的 G 和 G' 恰好相差一条边，也就是说， $V = V'$ 和 E 以及 E' 恰好相差一个元素，则称它们<姓名 id='17'>边相邻 。如果一个图可以通过移除另一个图中的一个节点及其相邻边得到，则称这两个图节点相邻 。

如果数据集是图，且采用边（分别为节点）邻接关系，则将差分隐私算法简称为边隐私 （分别为节点隐私）。

2.1 本地模型

差分隐私的定义隐含地假设存在一位可信数据管理者，它能够访问数据、在数据上运行隐私算法，并发布结果。这种设置称为差分隐私的中心化模型 。相比之下，在差分隐私的本地模型 中，参与计算的每个参与方都持有自己的数据。参与方之间的交互由算法 $\mathcal{A}$ 协调，该算法通过局部随机化器访问数据。局部随机化器是一种在单个参与方的数据上运行的差分隐私算法。在图数据集的背景下，输入图按如下方式分布在各参与方之间：每个参与方对应图中的一个节点，其数据是图的邻接矩阵中相应的一行。在每轮交互中，算法 $\mathcal{A}$ 为每个参与方分配一个或多个局部随机化器，这些随机化器可以依赖于先前各轮中获得的信息。

我们将 [JMNR19、KLN+11] 中的本地差分隐私定义推广到图的情形。考虑一个无向图 $G = ([n], E)$，它由一个 $n \times n$ 邻接矩阵 A表示。每个参与方 $i \in [n]$持有 i的第 行，记作 $\mathbf{a}_{i*}$。 我们有时将 $\mathbf{a}_{i*}$ 称为参与方 的邻接向量 i。 矩阵 A 的各项记作 a_{ij}，其中 $i, j \in [n]$.

定义 *2.3*（局部随机化器）。设 $\varepsilon > 0$ 和 $\delta \in [0,1)$。*(ε, δ)*-局部随机化器 R*:* $\{0,1\}^n \to \mathcal{Y}$是一个*$(\varepsilon, \delta)$*-边差分隐私算法，其输入为一个节点的邻居集合，以邻接向量 $a \in \{0,1\}^n$表示。换言之，对于所有相差一位的 a 和 a'，以及所有输出集合 $Y \subseteq \mathcal{Y}$，均有 $\Pr[R(a) \in Y] \leq e^{\varepsilon} \cdot \Pr[R(a') \in Y] + \delta$。概率是对 R 的随机硬币取值计算的（但不对输入的选择计算）。当 $\delta = 0$时，我们称 R 为 ε-局部随机化器。

一个随机 化算法 $\mathcal{A}$ 在分布式图上是 (ε, δ)-LEDP，当且仅当其满足 定义 2.4 .

定义 2.4 （局部边差分隐私）。 一个交互记录 π 是一个由 *5* 元组组成的向量$(S_U^t, S_R^t, S_\varepsilon^t, S_\delta^t, S_Y^t)$——对每一轮 t，记录所选参与方集合、分配的随机化器集合、随机化器隐私参数集合以及产生的随机化输出。设 S_π 为所有交互记录的集合， S_R 为所有随机化器的集合。设 $\perp$ 表示一个特殊字符，指示计算已停止。在此模型中，算法 是一个函数 $\mathcal{A}$: $S_\pi \to (2^{[n]} \times 2^{S_R} \times 2^{\mathbb{R}^{\geq 0}} \times 2^{\mathbb{R}^{\geq 0}}) \cup \{\perp\}$，将交互记录映射到参与方、随机化器和随机化器隐私参数的集合。交互记录的长度以 t为索引，即为其轮复杂度。

给定 $\varepsilon \geq 0$和 $\delta \in [0,1)$，若在（分布式）图 G上的随机化算法 $\mathcal{A}$满足以下条件：输出 $\mathcal{A}$ 所生成的完整交互记录的算法在图 G. 上满足(ε, δ)-边差分隐私，则称其为(ε, δ)-局部边差分隐私（***LEDP***）。当 $\delta = 0$时，我们称 $\mathcal{A}$ 为 ε-*LEDP*。

如果 $t = 1$，也就是说，只有一轮，那么 $\mathcal{A}$ 就称为非交互式。否则， $\mathcal{A}$ 就称为交互式。

注意，非交互式LEDP算法由每个节点对应的一个局部随机化器，以及一个后处理算法 $\mathcal{P}$ 构成；该算法将局部随机化器的输出作为输入。

我们使用一种称为随机响应的局部算法，该算法最初由 [War65], 提出，后来被改造为适用于差分隐私 [KLN+11]。

定义 2.5 （随机响应）。 给定隐私参数 $\varepsilon > 0$ 和一个 k比特向量$\mathbf{a}$，算法RandomizedResponse$_\varepsilon$（$\mathbf{a}$）输出一个 k比特向量，其中对于每个 $i \in [k]$，比特 i 以 $\frac{e^\varepsilon}{e^\varepsilon+1}$ 的概率为 a_i，否则以$1 - a_i$ 为值。

定理 2.6 （随机响应是 ε-LR）。 随机响应是一种 ε-局部随机化器。

其他隐私工具可见于附录A。

3 非交互式下界

在本节中，我们证明定理 1.1，现于此处正式重述。

定理 3.1。 存在一个图族，使得对于任何非交互式 (ε, δ)-*LEDP* 算法，若其满足$\varepsilon \in (0, \frac{1}{20})$和 $\delta \in [0, \frac{1}{100})$，以该图族中的一个 n节点图作为输入，并以至少 $1 - \frac{1}{3^6 \cdot 2^7}$ 的概率将图中的三角形数量近似到加性误差 α 以内，则必须有 $\alpha = \Omega(n^2)$。

从高层次来看，该下界的证明方法是：说明一个用于计数三角形的本地非交互式算法，可以被用来对秘密数据集实施重构攻击 X，而这发生在差分隐私的中心化模型中。Dinur和Nissim的一项突破性结果 [DN03]——并在后续工作中得到推广[DMT07、DY08、KRSU10、KRS13、De12]——表明，如果一个算法对包含 N 比特的敏感数据集回答过多随机线性查询，且答案过于准确，那么就可以重构出该数据集中的一个较大的常数比例。这被称为“重构攻击”。具体而言，Dinur和Nissim表明，回答误差在 $\pm O(\sqrt{N})$ 以内的 N 个随机线性查询就足以进行重构。众所周知，如果某个算法对秘密数据集的输出可用于重构，那么该算法就不是差分隐私的。这种推理方式给出了任何回答 N 个随机线性查询的差分隐私算法，其加性误差至少为 $\Omega(\sqrt{N})$。

假设我们可以证明，一个加性误差为 $O(n^2)$ 的 LEDP 三角形计数算法可用于构造一个差分隐私算法，以在某个大小为 n 的数据集上回答 n 个线性查询，且加性误差为 $O(\sqrt{n})$——那么根据上述结果，我们就会得出与该算法的隐私性相矛盾的结论。事实上，三角形计数算法确实可用于回答单个线性查询，但主要挑战在于，Dinur和Nissim 的重构攻击要求回答的不是一个，而是 n个关于同一数据集的线性查询。设 $\mathcal{A}$ 是一个 (ε, δ)-LEDP 三角形计数算法。如果我们天真地尝试以黑盒方式，每次回答一个线性查询 X 时都重新调用一次三角形计数算法，就会调用 $\mathcal{A}$ 共 n 次。这反过来会导致隐私参数随 n 线性增长，使隐私泄露变得毫无意义。也就是说，我们将得到如下形式的结果：一个具有较小加性误差的 (ε, δ)-LEDP 三角形计数算法，意味着存在一个具有较小加性误差的 $(O(\varepsilon n), O(n\delta))$-差分隐私算法，可用于回答线性查询。由于后一结论过于薄弱，无法与 Dinur和Nissim 的结果结合使用，因此我们采用了另一种方法。

为了避免进行 n 次三角形计数算法调用，我们针对秘密数据集 X开发了一种新型重构攻击，其中允许的线性查询集合具有特殊的组合结构。我们将这类新型查询称为外积查询。我们证明，若能访问一个 (ε, δ)-LEDP 算法 $\mathcal{A}$，该算法对三角形数量的近似误差至多为 $O(n^2)$ 加性误差，则我们可以设计一个 $(2\varepsilon, 2\delta)$-差分隐私算法 $\mathcal{B}$，用于回答数据集 X 上的 $\Theta(n^2)$ 个外积查询，该数据集大小为 $N = n^2$，且其中固定比例的查询回答误差为 $O(n)$ 加性误差。（数据集大小为 n^2，因此渐近来看，随机查询的数量和所需精度与 Dinur-Nissim 攻击相同。）核心思路是，不将 $\mathcal{A}$ 当作黑盒使用，而是以“灰盒”方式使用。这使我们能够回答全部 $\Theta(n^2)$ 个查询，而不降低 $\mathcal{B}$ 的隐私参数。进而，我们可以重构 X，这与算法 $\mathcal{B}$ 的隐私性相矛盾，因此也与算法 $\mathcal{A}$ 的隐私性相矛盾。因此，我们得出结论：任何 LEDP 三角形计数算法都必须具有 $\Omega(n^2)$ 加性误差。

本节其余部分的第 3.1 节组织如下。在第 3.1 节中，我们定义外积查询，并证明一个用于三角形计数、加性误差较小的（ε，δ）差分隐私算法 $\mathcal{A}$ 可用于构造一个

（2ε， 2δ）差分隐私算法 $\mathcal{B}$，用于以较小的加性误差回答外积查询。在第 3.2 节中，我们证明随机外积查询的一个反集中结果。在第 3.3 节中，我们利用该反集中结果证明，算法 $\mathcal{B}$ 若能准确回答敏感数据集 Θ（n^2）上的外积查询，便可用于重构 $X \in \{0,1\}^{n\times n}$ 的大部分内容 X，并完成定理 3.1的证明。

3.1 从外积查询到三角形计数的归约

在本节中，我们证明引理 3.3，这是我们归约的核心。该引理表明，给定一个可访问的（ε， δ）LEDP 算法 $\mathcal{A}$，该算法能够以较低的加性误差近似三角形数量，我们可以构造一个（2ε， 2δ）差分隐私算法 $\mathcal{B}$ （在中心化模型中），从而准确回答敏感数据集 Θ（n^2）个外积查询 X。我们首先正式定义这一类新查询。

