一种用于流程挖掘隐私保护的轻量级加密方法

莫赫森·卡泽米安，IEEE会员；马库斯·赫尔弗特

arXiv:2304.03579v1 [cs.CR] 7 Apr 2023

摘要——商业智能、企业管理和数据科学领域的新型技术成果均以实时且复杂的虚拟网络为基础。大量组织之间共享数据，由此形成计算复杂度高的系统，这是当前商业网络的重要特征之一。业务流程的发现、一致性检查和改进均通过生成的事件日志实现。在此方面，工业领域流程挖掘中一个被忽视的挑战是隐私保护。为满足当前数字商业技术的需求，在计算复杂度较低的结构中保护数据隐私，本文提出了一种基于哈尔变换和私钥的新型轻量级加密方法。我们从密码学、计算复杂度和结构脆弱性方面，将所提出的方法与众所周知的同态密码系统和沃尔什-哈达玛加密（WHE）进行比较。分析结果表明，与上述两种密码系统相比，所提出的方法能够以显著更低的复杂度、更高的准确率对事件日志进行匿名化处理。

索引词——隐私与安全，流程挖掘，哈尔变换，数据加密，医疗保健。

一、引言

5G及后续技术的到来为所有公司和组织提供了许多令人着迷的可能性，使其智能化和数字化业务得以成为现实 [1]。这些进步旨在基于数字环境传输所需信息。在行业中，机器与制造执行系统之间，以及仓库机器人与相关管理系统之间传递信息，是该领域的两个示例 [2]。因此，在当前的数字化世界中，挖掘任何组织中的通信，以改进已执行流程至关重要。

流程挖掘（PM）指一系列方法，旨在利用生成的信息，从任何组织的已执行流程中获取有价值的见解 [3], [4]。流程挖掘技术涉及数据科学领域（即数据挖掘和预测分析等领域）以及流程科学领域（例如业务流程管理（BPM）[5], [6]），旨在根据事件日志发现瓶颈并提升系统的整体性能。

目前，流程挖掘技术已在工业界和学术界得到广泛发展。其中一个最重要的领域是医疗保健，它能够改善医疗服务。

M. Kazemian 任职于爱尔兰梅努斯大学的 Adapt 研究中心和创新价值研究所（IVI）（电子邮箱：mohsenkazemian@yazd.ac.ir）。M. Helfert 任职于爱尔兰梅努斯大学的 Adapt 研究中心、创新价值研究所和商学院（电子邮箱：markus.helfert@mu.ie）。

利益相关者可以从事件日志 [7] 中获得有价值的洞察，例如识别活动的实际顺序以及资源的参与情况。医疗保健领域的事件日志可能包含患者的敏感和关键健康信息，例如案例 ID、活动、时间戳、年龄、诊断和治疗代码。

在早期研究中，人们就意识到流程挖掘需要考虑隐私保护问题 [8]。然而，直到最近，流程挖掘领域通常都忽视了隐私问题。如今，随着实时和虚拟网络架构的出现，充分保护数据隐私与安全对于流程挖掘领域而言比以往任何时候都更加重要。在这一领域，一个重要目标是对数据日志进行加密，从而在存在不可信环境、接口和组织的情况下，确保在应用流程挖掘算法之前信息始终安全 [9]。鉴于当前新兴网络本身复杂得多，采用低复杂度算法开发隐私保护方法至关重要。鉴于安全级别、硬件架构、处理速度和数据量等可变参数，是选择加密方法和流程挖掘技术时的关键参数。

A. Related Work

隐私与安全保护相关议题在数据挖掘这一通用领域中早已广为人知。目前，已有一些文章初步尝试应对流程挖掘领域中与隐私保护相关的一些显著挑战[8], [10]。文献 [9] 提出了一种方法，能够在外包流程挖掘的同时，确保事件日志和已执行流程的机密性。这项工作提供的隐私保护方法非常薄弱，可能容易受到先进去匿名化方案的攻击。Saavedra 等人 [11] 提出了一种模型，旨在掩盖包含敏感值的事件日志（例如，删除 {12345678}中的最后 4 个字符）。然而，该方法存在一些缺点，例如操作复杂度高，以及在应用流程挖掘技术之前无法确定能否正确还原原始数据，因此难以接受。Liu 等人 [12] 提出了一种基于访问控制、面向跨组织业务的隐私保护框架。使用加密数据进行流程挖掘是这一研究领域的主要目标之一。然而，据我们所知，这一目标尚未实现。为此，同态加密（HE）[13] 旨在允许对加密数据执行算术运算。因此，设计一种能够处理同态加密数据的流程挖掘算法

是一个全新的研究领域。然而，同态加密的计算复杂度极高，且架构延迟较高，因此并非适用于所有应用的通用解决方案。此外，通用流程挖掘算法并非为高效处理同态加密而设计。因此，开发一种计算复杂度较低、且能与现有流程挖掘算法配合使用的方法，值得期待。