定义 3.2 （外积查询）。设 $X \in \{0,1\}^{n\times n}$。对 X 的一次外积查询指定两个长度为 n、元素取自 $\{-1,1\}$ 的向量 A和 B，并返回 $A^T XB$，即 $\sum_{i,\ j\in[n]} A_i X_{ij} B_j$。

令 γ 为所需的重构参数；若我们正确重构了 X 中至少 $(1-\gamma)n^2$ 比特，则表示攻击成功。（稍后在第 3.3 节中， γ 将设为 $1/9$，查询数量 k将设为 $\Theta(n^2)$。）

引理 3.3 （通过三角形计数回答外积查询）。设 $\varepsilon, \delta > 0$ 和 $\gamma \in (0, 1/2)$。假设存在一个非交互式 (ε, δ)-*LEDP* 算法 $\mathcal{A}$，对于每个 $3n$ 节点图，以至少 $1 - \frac{\gamma^2}{9\cdot 128}$ 的概率近似三角形数量，且加性误差至多为 $\frac{\sqrt{\gamma} n^2}{20}$。那么，在中心化模型中存在一个 $(2\varepsilon, 2\delta)$-差分隐私算法 $\mathcal{B}$，对于每个秘密数据集 $X \in \{0,1\}^{n\times n}$ 和任意一组 k 个外积查询 $(A^{(1)}, B^{(1)}) \ldots, (A^{(k)}, B^{(k)})$，给出答案 $a_1, \ldots, a_k$，满足

$$\Pr\left[\left|\left\{\ell \in [k] : \left|(A^{(\ell)})^T X B^{(\ell)} - a_\ell\right| > \frac{\sqrt{\gamma} n}{4}\right\}\right| > \frac{\gamma^2 k}{64}\right] \le \frac{1}{6}.$$

也就是说，对于任意数据集 X 和一组 k 个外积查询，算法 $\mathcal{B}$ 至多以 $5/6$ 的概率对其中 k 个查询给出不准确的答案；这里，不准确是指加性误差大于 $\frac{\sqrt{\gamma} n}{4}$。

证明。考虑引理前提中描述的算法 $\mathcal{A}$。由于 $\mathcal{A}$ 是局部非交互式的，它由每个顶点 v对应的一个局部随机化器 R_v(**a**)以及一个后处理算法 $\mathcal{P}$指定。每个随机化器将邻接向量 $\mathbf{a} \in \{0,1\}^n$ 作为输入，并将其输出传递给 $\mathcal{P}$. 接下来，我们定义算法 $\mathcal{B}$，该算法给定一个敏感数据集 X 和一组 k 外积查询后，将随机化器和后处理算法作为子程序，以获得这些外积查询的准确答案。

固定一个数据集 $X \in \{0,1\}^{n\times n}$。对于每个外积查询（$A$， B），算法 $\mathcal{B}$构造若干个对应的查询 图。所有查询图都定义在同一个顶点集 V上，该顶点集的大小为 $3n$，并划分为三个集合 U_1、U_2、和 W，其大小为 n。对于 U_t中的顶点， $t \in \{1,2\}$对应的顶点记为 u_{t1}、……、u_{tn}。 W中的顶点记为 w_1、……、w_n。示意见图 1。

算法 $\mathcal{B}$ 首先构造一个二部图 G_X，其两部分为 U_1 和 U_2，并以 X 作为邻接矩阵；也就是说，对于每个满足 $X_{ij} = 1.$ 的 i、$j \in [n]$，它都会添加一条边 (u_{1i}, u_{2j})。我们称 G_X为秘密子图，因为它会作为子图包含在每个查询图中，并且它是该图中唯一包含原始敏感数据集 X信息的部分。注意， G_X 不依赖于外积查询。每个查询子图的其余边连接 $U_1 \cup U_2$ 与 W，并且因查询图而异，因此最终得到的图整体上是三部图。对于每个 $v \in U_1 \cup U_2$，令 $\Gamma_X(v)$ 表示秘密子图 G_X中 v 的邻居。该构造的一个关键思路是，秘密子图 G_X 中的每个节点在每个查询图中都只会有两种可能的邻居集合之一。这使算法 $\mathcal{B}$ 能够通过对 $U_1 \cup U_2$中的每个顶点仅调用局部随机化器两次，来模拟所有查询图上的三角形计数计算。对于每个顶点 $v \in U_1 \cup U_2$，算法 $\mathcal{B}$两次运行其局部随机化器 $R_v(\cdot)$：一次使用由 $\Gamma_X(v)$指定的邻接表，另一次使用由 $\Gamma_X(v)$指定的邻接表 $\cup W$。随后，算法 $\mathcal{B}$ 将前一次调用的输出记录为 $r_0(v)$，将后一次调用的输出记录为 $r_1(v)$。

根据定理 A.1，仅输出所有 $4n$ 本地随机化器的响应向量的算法，通过2ε， 2δ组合可知是 X -差分隐私，因为的每个比特都被编码为一条潜在的边，并在其端点的两次

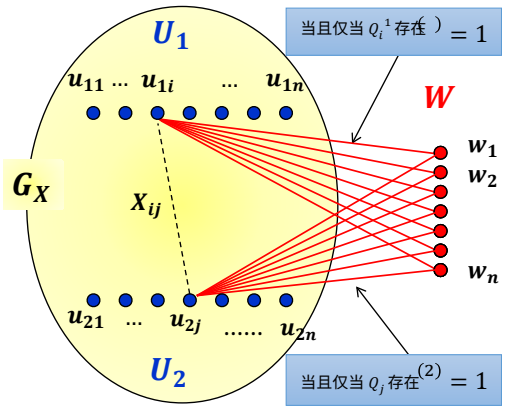


图1：查询图 $G_{X,Q}$ 的构造。部分 U_1, U_2, W 各由 n 个节点组成。当且仅当 $X_{ij} = 1$ 时，虚线表示一条边。只有子图 G_X （由 $U_1 \cup U_2$ 诱导）包含秘密信息。

随机化器执行中使用，其中每次执行（所有随机化器）均为 (ε, δ)-LEDP。在剩余步骤中，算法 $\mathcal{B}$ 仅对响应向量进行后处理，因此它是 $(2\varepsilon, 2\delta)$-差分隐私。

接下来，我们将介绍如何对响应向量进行后处理，以获得外积查询 (A, B) 的答案。对于每个外积查询，算法 $\mathcal{B}$ 将首先获取三个线性查询的答案，我们称之为 子矩阵查询。子矩阵查询的定义与外积查询相同，不同之处在于向量 A 和 B 的元素属于 $\{0,1\}$，而非 $\{-1,1\}$。接下来，我们将说明如何在 X 上回答子矩阵查询，并将每个外积查询如何通过子矩阵查询进行模拟的描述留待 命题 3.5 给出。

为回答数据集 X 上的子矩阵查询 $Q = (Q^{(1)}, Q^{(2)})$，算法 $\mathcal{B}$ 按如下方式将秘密子图 G_X 补全为查询图 $G_{X,Q}$。对于每个顶点 $u_{ti} \in U_1 \cup U_2$，其中 $t \in \{1,2\}$ 且 $i \in [n]$，根据 $Q^{(t)}$ 确定要添加的边：具体而言，若 $Q_i^{(t)} = 1$，则从 u_{ti} 向 W 中的所有顶点添加边。接下来的命题给出了图 $G_{X,Q}$ 中的三角形数量与子矩阵查询 Q 的答案之间的关系。

命题 3.4。 图 $G_{X,Q}$ 中的三角形数量等于 $n \cdot (Q^{(1)})^T X Q^{(2)}$.

证明。 注意，图 $G_{X,Q}$ 是一个三部图，其部分为（U_1, U_2, W），因此每个三角形都必须在每个部分各包含一个顶点。子矩阵查询 $Q = (Q^{(1)}, Q^{(2)})$ 的答案为

$$(Q^{(1)})^T X Q^{(2)} = \sum_{i,j \in [n]} Q_i^{(1)} Q_j^{(2)} X_{ij}.$$

对于求和中的每一项，当且仅当 $Q_i^{(1)} = Q_j^{(2)} = 1$. 时，$u_{1i}$ 和 u_{2j} 都与 W 中的所有节点相邻。如果图中存在边（u_{1i} u_{2j} n $Q_i^{(1)} = Q_j^{(2)} = X_{ij} =$，），则会 1 n 形成三角形。因此，每个对应于三角形的项都具有如下形式：（u_{1i}, u_{2j}, w_ℓ），其中 $\ell \in [n]$。其他所有项都不会形成三角形，因为要么 $X_{ij} = 0$，此时边（u_{1i}, u_{2j}）不存在于图中；要么 $Q_i^{(1)} = 0$ 或 $Q_j^{(2)} = 0$，，此时 u_{1i} 和 u_{2j} 没有公共邻居。 □

为回答子矩阵查询 Q，算法 $\mathcal{B}$ 在相应的查询图 $G_{X,Q}$ 上模拟调用三角形计数算法 $\mathcal{A}$。首先，$\mathcal{B}$ 使用图 $G_{X,Q}$. 指定的邻接向量，运行 W 中各顶点的局部随机化器。请注意，这些顶点无法获取任何私人信息，因此此操作不会影响隐私。对于每个顶点 $u_{ti} \in U_1 \cup U_2$，其中 $t \in \{1,2\}$ 和 $i \in [n]$，算法 $\mathcal{B}$ 使用先前运行的随机化器所产生的结果 $r_b(u_{ti})$，其中 $b = Q_i^{(t)}$。例如，若 $Q_i^{(1)} = 0$，则 $\mathcal{B}$ 使用结果 $r_0(u_{1i})$；若 $Q_i^{(1)} = 1$，则使用结果 $r_1(u_{1i})$。现在，算法 $\mathcal{B}$ 已获得图 $G_{X,Q}$ 上所有顶点随机化器的结果，并直接对这些结果运行后处理算法 $\mathcal{P}$。为得到子矩阵查询的答案，$\mathcal{B}$ 将 $\mathcal{P}$ 的输出除以 n。

最后，算法 $\mathcal{B}$ 通过获取三个子矩阵查询的答案，按照以下命题的说明回答每个外积查询。

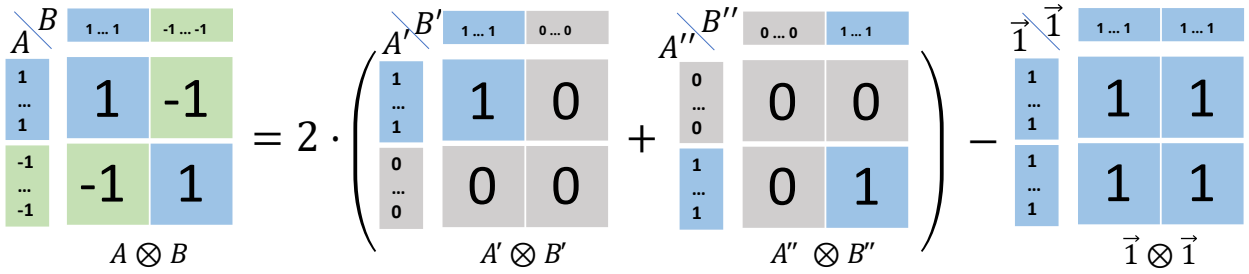


图 2：每个外积查询都可以使用三个子矩阵查询进行模拟。为便于说明，将所有向量的元素重新排列，使相同的值归为一组。外积表示为 $\otimes$.

命题 3.5. 对 X 的外积查询可以通过对 X 进行三次子矩阵查询来模拟。此外，如果这三次子矩阵查询的答案的加性误差均不超过 α，那么外积查询的答案的加性误差不超过 5α.

证明。 考虑对一个 $n \times n$ 矩阵 X 指定的外积查询 A，$B \in \{-1,1\}^n$. 定义 n 比特向量 $A' = \frac{1}{2}(A + \vec{1})$ 和 $A'' = \frac{1}{2}(-A + \vec{1})$，其中 $\vec{1}$ 表示一个长度为 n. 的全 1 向量。定义 B' 和 B'' 的方式类似。于是，如 图 2所示，

$$A^T X B = 2((A')^T X B' + (A'')^T X B'') - \vec{1}^T X \vec{1}.$$

也就是说，外积查询 $(A,\ B)$ 的答案可以根据子矩阵查询 $(A',\ B')$、$(A'',\ B'')$ 和 $(\vec{1},\ \vec{1})$ 的答案计算得出，而加性误差从 α 增加到 5α，如上所述。 □

接下来证明以下命题。

命题 3.6。 设 $\mathcal{A}$ 如引理 3.3的前提所述。对于每个秘密数据集 $X \in \{0,1\}^{n \times n}$ 以及任意一组 k 外积查询 $\{(A^{(\ell)},\ B^{(\ell)})\}_{\ell \in [k]}$，算法 $\mathcal{B}$ 给出答案 $a_1, \ldots, a_k$ 满足

$$\Pr\left[\left|\left\{\ell \in [k] : \left|(A^{(\ell)})^T X B^{(\ell)} - a_\ell\right| > \frac{\sqrt{\gamma} n}{4}\right\}\right| > \frac{\gamma^2 k}{64}\right] \leq \frac{1}{6}.$$

也就是说，“错误地”回答的外积查询数量超过 $\frac{\gamma^2 k}{64}$ 的概率至多为 $1/6$。

证明。 根据关于 $\mathcal{A}$ 的假设，对于每个图 G，算法 $\mathcal{A}$ 以至少 $1 - \frac{\gamma^2}{9 \cdot 128}$ 的概率返回图 G 中的三角形数量，且加性误差至多为 $\frac{\sqrt{\gamma} n^2}{20}$。

给定秘密数据集 X 和 k 个外积查询，算法 $\mathcal{B}$ 首先为这些外积查询创建 k 组三个一组的子矩阵查询。然后，$\mathcal{B}$ 将 $\mathcal{A}$ 用作灰盒，同时回答所有 $3k$ 个子矩阵查询。回想一下，这是通过对持有私人信息的顶点（即属于部分 U_1，U_2的顶点）调用局部随机化器来实现的：针对某个查询中与该顶点对应的比特的每个可能取值，分别调用一次。然后，对每个单独的子矩阵查询，在 W 中具有与该特定查询图相对应的邻接表的 n 个顶点中的每一个顶点上调用一个局部随机化器。接着，为了回答每个特定的子矩阵查询，算法 $\mathcal{B}$ 将 W 中各顶点的新输出与运行随机化器于 $U_1 \cup U_2$ 上所得的、对应于该特定查询的已存储输出相结合，并对这个由 $3n$ 个输出组成的向量调用后处理算法 $\mathcal{P}$。最后，$\mathcal{B}$ 将 $\mathcal{P}$的答案除以n，以得到子矩阵查询的答案。

$\mathcal{P}$ 的每次调用通过 $\mathcal{B}$ 模拟一次三角形计数计算。总体而言，我们有 $3k$ （依赖的）次模拟的三角形计数计算。根据 引理 3.3前提中关于 $\mathcal{A}$的假设，后处理算法 $\mathcal{P}$ 对每次模拟的三角形计数计算给出不准确答案（即加性误差超过 $\frac{\sqrt{\gamma} n^2}{20}$）的概率至多为 $\frac{\gamma^2}{9 \cdot 128}$（其中，该概率取自单个 $3n$ 局部随机化器的随机硬币以及 $\mathcal{P}$的随机硬币）。总体而言，共有$3k$ （依赖的）次模拟，因此，对于 $\mathcal{A}$返回的加性误差大于 $\frac{\sqrt{\gamma} n^2}{20}$ 的模拟三角形计数计算，其期望数量至多为 $\gamma^2 \cdot (\frac{3k}{9 \cdot 128}) = \frac{\gamma^2 \cdot k}{6 \cdot 64}$。因此，根据马尔可夫不等式，模拟三角形查询中不准确查询的数量超过 $\frac{\gamma^2 \cdot k}{64}$ 的概率至多为 $\frac{1}{6}$。

在以下事件发生的条件下：至多 $\frac{\gamma^2 \cdot k}{64}$ 次三角形计数计算的答案不准确，因此其余计算的答案误差至多为 $\alpha = \frac{\sqrt{\gamma} n^2}{20}$，将该事件记为$E$。回顾一下，每次三角形计数计算都用于回答二个子矩阵查询，并且根据命题 3.4，如果某次三角形计数计算的答案具有加性误差 α，则相应的子矩阵查询答案具有加性误差 α/n。因此，在上述条件下，至多 $\frac{\gamma^2 \cdot k}{64}$ 个子矩阵查询的答案具有大于 α/n的加性误差。三角形计数计算的每个不准确答案至多会影响一个外积查询的答案。此外，根据命题 3.5，如果用于计算单个外积查询的全部三个子矩阵查询的答案的误差均在 α/n以内，则该外积查询的答案误差在 $5\alpha/n$以内。因此，在上述条件下，至多 $\frac{\gamma^2 \cdot k}{64}$ 个外积查询的答案具有大于 $5\alpha/n = \frac{\sqrt{\gamma} n}{4}$ 的加性误差。由于事件 E 发生的概率至少为 $5/6$，因此我们得到：以至少 $5/6$的概率，外积查询中答案具有大于 $\frac{\sqrt{\gamma} n}{4}$ 的加性误差的查询比例至多为 $\frac{\gamma^2 \cdot k}{64}$，从而 命题 3.6 成立。□

至此，引理 3.3的证明完成。□

3.2 随机外积查询的反集中

在本节中，我们证明引理 1.4。为了分析我们的重构攻击，我们将考虑真实数据集 X 和一个潜在的重构数据集 Y。令 M 表示 $X - Y$。那么，对于一个外积查询（A, B），该查询在数据集 X 上的答案与在数据集 Y上的答案之差为 $A^T XB - A^T YB = A^T MB$. 本节的主要结果表明，如果 X和 Y在许多条目上不同（即， M 含有许多非零条目），那么一个随机外积查询很可能在 X 和 Y上产生显著不同的答案。

证明引理 *1.4*。 令 $Z_{ij} = A_i B_j$对所有 i, $j \in [n]$, 以及 $U = A^T MB$. 我们通过计算 U的期望、二阶矩和四阶矩来证明该引理，然后将帕利–齐格蒙德不等式（见定理 A.5）应用于 U^2.

由 A_i 和 B_j 对所有 $i, j \in [n]$, 相互独立可知，我们有 $\mathbb{E}[Z_{ij}] = \mathbb{E}[A_i] \cdot \mathbb{E}[B_j] = 0$，且 $\mathbf{Var}[Z_{ij}] = \mathbb{E}[Z_{ij}^2] = \mathbb{E}[A_i^2 B_j^2] = 1$。根据 U 的定义以及期望的线性性，

$$\mathbb{E}[U] = \mathbb{E}[A^T MB] = \mathbb{E}\Big[\sum_{i,j\in[n]} M_{i,j} Z_{i,j}\Big] = \sum_{i,j\in[n]} M_{i,j}\mathbb{E}[Z_{i,j}] = 0.$$

注意，随机变量 Z_{ij} 两两独立。这是随机外积查询的一项重要特征，也是使用它们而非子矩阵查询的主要原因。这一特征大大简化了分析。由于 U 是无偏的， $\mathbb{E}[U^2] = \mathbf{Var}[U]$。由 Z_{ij} 的两两独立性，

$$\mathrm{Var}[U] = \mathrm{Var}\Big[\sum_{i,j\in[n]} M_{ij} Z_{ij}\Big] = \sum_{i,j\in[n]} M_{ij}^2 \ \mathrm{Var}[Z_{ij}] = \sum_{i,j\in[n]} M_{ij}^2 = m.$$

接下来，我们给出 U 四阶矩的一个上界。

断言 3.7。 $\mathbb{E}[U^4] \le 9n^4$.

证明。 我们使用 U 的定义，将其展开为求和，并展开乘积中的各项：

$$\begin{aligned}\mathbb{E}[U^4] = \mathbb{E}[(A^T MB)^4] &= \mathbb{E}\Big[\Big(\sum_{i,j\in[n]} M_{ij} Z_{ij}\Big)^4\Big] \\ &= \sum_{(i_1,j_1),\ldots,(i_4,j_4)\in[d]\times[d]} M_{i_1 j_1} M_{i_2 j_2} M_{i_3 j_3} M_{i_4 j_4}\ \mathbb{E}[Z_{i_1 j_1} Z_{i_2 j_2} Z_{i_3 j_3} Z_{i_4 j_4}], \qquad (1)\end{aligned}$$

其中式（1）是利用期望的线性性得到的。接下来，我们计算式（1）中乘积的期望：

$$\begin{aligned}\mathbb{E}[Z_{i_1j_1}Z_{i_2j_2}Z_{i_3j_3}Z_{i_4j_4}] &= \mathbb{E}[A_{i_1}B_{j_1}A_{i_2}B_{j_2}A_{i_3}B_{j_3}A_{i_4}B_{j_4}] \\ &= \mathbb{E}[A_{i_1}A_{i_2}A_{i_3}A_{i_4}]\ \mathbb{E}[B_{j_1}B_{j_2}B_{j_3}B_{j_4}],\end{aligned}$$

其中，最后一个等式由对所有 $i, j \in [n]$. 而言 A_i 和 B_j 相互独立推出。若元组 (i_1, i_2, i_3, i_4) 中至少有一个指标只出现一次，则表达式 $\mathbb{E}[A_{i_1}A_{i_2}A_{i_3}A_{i_4}]$ 为 0，因为在这种情况下，我们可以利用对应因子 A_i 与其余因子的独立性，将该表达式写成$\mathbb{E}[A_i]$ 乘以其余因子乘积的期望。由于对所有 $i \in [n]$ 都有 $\mathbb{E}[A_i] = 0$，因此整个表达式的值为 0。

注意，如果其中一个因子恰好出现三次，那么另一个因子恰好出现一次。因此，剩下的情况是每个因子都出现偶数次。如果有两个因子，比如 A_i和 A_j，各出现两次，那么该表达式的值为 $\mathbb{E}[A_i^2A_j^2] = 1$。当 $i = j$时，该表达式的值也为 1。

因此，公式 (1) 中的每一项要么为 0，要么为 1。由对称性可知，只有当元组 (i_1, i_2, i_3, i_4) 中的每个索引以及元组 (j_1, j_2, j_3, j_4) 中的每个索引都出现偶数次时，该项才可能为 1。下面还需给出这类项数量的上界。选择两个不同的 i-索引有 $\binom{n}{2}$ 种方式，它们在 4 元组中的位置有 $\binom{4}{2} = 6$ 种可能。此外，在 4 元组中选择一个出现 4 次的索引有 n 种方式。因此，非零 $\mathbb{E}[A_{i_1}A_{i_2}A_{i_3}A_{i_4}]$ 的可能数量至多为 $3n^2$。对于 $\mathbb{E}[B_{j_1}B_{j_2}B_{j_3}B_{j_4}]$，同样的上界也成立。因此，公式 (1) 中等于 1 的项数至多为 $9n^4$。所以，该和的值至多为 $9n^4$。至此，命题 3.7 得证。

□

由于 U^2 是一个具有有限方差的非负随机变量，Paley-Zygmund 不等式给出，对于所有 $\theta \in [0, 1]$,

$$\Pr\left[U^2 > \theta\ \mathbb{E}[U^2]\right] \geq (1-\theta)^2\frac{(\mathbb{E}[U^2])^2}{\mathbb{E}[U^4]} \geq (1-\theta)^2\frac{m^2}{9n^4} \geq (1-\theta)^2\frac{(\gamma n^2)^2}{9n^4} = (1-\theta)^2\frac{\gamma^2}{9},$$

其中最后一个不等式 利用了引理中给出的界 $m \geq \gamma n^2$。最后，我们令 $\theta = 1/4$，得到 :

$$\Pr\left[|A^TMB| > \frac{\sqrt{m}}{2}\right] = \Pr\left[|U| > \frac{\sqrt{m}}{2}\right] = \Pr\left[U^2 > \frac{m}{4}\right] \geq \frac{3^2}{4^2}\frac{\gamma^2}{9} = \frac{\gamma^2}{16},$$

从而完成 引理 1.4的证明。 □

3.3 使用外积查询的重构攻击

为简化本节中的符号表示，我们将数据集和外积查询表示为向量。形式上，X 表示原始敏感数据集的向量化结果，即 $\{0,1\}^{n^2}$. 中的一个向量。对于外积查询（A, B），我们令 $Q \in \{0,1\}^{n^2}$ 表示 $A \otimes B$的向量化结果，即 A和 B的外积。（换言之， Q 是 A 和 B的克罗内克积。）于是，该查询的答案就是点积 $Q \cdot X$.

在本节中，我们定义并分析攻击者的算法 $\mathcal{C}$，并完成定理 3.1的证明。攻击者 $\mathcal{C}$ 在敏感数据集 X 上运行第 3.1 节 中的算法 $\mathcal{B}$，并使用一组 k 个随机外积查询 Q_1、……、Q_k，得到答案 a_1、……、a_k.。对于所有 $\ell \in [k]$，如果答案 a_ℓ 对于数据集 Y满足 $|Q_\ell \cdot Y - a_\ell| \leq \frac{\sqrt{\gamma}n}{4}$，我们就称其为准确答案；否则，称 a_ℓ 对于 Y是不准确的。攻击者 $\mathcal{C}$输出任意一个数据集$Y^* \in \{0,1\}^{n^2}$，使得 a_1、……、a_k 中至多有 $\frac{\gamma^2 k}{64}$ 个答案对于 Y^*. 是不准确的。根据引理 3.3， X 满足这一要求的概率至少为 $\frac{5}{6}$。若这一事件发生，算法 $\mathcal{C}$ 将能够输出某个 Y^*.（否则，攻击失败。）

接下来，我们分析攻击。令 $\|X - Y\|_1$ 表示数据集 X 和 Y 之间的汉明距离。若数据集 Y是坏的，则称其为坏数据集；也就是说，它与 X 在超过 γn^2 个条目上不同。我们将证明， $\mathcal{C}$ 不太可能选择坏数据集作为 Y^*.。固定一个坏数据集 Y。令 $M = X - Y$，并注意到 M 有 $m > \gamma n^2$ 个非零条目。若查询集 $\{Q_1, \ldots, Q_k\}$ 能够捕获数据集 Y，则称该查询集为捕获该数据集；也就是说，$(|Q_1 \cdot M|$、……、$|Q_k \cdot M|)$ 中超过 $\frac{\gamma^2 k}{32}$ 个条目大于 $\frac{\sqrt{\gamma}n}{2}$。

引理 3.8。假设攻击者 $\mathcal{C}$ 进行了 $k = \frac{128n^2}{\gamma^2}$ 次均匀随机外积查询。那么，存在一个未被攻击者的查询集捕获的坏数据集的概率至多为 $\frac{1}{6}$。

证明。考虑一组 k 均匀随机外积查询 $\{Q_\ell\}_{\ell\in[k]}$。固定一个坏数据集 Y。于是 $\|X - Y\|_1 > \gamma n^2$。令 $M = X - Y$.

对于每个 $\ell \in [k]$，令 $\chi_\ell = 1$ 如果 $|Q_\ell \cdot M| > \frac{\sqrt{\gamma}n}{2}$，否则令 $\chi_\ell = 0$。此外，令 $\chi = \sum_{\ell=1}^{k} \chi_\ell$. 根据定义，差向量 $M = X - Y$ 具有多于 γn^2 个非零条目。根据 引理 1.4中的反集中界，$\Pr\left[|Q_\ell \cdot M| > \frac{\sqrt{\gamma}n}{2}\right] \geq \frac{\gamma^2}{16}$。因此，$\mathbb{E}[\chi_\ell] \geq \frac{\gamma^2}{16}$ 根据切尔诺夫界（见定理 A.4），对于 $k = \frac{128n^2}{\gamma^2}$ 以及对于 $n \geq 3$，我们有：

$$\Pr\left[\chi \leq \frac{\gamma^2 \cdot k}{32}\right] \leq \exp\left(-\frac{\gamma^2 k}{128}\right) = \exp\left(-n^2\right) < \frac{1}{6 \cdot 2^{n^2}}.$$

因此，集合 $\{Q_\ell\}_{\ell\in[k]}$ 未能捕获每个特定不良数据集的概率至多为 $\frac{1}{6\cdot 2^{n^2}}$。对至多 2^{n^2} 个不良数据集应用并集界，存在某个未被攻击者的查询捕获的不良数据集的概率至多为 1/6。

□

引理 3.9 (重构引理)。如果算法 $\mathcal{B}$ 在除至多 $\frac{\gamma^2 k}{64}$ 个答案外的所有答案上，加性误差至多为 $\frac{\sqrt{\gamma}n}{4}$，且它使用的查询集能捕获所有坏数据集 Y，那么重构攻击是成功的；也就是说，攻击者 $\mathcal{C}$ 输出的 Y^* 与 X 至多有 γn^2 个条目不同，即 $\|X - Y^*\|_1 \leq \gamma n^2$。

证明。根据引理的第一个前提，数据集 X 与答案 a_ℓ 中至多 $\frac{\gamma^2 k}{64}$ 个“意见不一致”。因此，攻击者 $\mathcal{C}$ 必然会输出某个数据集 Y^*。为导出矛盾，假设 Y^* 是不良数据集。令 $\{Q_\ell\}_{\ell\in[k]}$ 为 $\mathcal{B}$ 选择的查询集。令 $M^* = X - Y^*$ 为差向量。根据三角不等式，$|Q_\ell M^*| = |Q_\ell X - Q_\ell Y^*| \leq |Q_\ell X - a_\ell| + |Q_\ell Y^* - a_\ell|$。根据引理中的第一个假设，对于除至多 $\frac{\gamma^2 k}{64}$ 个查询之外的所有查询，$|Q_\ell X - a_\ell| \leq \frac{\sqrt{\gamma}n}{4}$。根据攻击 $\mathcal{C}$ 的描述，输出 Y^* 满足：对于除至多 $\frac{\gamma^2 k}{64}$ 个查询之外的所有查询，$|Q_\ell Y^* - a_\ell| \leq \frac{\sqrt{\gamma}n}{4}$。因此，对于除至多 $\frac{\gamma^2 k}{32}$ 个查询之外的所有查询，$|Q_\ell M^*| \leq |Q_\ell X - a_\ell| + |Q_\ell Y^* - a_\ell| \leq \frac{\sqrt{\gamma}n}{2}$。由于 $\{Q_\ell\}_{\ell\in[k]}$ 能捕获所有不良数据集，因此它尤其能捕获 Y^*，因为 Y^* 是不良数据集。根据捕获的定义，对于超过 $\frac{\gamma^2 k}{32}$ 个值 $Q_\ell M^*$，$|Q_\ell M^*| > \frac{\sqrt{\gamma}n}{2}$。由此得出矛盾，说明 Y^* 是良好数据集。

□

证明 定理 3.1 的最后一个要点是下面这个引理，它基于 [De12] 的一个论证。任何输出其秘密数据集较大部分的算法，都肯定不具备隐私性，无论采用何种合理的隐私定义。 引理 3.10 指出，这样的算法不满足差分隐私。

引理 3.10。设 $\mathcal{C}$ 是一个算法，其输入为秘密数据集 X（属于 $\{0,1\}^N$），输出为同一集合中的一个向量 $\{0,1\}^N$。若 $\mathcal{C}$ 满足 (ε, δ)- 差分隐私，且 X 在 $\{0,1\}^N$ 中均匀分布，则

$$\mathbb{E}\left[\|\mathcal{C}(X) - X\|_1\right] \geq e^{-\varepsilon}\left(\tfrac{1}{2} - \delta\right)N\,.$$

上面的引理 3.10 仅对 $\|\mathcal{C}(X) - X\|_1$ 的期望给出了界。 [De12] 中更精巧的论证得出了紧得多的集中性结果。这里我们采用较简单的版本，因为它便于进行自成一体的阐述。

证明。固定一个索引 $i \in [N]$ 和一个比特 $r \in \{0,1\}$。令 $X_{i\to r}$ 表示将 X 的第 i 个条目替换为比特 r 后得到的向量。

考虑这对随机变量（X，$\mathcal{C}(X)$）。由于 $\mathcal{C}$ 满足（ε，δ）-差分隐私，因此它的分布与这对变量（$X_{i\to R}$，$\mathcal{C}(X)$）相似，其中 R 是一个与其他值独立的均匀随机比特。具体而言，对于任意事件 $E \subseteq \{0,1\}^N \times \{0,1\}^N$，

$$\Pr[(X_{i\to R}, \mathcal{C}(X)) \in E] \leq e^{\varepsilon} \Pr[(X, \mathcal{C}(X)) \in E] + \delta.$$

将该不等式应用于事件 $E_i = \{(x,\ y):\ x_i \neq y_i\}$，可得

$$\tfrac{1}{2} = \Pr[\mathcal{C}(X)_i \neq R] \leq e^{\varepsilon} \Pr[\mathcal{C}(X)_i \neq X_i] + \delta \quad \text{and thus} \quad \Pr[\mathcal{C}(X)_i \neq X_i] \geq e^{-\varepsilon}(\tfrac{1}{2} - \delta)\,.$$

汉明距离 $\|\mathcal{C}(X) - X\|_1$ 是事件 $\mathcal{C}(X)_i \neq X_i$ 对应的指示随机变量之和。根据期望的线性性，汉明距离的期望至少为 $e^{-\varepsilon}\left(\frac{1}{2} - \delta\right)N$。

□

最后，我们使用 引理 3.3 和 3.8来 3.10 完成主定理的证明 .

定理的证明 *3.1*。 令 $\gamma = \frac{1}{9}$. 反设，对于定理陈述中的某些 ε 和 δ，存在一个 $(\varepsilon,\ \delta)$-LEDP 算法 $\mathcal{A}$，对每个 $3n$ 节点图，都能以至少 的概率将图中的三角形数量近似到加性误差 $\alpha = \frac{\sqrt{\gamma n^2}}{20}$ $1 - \frac{\gamma^2}{9\cdot 128}$ 以内。 $1 - \frac{1}{3^6 \cdot 2^7}$ 3 3, $2\varepsilon\ 2\delta$ $\mathcal{B}$ X

于是，根据引理 。，存在一个 (，)-差分隐私算法，对于每个秘密数据集和任意一组 k 个外积查询，以至少 5 6 的概率，在至多 $\frac{\gamma^2 k}{64}$ 个查询中给出不准确的答案（即加性误差超过 $\frac{\sqrt{\gamma n}}{4}$）。根据引理 3.8，攻击者 $\mathcal{C}$ 选取的一组 $k = \frac{128n^2}{\gamma^2}$ 个随机外积查询未能捕获所有不良数据集的概率至多为 16。根据并集界，以至少 23 的概率，攻击者 $\mathcal{C}$ 满足引理 3.9 的前提，且所选查询能够捕获所有不良数据集。因此，以至少 23 的概率，攻击者 $\mathcal{C}$ 输出一个数据集 Y^*，它在至少 $(1-\gamma)n^2$ 个条目上与 X 一致。因此，期望汉明距离 $\mathbb{E}[\|X - Y^*\|_1]$ 至多为 $\frac{2}{3}\cdot\gamma n^2 + \frac{1}{3}n^2 = \frac{1+2\gamma}{3}n^2$。当 $\gamma = \frac{1}{9}$ 时，期望距离小于 $0.41n^2$。

回忆一下，攻击者 $\mathcal{C}$ 在秘密数据集 X 上运行 $(2\varepsilon, 2\delta)$-差分隐私算法 $\mathcal{B}$，然后对 $\mathcal{B}$的输出进行后处理。因此，$\mathcal{C}$ 满足 $(2\varepsilon, 2\delta)$-差分隐私，我们可以应用引理 3.10 得出，汉明距离 $\mathbb{E}\|\mathcal{C}(X) - X\|_1$ 的期望至多为 $e^{-2\varepsilon}(\frac{1}{2} - 2\delta)n^2$。由于根据假设，$\varepsilon \le 1/20$ 和 $\delta \le 1/100$，我们有 $\mathbb{E}\|\mathcal{C}(X) - X\|_1 \ge 0.43n^2$。这与上述 $0.41n^2$ 的上界矛盾。 □

4 非交互式三角形计数

在本节中，我们证明定理 1.2。在算法 1中，我们给出一种基于随机响应的简单算法，用于对图中的三角形数量进行无偏估计。我们用 $\binom{[n]}{\ell}$ 表示无序的 ℓ元顶点组的集合。设（固定）变量 $\mathbb{1}_{\{i,j,k\}}$ 在 $\{i,j,k\}$ 是 G 中的一个三角形时取 1，否则取 0。类似地，设 $\mathbb{1}_{\{i,j\}} = 1$若 $\{i,j\} \in E$则取 $\mathbb{1}_{\{i,j\}} = 0$，否则取 $\mathbb{1}_{\{i,j,k\}} = \mathbb{1}_{\{i,j\}}\cdot\mathbb{1}_{\{i,k\}}\cdot\mathbb{1}_{\{j,k\}}$

我们后处理过程的主要思路是对随机响应返回的带噪边进行重新缩放，使得每条边 $\{i,j\}$ 对应的重新缩放后的随机变量 $Y_{\{i,j\}}$ 的期望恰好为 $\mathbb{1}_{\{i,j\}}$.

算法 1 通过随机响应进行三角形计数

输入： 由 $n \times n$ 邻接矩阵 A 表示的图 G $G = ([n],\ E)$，其条目为 a_{ij}，$\varepsilon > 0$。**输出：** 图 G中的近似三角形数量。

1: **对于** $i = 1$ 到 n **执行**2: **发布** $(X_{\{i,i+1\}},\ \cdots,\ X_{\{i,n\}})$ ← 随机响应$_\varepsilon$($[a_{i(i+1)}, \ldots, a_{in}]$)。 3: 对于所有 $\{i,j\} \in \binom{[n]}{2}$，设置 $Y_{\{i,j\}} \leftarrow \frac{X_{\{i,j\}}\cdot(\ e^\varepsilon+1\ -1\ )}{e^\varepsilon-1}$。 4: 对于所有 $\{i,j,k\} \in \binom{[n]}{3}$，设置 $Z_{\{i,j,k\}} \leftarrow Y_{\{i,j\}}\cdot Y_{\{j,k\}}\cdot Y_{\{i,k\}}$。 5: **返回** $\widehat{T} \leftarrow \sum_{\{i,j,k\}} Z_{\{i,j,k\}}$。 $\{i,j,k\}\in\binom{[n]}{3}$

算法 1 是 ε-LEDP，由 定理 2.6给出。我们首先证明，该算法提供了三角形数量的无偏估计。

引理 4.1。 算法 *1* 返回输入图中三角形数量的无偏估计 .

证明。 令所有随机变量的定义均与算法 1相同。我们计算它们的期望。随机变量 $X_{\{i,j\}}$ 是图中由随机响应返回的含噪边 $\{i,j\}$ 是否存在的指示变量。若边 $\{i,j\} \in E$，则 $\mathbb{E}\left[X_{\{i,j\}}\right] = \frac{e^\varepsilon}{e^\varepsilon+1}$；否则，$\mathbb{E}\left[X_{\{i,j\}}\right] = \frac{1}{e^\varepsilon+1}$。现在，我们计算缩放后随机变量 $Y_{\{i,j\}} = \frac{X_{\{i,j\}}\cdot(\ e^\varepsilon+1\ -1)}{e^\varepsilon-1}$的期望。若 $\{i,j\} \in E$，则期望为 $\mathbb{E}\left[Y_{\{i,j\}}\right] = \frac{\mathbb{E}[X_{\{i,j\}}]\cdot(\ e^\varepsilon+1\ -1\ =)(1-\frac{1}{e^\varepsilon+1})\cdot(e^\varepsilon+1)}{e^\varepsilon-1} = 1$。若 $\{i,j\} \notin E$，则期望为 $\mathbb{E}\left[Y_{\{i,j\}}\right] = \frac{\frac{1}{e^\varepsilon+1}\cdot(e^\varepsilon+1\ -1}{e^\varepsilon-1} = 0$。因此，$Y_{\{i,j\}}$ 的期望值为 $\mathbb{1}_{\{i,j\}}$，如所期望。

接下来，根据 $Z_{\{i,j,k\}}$ 的定义以及随机变量 $Y_{\{i,j\}}$ 的相互独立性，

$$\mathbb{E}[Z_{\{i,j,k\}}] = \mathbb{E}[Y_{\{i,j\}}]\cdot\mathbb{E}[Y_{\{j,k\}}]\cdot\mathbb{E}[Y_{\{i,k\}}] = \mathbb{1}_{\{i,j\}}\cdot\mathbb{1}_{\{i,k\}}\cdot\mathbb{1}_{\{j,k\}} = \mathbb{1}_{\{i,j,k\}}.$$

^最后，我们计算三角形估计器 T. 的期望。令 T 为图 G G 中的三角形数量。^注意，$T = \sum_{\{i,j,k\} \in \binom{[n]}{3}} \mathbb{1}_{\{i,j,k\}}$. 根据 T 的定义以及期望的线性性，

$$\mathbb{E}\left[\widehat{T}\right] = \mathbb{E}\Big[\sum_{\{i,j,k\}\in\binom{[n]}{3}} Z_{\{i,j,k\}}\Big] = \sum_{\{i,j,k\}\in\binom{[n]}{3}} \mathbb{E}\Big[Z_{\{i,j,k\}}\Big] = \sum_{\{i,j,k\}\in\binom{[n]}{3}} \mathbb{1}_{\{i,j,k\}} = T, \tag{2}$$

^完成证明 T 是无偏的。 □

^我们计算由 T 算法算法 1返回的非交互式三角形计数估计器的方差。

引理 4.2。 *算法 1* 返回一个近似三角形计数 $\widehat{T}$，满足 $\mathrm{Var}\left[\widehat{T}\right] = \Theta\left(\frac{C_4}{\varepsilon^2} + \frac{n^3}{\varepsilon^6}\right)$ 其中C_4 是4 长度的环在 G中的数量。

证明。 我们计算算法 1中定义的所有随机变量的方差。首先，由于 $X_{\{i,j\}}$ 是伯努利变量，$\mathrm{Var}[X_{\{i,j\}}] = \frac{e^\varepsilon}{e^\varepsilon+1()^2}$。 然后，$\mathrm{Var}[Y_{\{i,j\}}] = \frac{(e^\varepsilon+1)^2}{(e^\varepsilon-1)^2} \cdot \mathrm{Var}[X_{\{i,j\}}] = \frac{e^\varepsilon}{e^\varepsilon-1()^2}$。 为了计算 $\mathrm{Var}[Z_{\{i,j,k\}}]$，我们计算 $\mathbb{E}[Y^2_{\{i,j\}}] = \mathrm{Var}[Y_{\{i,j\}}] + \mathbb{E}[Y_{\{i,j\}}]^2 = \frac{e^\varepsilon}{e^\varepsilon-1()^2} + \mathbb{1}_{\{i,j\}} = \Theta\left(\frac{1}{\varepsilon^2}\right)$。 然后

$$\mathrm{Var}[Z_{\{i,j,k\}}] = \mathbb{E}[Z^2_{\{i,j,k\}}] - \mathbb{E}[Z_{\{i,j,k\}}]^2 = \mathbb{E}[Y^2_{\{i,j\}}] \cdot \mathbb{E}[Y^2_{\{j,k\}}] \cdot \mathbb{E}[Y^2_{\{i,k\}}] - \mathbb{1}^2_{\{i,j,k\}} = \Theta\left(\frac{1}{\varepsilon^6}\right). \tag{3}$$

^为计算 $\mathrm{Var}[T]$，我们考虑如下变量替换： $U_{\{i,j,k\}} = Z_{\{i,j,k\}} - \mathbb{1}_{\{i,j,k\}}$。 注意到$\mathbb{E}[U_{\{i,j,k\}}] = 0$，且由于 $\mathbb{1}_{\{i,j,k\}}$ 是固定的，$\mathrm{Var}\left[U_{\{i,j,k\}}\right] = \mathrm{Var}\left[Z_{\{i,j,k\}}\right]$。 类似地，由于 T 是固定的，$\mathrm{Var}\left[\widehat{T}-T\right] = \mathrm{Var}\left[\widehat{T}\right]$，

$$\mathrm{Var}\left[\widehat{T}\right] = \mathrm{Var}\left[\widehat{T} - T\right] = \mathrm{Var}\Big[\sum_{\{i,j,k\}\in\binom{[n]}{3}} \left(Z_{\{i,j,k\}} - \mathbb{1}_{\{i,j,k\}}\right)\Big] = \mathrm{Var}\Big[\sum_{\{i,j,k\}\in\binom{[n]}{3}} U_{\{i,j,k\}}\Big]. \tag{4}$$

为分析 $\mathrm{Var}\Big[\sum_{\{i,j,k\}\in\binom{[n]}{3}} U_{\{i,j,k\}}\Big]$，注意到两个变量 $U_{\{i,j,k\}}$ 和 $U_{\{r,q,s\}}$ 的协方差为零，当三元组 $\{i,j,k\}$ 和 $\{r,q,s\}$ 至多共享一个顶点时（因为它们没有共享边）。令 P_2 为变量对的集合 $U_{\{i,j,k\}}$，其中对应的三元组恰好共享两个节点。根据 公式（4），

$$\begin{aligned}
\mathrm{Var}\left[\widehat{T}\right] &= \sum_{\{i,j,k\}\in\binom{[n]}{3}} \mathrm{Var}\left[U_{\{i,j,k\}}\right] + \sum_{U_{\{i,j,k\}},U_{\{j,k,l\}}\in P_2} \mathbb{E}\left[U_{\{i,j,k\}} \cdot U_{\{j,k,l\}}\right] \\
&= \sum_{\{i,j,k\}\in\binom{[n]}{3}} \Theta\left(\frac{1}{\varepsilon^6}\right) + \sum_{U_{\{i,j,k\}},U_{\{j,k,l\}}\in P_2} \mathbb{E}[(Z_{\{i,j,k\}} - \mathbb{1}_{\{i,j,k\}})(Z_{\{j,k,l\}} - \mathbb{1}_{\{j,k,l\}})] \\
&= \Theta\left(\frac{n^3}{\varepsilon^6}\right) + \sum_{U_{\{i,j,k\}},U_{\{j,k,l\}}\in P_2} \left(\mathbb{E}[Y_{\{i,j\}} \cdot Y^2_{\{j,k\}} \cdot Y_{\{i,k\}} \cdot Y_{\{l,j\}} \cdot Y_{\{l,k\}}] - \mathbb{1}_{\{i,j,k\}}\mathbb{1}_{\{j,k,l\}}\right) \\
&= \Theta\left(\frac{n^3}{\varepsilon^6}\right) + \sum_{U_{\{i,j,k\}},U_{\{j,k,l\}}\in P_2} \left(\mathbb{1}_{\{i,j\}} \cdot \mathbb{E}[Y^2_{\{j,k\}}] \cdot \mathbb{1}_{\{i,k\}} \cdot \mathbb{1}_{\{l,j\}} \cdot \mathbb{1}_{\{l,k\}} - \mathbb{1}_{\{i,j,k\}}\mathbb{1}_{\{j,k,l\}}\right) \\
&= \Theta\left(\frac{n^3}{\varepsilon^6} + \frac{C_4}{\varepsilon^2}\right).
\end{aligned} \tag{5}$$

该 C_4 项公式 (5) 源于以下事实： $\mathbb{E}[Y^2_{\{j,k\}}] = \Theta\left(\frac{1}{\varepsilon^2}\right)$，并且 $\mathbb{1}_{\{i,j\}} \cdot \mathbb{1}_{\{i,k\}} \cdot \mathbb{1}_{\{l,j\}} \cdot \mathbb{1}_{\{l,k\}} = 1$当且仅当（$i$， j，k， l）是一个长度为四的环。公式 (5) 给出了我们的最终界。 □

定理的证明 *定理 1.2*。 定理 1.2 直接由以下结果得出：引理 4.1 和4.2其中，通过方差可以得到以常数概率成立的加性误差。 □

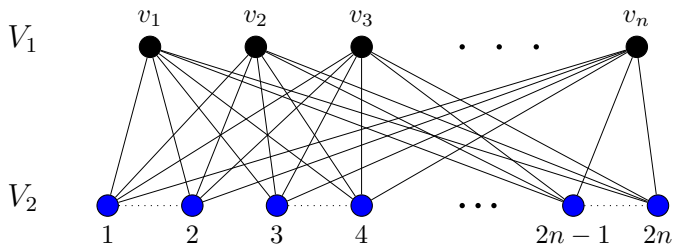


图 3：交互式 $\Omega(n^{3/2})$ 下界的一个实例由一个完全二分图构成，其两部分 V_1, V_2 的大小分别为 n 和 $2n$；此外，当且仅当秘密输入位 $X_i = 1$为真时，每对 $\{2i-1, 2i\}$ 之间都有一条边。

5 交互式下界

在本节中，我们给出每个 $\binom{n^{3/2}}{\varepsilon}$ 算法在加性误差方面的 Ω 下界 ε，该算法基于 LEDP 估计图中的三角形数量，正式表述见 定理 1.3。我们将问题归约到 LDP 模型下的求和计算问题。

定义 5.1 (求和函数)。 令 SUM_n 为以下函数。对于所有 $x_1, \ldots, x_n \in \{0,1\}$,$SUM_n(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i$.

已证明，该问题的加性误差下界为 $\Omega(\sqrt{n}/\varepsilon)$ [JMNR19，arxiv v2] 的定理 5.3。我们代入 $\alpha = \alpha_0/n$ 和 $\beta = \varepsilon\alpha_0/n$，得到以下引理。

引理 5.2 ([CSS12, BNO08, JMNR19])。 存在一个常数 $c > 0$，使得对于每个 $\varepsilon \in (0,1)$、$n \in \mathbb{N}$、$\alpha_0 \in (0, n]$ 以及 $\delta \in \left[0, \frac{1}{10^5} \cdot n^3 \, e^{\varepsilon^3 \alpha_0^2 \ln(n^2/\varepsilon\alpha_0)}\right]$，若 $\mathcal{B}$ 是一个 (ε, δ)-*LDP* 算法，其中每个参与方 i 接收输入$x_i \in \{0,1\}$，且 $\mathcal{B}$ 以至少 $2/3$的概率将 SUM_n 估计到加性误差 α_0 以内，则 $\alpha_0 \geq c \cdot \sqrt{n}/\varepsilon$。

定理*定理 1.3*的证明。我们从 SUM_n 本地模型中的问题归约而来。根据引理 5.2，任何近似计算 SUM_n 且加性误差至多为 α_0 的算法（可能具有交互性，且成功概率为足够高的常数）都必须具有 $\alpha_0 = \Omega(\binom{\sqrt{n}}{\varepsilon})$。在我们的归约中，我们将把三角形计数算法的加性误差设为 $\alpha = \alpha_0 n$.

我们的归约是黑盒的。给定一个 SUM_n 实例，其中每个本地参与方持有向量 $(X_1, \ldots, X_n)$ 中的一个比特 X_i，参与方隐式构造如下图 G G。顶点集由两组节点组成，V_1 和 V_2，其中 V_1 的大小为 n，而 V_2 的大小为 $2n$。V_1 中的节点不包含任何秘密信息，可由任意本地参与方模拟。V_2 中的节点为 $[2n]$，每个参与方 $i \in [n]$ 负责模拟 V_2中的节点 $2i-1$ 和 $2i$。为构造 G的边集，我们在 V_1 和 V_2之间添加完全二分图的边。此外，V_2 中的每对节点 $(2i-1, 2i)$ 之间有一条边，当且仅当 $x_i = 1$成立。图 3 展示了示意图。

令 $S = x_1 + \ldots x_n$. 注意到，G 中的任意三角形都必须有两个顶点位于 V_2 中，并且一对匹配节点之间有一条边。每条这样的边恰好贡献 n 个三角形。因此，G 中三角形的总数为 $T = Sn$.

为导出矛盾，假设存在一个 (ε, δ)-LEDP 算法 $\mathcal{A}$，它对边数量的估计误差不超过 $o(\binom{n^{3/2}}{\varepsilon})$。我们在 G 上运行该算法。根据构造，参与方 i 可以模拟分配给它的两个节点，而任何人都可以模拟 V_1 中的节点。当算法得到三角形数量的估计值 T 后，它输出 $\hat{S} = \widehat{T}/n$. 如果 $\widehat{T} = T \pm o(\binom{n^{3/2}}{\varepsilon})$，则 $\hat{S} = \widehat{T}/n = T/n \pm o(\binom{\sqrt{n}}{\varepsilon}) = S \pm o(\binom{\sqrt{n}}{\varepsilon})$，这与引理 5.2矛盾。

此外，如果 $\mathcal{A}$是（ε，δ）-LEDP，那么该归约算法相对于秘密数据集 X是（ε，δ）-LDP。然而，后者与引理 5.2矛盾。因此，$\mathcal{A}$ 不可能存在。 □

致谢。 T.E. 得到了美国国家科学基金会 TRIPODS 项目（资助编号 DMS-2022448）和波士顿大学的支持。本研究的部分工作是在作者隶属于波士顿大学和麻省理工学院期间完成的。A.S. 部分得到了美国国家科学基金会资助项目 CCF-1763786 和 CNS-2120667，以及谷歌和苹果颁发的教师奖项的支持。我们感谢 Iden Kalemaj 和 Satchit Sivakumar 对我们结果的初始版本提出的宝贵意见。

参考文献

[AHD18] Mohammad Al Hasan 和 Vachik S Dave。大型网络中的三角形计数：综述。《*Wiley* 跨学科评论：数据挖掘与知识发现》，8(2):e1226，2018。[AKK19] Sepehr Assadi、Michael Kapralov 和 Sanjeev Khanna。通过边采样计数任意子图的简单亚线性时间算法。发表于*ITCS*，*LIPIcs*第 124 卷，页码 6:1–6:20。施洛斯达格施图尔—莱布尼茨信息中心，2019。[App17] 苹果差分隐私团队。大规模隐私差分学习。2017。[BBCG08] Luca Becchetti、Paolo Boldi、Carlos Castillo 和 Aristides Gionis。在海量图中进行局部三角形计数的高效半流式算法。发表于第 *14* 届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，页码 16–24，2008。[BBDS13] Jeremiah Blocki、Avrim Blum、Anupam Datta 和 Or Sheffet。通过受限敏感度对社交网络进行差分隐私数据分析。载于 Robert D. Kleinberg 编，《理论计算机科学创新，*ITCS '13*，美国加利福尼亚州伯克利，*2013* 年 *1* 月 *9–12* 日》，页码 87–96。ACM，2013。[BEL+22] Amartya Shankha Biswas、Talya Eden、Quanquan C. Liu、Ronitt Rubinfeld 和 Slobodan Mitrovic。小子图计数的大规模并行算法。245:39:1–39:28，2022。[BEM+17] Andrea Bittau、 ́Ulfar Erlingsson、Petros Maniatis、Ilya Mironov、Ananth Raghunathan、David Lie、Mitch Rudominer、Ushasree Kode、Julien Tinnes 和 Bernhard Seefeld。Prochlo：为众包分析提供强隐私保护。发表于第 *26* 届操作系统原理研讨会论文集，SOSP '17，页码 441–459，美国纽约州纽约，2017。美国计算机协会。[BKV17] Prakash Balachandran、Eric D. Kolaczyk 和 Weston D. Viles。低发生率测量误差对大型网络中子图计数的传播影响。《机器学习研究杂志》，18(61):1–33，2017。[BNO08] Amos Beimel、Kobbi Nissim 和 Eran Omri。分布式隐私数据分析：同时解决如何分析和分析什么。发表于国际密码学年会，页码 451–468。施普林格，2008。[BS16] Mark Bun 和 Thomas Steinke。集中差分隐私：简化、扩展与下界。第 9985 卷，页码 635–658，2016 年 11 月。[CEI+21] Justin Y Chen、Talya Eden、Piotr Indyk、Honghao Lin、Shyam Narayanan、Ronitt Rubinfeld、Sandeep Silwal、Tal Wagner、David Woodruff 和 Michael Zhang。利用预测在图流中进行三角形和四环计数。发表于国际学习表征会议，2021。[Che52] Herman Chernoff。基于观测值之和的假设检验渐近效率度量。《数理统计年刊》，页码 493–507，1952。[CJK+18] Graham Cormode、Somesh Jha、Tejas Kulkarni、Ninghui Li、Divesh Srivastava 和 Tianhao Wang。大规模隐私保护：本地差分隐私的实践。发表于 *2018* 年国际数据管理会议论文集，SIGMOD '18，页码 1655–1658，美国纽约州纽约，2018。美国计算机协会。

[CKY22] Jinyuan Chang、Eric D. Kolaczyk 和 Qiwei Yao。噪声网络中的子图密度估计。《美国统计学会期刊》，117(537):361–374，2022。[CSS12] TH Hubert Chan、Elaine Shi 和 Dawn Song。差分隐私多方聚合的最优下界。发表于欧洲算法研讨会，页码 277–288。施普林格，2012。[CZ13] Shixi Chen 和 Shuigeng Zhou。递归机制：迈向节点差分隐私和不受限制的连接。载于 Kenneth A. Ross、Divesh Srivastava 和 Dimitris Papadias 编，《ACM SIGMOD 数据管理国际会议论文集》，SIGMOD 2013，美国纽约州纽约，2013 年 6 月 22–27 日，页码 653–664。ACM，2013。[De12] Anindya De。差分隐私中的下界。发表于密码学理论会议，页码 321–338。施普林格，2012。[DJW13] John Duchi、Michael Jordan 和 Martin Wainwright。本地隐私与统计极小极大速率。发表于电气电子工程师学会计算机科学基础研讨会，FOCS '13，页码 429–438，美国加利福尼亚州伯克利，2013。https://arxiv.org/abs/1302.3203。[DJW+18] John C Duchi、Michael I Jordan、Martin J Wainwright 等。本地隐私估计的极小极大最优程序。《美国统计学会期刊》，113(521):182–201，2018。[DKM+06] Cynthia Dwork、Krishnaram Kenthapadi、Frank McSherry、Ilya Mironov 和 Moni Naor。我们的数据，我们自己：通过分布式噪声生成实现隐私保护。载于 Serge Vaudenay 编，《密码学进展——EUROCRYPT 2006：第 25 届密码技术理论与应用国际年会，俄罗斯圣彼得堡，2006 年 5 月 28 日至 6 月 1 日，论文集》，计算机科学讲义第 4004 卷，页码 486–503。施普林格，2006。[DKY17] Bolin Ding、Janardhan Kulkarni 和 Sergey Yekhanin。以隐私保护方式收集遥测数据。发表于第 31 届神经信息处理系统国际会议论文集，NIPS'17，页码 3574–3583，美国纽约州雷德胡克，2017。Curran Associates Inc.。[DL09] Cynthia Dwork 和 Jing Lei。差分隐私与稳健统计。发表于第 41 届 ACM 理论计算研讨会年会论文集，STOC '09，页码 371–380，美国纽约州纽约，2009。美国计算机协会。[DLR+22] Laxman Dhulipala、Quanquan C. Liu、Sofya Raskhodnikova、Jessica Shi、Julian Shun 和 Shangdi Yu。通过局部可调整图算法实现差分隐私：k-core 分解、低出度排序和最稠密子图。发表于第 63 届电气电子工程师学会计算机科学基础年会，FOCS 2022，美国科罗拉多州丹佛，2022 年 10 月 31 日至 11 月 3 日，页码 754–765。电气电子工程师学会，2022。[DMNS06] Cynthia Dwork、Frank McSherry、Kobbi Nissim 和 Adam Smith。为隐私数据分析校准噪声与敏感度。发表于第三届密码学理论会议论文集，TCC'06，页码 265–284，德国柏林、海德堡，2006。施普林格出版社。[DMT07] Cynthia Dwork、Frank McSherry 和 Kunal Talwar。隐私的代价与 LP 解码的局限。发表于第 39 届 ACM 理论计算研讨会年会论文集，页码 85–94。ACM，2007。[DN03] Irit Dinur 和 Kobbi Nissim。在保护隐私的同时披露信息。载于 Frank Neven、Catriel Beeri 和 Tova Milo 编，《第二十二届 ACM SIGACT-SIGMOD-SIGART 数据库系统原理研讨会论文集》，美国加利福尼亚州圣迭戈，2003 年 6 月 9–12 日，页码 202–210。ACM，2003。[DRV10] Cynthia Dwork、Guy N. Rothblum 和 Salil Vadhan。增强与差分隐私。发表于 2010 年 IEEE 第 51 届计算机科学基础年会论文集，FOCS '10，页码 51–60，美国，2010。电气电子工程师学会计算机学会。

[DY08] Cynthia Dwork 和 Sergey Yekhanin。针对统计披露控制机制的新型高效攻击。载于《国际密码学年会》，第 469–480 页。施普林格，2008 年。[EGS03] Alexandre V. Evfimievski、Johannes Gehrke 和 Ramakrishnan Srikant。在隐私保护数据挖掘中限制隐私泄露。载于 Frank Neven、Catriel Beeri 和 Tova Milo（编），《第二十二届 ACM SIGACT-SIGMOD-SIGART 数据库系统原理研讨会论文集》，2003 年 6 月 9–12 日，美国加利福尼亚州圣迭戈，第 211–222 页。美国计算机协会，2003 年。[ELRS17] Talya Eden、Amit Levi、Dana Ron 和 C Seshadhri。亚线性时间内近似计算三角形数量。《SIAM 计算杂志》，46(5):1603–1646，2017 年。[EM02] Jean-Pierre Eckmann 和 Elisha Moses。共链曲率揭示万维网上隐藏的主题层。《美国国家科学院院刊》，99(9):5825–5829，2002 年。[EPK14] Ulfar Erlingsson、Vasyl Pihur 和 Aleksandra Korolova。Rappor：随机化、可聚合且保护隐私的序数响应。载于《2014 年 ACM SIGSAC计算机与通信安全会议论文集》，CCS '14，美国纽约州纽约，2014 年，第 1054–1067 页。美国计算机协会。[FDBV01] Ill´es J Farkas、Imre Der´enyi、Albert-L´aszl´o Barab´asi 和 Tamas Vicsek。“真实世界”图的谱：超越半圆定律。《物理评论 E》，64(2):026704，2001 年。[GLCZ18] Tianchong Gao、Feng Li、Yu Chen 和 XuKai Zou。基于 HRG 模型对在线社交网络进行本地差分隐私匿名化。《电气电子工程师学会计算社会系统汇刊》，5(4):1009–1020，2018 年。[GRU12] Anupam Gupta、Aaron Roth 和 Jonathan R. Ullman。迭代构造与隐私数据发布。载于 Ronald Cramer（编），《密码学理论——第九届密码学理论会议，TCC 2012》，意大利西西里岛陶尔米纳，2012 年 3 月 19–21 日，论文集，《计算机科学讲义》第 7194 卷，第 339–356 页。施普林格，2012 年。[IMC21] Jacob Imola、Takao Murakami 和 Kamalika Chaudhuri。图统计量的局部差分隐私分析。载于《第 30 届 USENIX 安全研讨会，USENIX Security 2021》，2021 年 8 月 11–13 日，第 983–1000 页，2021 年。[IMC22a] Jacob Imola、Takao Murakami 和 Kamalika Chaudhuri。本地差分隐私下高效通信的三角形计数。载于《第 31 届 USENIX 安全研讨会，USENIX Security 2022》，2022 年 8 月 10–12 日，2022 年。[IMC22b] Jacob Imola、Takao Murakami 和 Kamalika Chaudhuri。洗牌模型下差分隐私的三角形与 4-环计数。载于《2022 年 ACM SIGSAC计算机与通信安全会议论文集》，CCS '22，美国纽约州纽约，2022 年，第 1505–1519 页。美国计算机协会。[JMNR19] Matthew Joseph、Jieming Mao、Seth Neel 和 Aaron Roth。交互性在本地差分隐私中的作用。载于 David Zuckerman（编），《第 60 届电气电子工程师学会年度计算机科学基础研讨会，FOCS 2019》，美国马里兰州巴尔的摩，2019 年 11 月 9–12 日，第 94–105 页。电气电子工程师学会计算机学会，2019 年。[KLN+11] Shiva Prasad Kasiviswanathan、Homin K Lee、Kobbi Nissim、Sofya Raskhodnikova 和 Adam Smith。我们能以隐私保护方式学到什么？《SIAM 计算杂志》，40(3):793–826，2011 年。[KNRS13] Shiva Prasad Kasiviswanathan、Kobbi Nissim、Sofya Raskhodnikova 和 Adam D. Smith。使用节点差分隐私分析图。载于 Amit Sahai（编），《密码学理论——第十届密码学理论会议，TCC 2013》，日本东京，2013 年 3 月 3–6 日，论文集，《计算机科学讲义》第 7785 卷，第 457–476 页。施普林格，2013 年。

[KP17] John Kallaugher 和 Eric Price。用于三角形计数的混合采样方案。载于《年度 ACM-SIAM 离散算法研讨会论文集》，第 1778–1797 页，2017 年。[KRS13] Shiva Prasad Kasiviswanathan、Mark Rudelson 和 Adam D. Smith。线性重构攻击的威力。载于 Sanjeev Khanna 编，《第二十四届年度 ACM-SIAM 离散算法研讨会论文集》（SODA 2013），美国路易斯安那州新奥尔良，2013 年 1 月 6—8 日，第 1415–1433 页。SIAM，2013 年。[KRSU10] Shiva Prasad Kasiviswanathan、Mark Rudelson、Adam Smith 和 Jonathan Ullman。私密发布列联表的代价，以及相关行随机矩阵的谱。载于《第 42 届 ACM 计算理论研讨会论文集》（STOC '10），第 775–784 页。ACM，2010 年。[KRSY14] Vishesh Karwa、Sofya Raskhodnikova、Adam D. Smith 和 Grigory Yaroslavtsev。图结构的隐私分析。《ACM 数据库系统汇刊》，39(3)：22:1–22:33，2014 年。[KSK14] Vishesh Karwa、Aleksandra B. Slavkovi´c 和 Pavel Krivitsky。差分隐私指数随机图。载于《统计数据库中的隐私》，第 143–155 页。施普林格国际出版公司，2014 年。[MSOI+02] Ron Milo、Shai Shen-Orr、Shalev Itzkovitz、Nadav Kashtan、Dmitri Chklovskii 和 Uri Alon。网络模体：复杂网络的简单构件。《科学》，298(5594)：824–827，2002 年。[MV20] Andrew McGregor 和 Sofya Vorotnikova。数据流模型中的三角形和四环计数。载于 ACM SIGMOD-SIGACT-SIGAI 数据库系统原理研讨会论文集，第 445–456 页，2020 年。[MVV16] Andrew McGregor、Sofya Vorotnikova 和 Hoa T Vu。用于数据流中三角形计数的更优算法。载于《第 35 届 ACM SIGMOD-SIGACT-SIGAI 数据库系统原理研讨会论文集》，第 401–411 页，2016 年。[NRS07] Kobbi Nissim、Sofya Raskhodnikova 和 Adam D. Smith。私密数据分析中的平滑敏感度与采样。载于 David S. Johnson 和 Uriel Feige 编，《第 39 届年度 ACM 计算理论研讨会论文集》，美国加利福尼亚州圣迭戈，2007 年 6 月 11—13 日，第 75–84 页。ACM，2007 年。完整论文：http://www.cse.psu.edu/~asmith/pubs/NRS07.[PDBL16] Arnau Prat-P´erez、David Dominguez-Sal、Josep M. Brunat 和 Josep Llu´ís Larriba-Pey。三加三，合为一体：由三角形驱动的社区检测。《ACM 知识发现与数据挖掘汇刊》，10(3)：22:1–22:42，2016 年。[PDFV05] Gergely Palla、Imre Der´enyi、Ill´es Farkas 和 Tam´as Vicsek。揭示自然与社会复杂网络中重叠的社区结构。《自然》，435(7043)：814–818，2005 年。[PSKP14] Ha-Myung Park、Francesco Silvestri、U Kang 和 Rasmus Pagh。带有保证的 MapReduce 三角形枚举。载于《第 23 届 ACM 信息与知识管理国际会议论文集》，第 1739–1748 页，2014 年。[PT12] Rasmus Pagh 和 Charalampos E Tsourakakis。彩色三角形计数及其 MapReduce 实现。《信息处理快报》，112(7)：277–281，2012 年。[PZ32] REAC Paley 和 Antoni Zygmund。单位圆内解析函数札记。载于《剑桥哲学学会数学学报》，第 28 卷，第 266–272 页。剑桥大学出版社，1932 年。[QYY+17] Zhan Qin、Ting Yu、Yin Yang、Issa Khalil、Xiaokui Xiao 和 Kui Ren。利用本地差分隐私生成合成去中心化社交图。载于《2017 年 ACM SIGSAC计算机与通信安全会议论文集》（CCS '17），第 425–438 页，美国纽约州纽约，2017 年。美国计算机协会。

[RS16] Sofya Raskhodnikova 和 Adam D. Smith。图的差分隐私分析。载于算法百科全书，第 543–547 页。2016 年。[SXK+19] Haipei Sun、Xiaokui Xiao、Issa Khalil、Yin Yang、Zhan Qin、Hui (Wendy) Wang 和 Ting Yu。利用去中心化差分隐私从扩展局部视图中分析子图统计量。载于*2019* 年美国计算机协会 *SIGSAC* 计算机与通信安全会议论文集，CCS '19，第 703–717 页，美国纽约州纽约市，2019 年。美国计算机协会。[War65] Stanley L Warner。随机响应：一种消除回避回答偏差的调查技术。美国统计学会期刊，60(309):63–69，1965 年。[YHA+20] Qingqing Ye、Haibo Hu、Man Ho Au、Xiaofeng Meng 和 Xiaokui Xiao。迈向局部差分隐私通用图度量估计。载于*2020* 年电气电子工程师学会第 *36* 届国际数据工程会议（*ICDE*），第 1922–1925 页，2020 年。

A 附加工具

A.1 隐私工具

差分隐私在组合和后处理中得以保持。

定理 A.1（组合定理，改编自 [DMNS06、DL09、DRV10]）。设 $\varepsilon > 0$ 且 $\delta \in [0, 1)$。如果算法 R 运行 （ε，δ）-局部随机化器 R_1，……，R_k，那么 R 是一个 （$k\varepsilon$，$k\delta$）-局部随机化器。

n　　扩展定理指出，对差分隐私算法输出进行后处理的结果是 D　　P.

定理 A.2（后处理 [DMNS06、BS16]）。设 $\mathcal{A}$ 是一个 （ε，δ）-差分隐私算法，且 $\mathcal{P}$ 是从 $Range$（$\mathcal{A}$）到任意集合的任意（随机）映射。在 $\mathcal{A}$ 的输出上运行 $\mathcal{P}$ 的算法是（ε，δ）-差分隐私算法。

为完整起见，我们给出讨论中心化模型时所采用的全局敏感度定义。

定义 A.3（全局敏感度 [DMNS06]）。设 $d \in \mathbb{N}$ 和 $f: \mathcal{D} \to \mathbb{R}^d$ 是定义域为 $\mathcal{D}$上的函数。f 的 ℓ_1-敏感度为 $\Delta f = \max_{x,\, x'} \|f(x) - f(x')\|_1$，其中最大值取遍所有相邻的 x、$x' \in \mathcal{D}$，相邻的具体定义视情况而定。

A.2 集中界

定理 A.4（切尔诺夫界 [Che52]）。设 $X_1, \ldots, X_n$ 是取值于 $\{0, 1\}$. 中的独立随机变量。令 $X = \sum_{i=1}^{n} X_i$ 且 $\mu = \mathbb{E}[X]$。则对于任意 $\delta > 0$,

$$\Pr\left[X < (1-\delta)\mu\right] \leq \exp\left(-\frac{\delta^2 \cdot \mu}{2}\right).$$

定理 A.5 (Paley–Zygmund 不等式 [PZ32])。若 $Z \geq 0$ 是具有有限方差的随机变量，则对于所有 $0 \leq \theta \leq 1$,

$$\Pr[Z > \theta \cdot \mathbb{E}[Z]] \geq (1-\theta^2)\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]}.$$