B. 主要贡献

本文提出了一种名为哈尔变换加密（HTE）的新方法，其结构具有较低的计算复杂度。HTE 利用哈尔矩阵 [14] 的简单架构和一个私钥，为数字和字符提供加密。因此，对于将乘法视为耗时操作的网络和处理器，该方法能够显著节省资源。由于流程挖掘技术不受领域限制，可应用于任何存在流程且拥有数据日志的组织，因此本文提出的方法适用于任何行业。不过，本文通过医疗保健领域对其进行了分析。本研究旨在通过一种新设计的系统模型，满足当前数字商业网络的重要需求，例如降低计算复杂度，以及为事件日志数据提供易于恢复的加密技术结构。

此外，为证明如果存在一种处理加密数据的流程挖掘算法，那么改进这一密码系统是必不可少的要求，本文研究了作为众所周知的同态密码系统之一的Paillier技术的计算复杂度。需要注意的是，我们提出的对称密码技术在应用流程挖掘算法之前需要先进行解密，这适用于当前的应用场景。

本文其余部分安排如下：第二节介绍一些预备定义，包括同态加密和哈尔变换的概念。第三节和第四节分别介绍系统模型和所提出的方案。第五节介绍结果与讨论，包括事件日志的加密和计算复杂度分析，随后第六节给出结论。

记号：在全文中，我们使用小写和大写粗体字母 $\otimes$、α^T 和 $\{\cdot\}$ 分别表示向量、矩阵、克罗内克积、 α 的转置和乘法运算符。

II. 定义

初步定义及所提出的方法定义如下：

定义 1（哈尔变换）：未归一化的哈尔酉矩阵由 ±1 和 0 组成，变换仅通过加法和减法运算实现，不涉及任何乘法。然而，哈尔变换使用归一化的哈尔矩阵，其中除了 ±1 和 0 [14] 之外还包含其他数值。虽然哈尔变换需要进行乘法运算，但由于其中存在 ±1 和 0，计算时间非常短。

矩阵结构。因此，在复杂度和延迟方面实现了显著节省。Haar函数 $\hbar_\upsilon(\varphi)$定义在区间 $[0,1]$ 中，且该函数的阶数 υ 可唯一分解为整数 a 和 b，如下所示：

$$\upsilon = 2^a + b - 1, \quad N = 2^l, \quad \upsilon \in \{0,1,...,N-1\}, \tag{1}$$

其中 $l \in \{1,2,...\}$、 $0 \le a \le l-1$、 $0 \le b \le 2^a$。Haar函数定义如下：

$$\hbar_0(\varphi) \equiv \hbar_{00}(\varphi) = \frac{1}{\sqrt{N}}, \quad \varphi \in [0,1], \tag{2}$$

$$\hbar_\upsilon(\varphi) \equiv \hbar_{ab}(\varphi) = \frac{1}{\sqrt{N}} \begin{cases} 2^{\frac{a}{2}} & \frac{b-1}{2^a} \le \varphi < \frac{b-\frac{1}{2}}{2^a}, \\ -2^{\frac{a}{2}} & \frac{b-\frac{1}{2}}{2^a} \le \varphi < \frac{b}{2^a}, \\ 0 & \text{otherwise in [0,1].} \end{cases} \tag{3}$$

N 阶哈尔变换矩阵由各行组成，这些行是根据先前的函数在点 $\varphi = \frac{e}{N}$ 处计算得到的，其中 $e \in \{0,1,...,N-1\}$。令 H 表示一个维度为 2^l 的Haar矩阵，$\check{}$y 表示向量 x 的 N 点哈尔变换，则其计算公式为：

$$\check{\mathbf{y}} = \mathbf{H}_{2^l}\mathbf{x}. \tag{4}$$

注意，Haar矩阵通常可以通过以下方程推导得到：

$$\mathbf{H}_{2^l} = \begin{bmatrix} \mathbf{H}_{2^{l-1}} \otimes [1,1] \\ \mathbf{I}_{2^{l-1}} \otimes [1,-1] \end{bmatrix}, \tag{5}$$

$$\mathbf{I}_{2^{l-1}} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}_{2^{l-1}\times 2^{l-1}}. \tag{6}$$

定义 2（Walsh-Hadamard（WH））：维度为 2^l 的 Walsh-Hadamard 方阵定义如下 [15]:

$$\bar{\mathbf{H}}_{2^{l+1}} = \bar{\mathbf{H}}_{2^l} \otimes \bar{\mathbf{H}}_2 = \begin{bmatrix} \bar{\mathbf{H}}_{2^l} & \bar{\mathbf{H}}_{2^l} \\ \bar{\mathbf{H}}_{2^l} & -\bar{\mathbf{H}}_{2^l} \end{bmatrix}, \tag{7}$$

with

$$\bar{\mathbf{H}}_2 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{8}$$

该矩阵也可用于加密策略。然而，Walsh-Hadamard矩阵中没有零元素，因此比Haar矩阵更复杂。

定义3（同态加密）：同态加密允许在加密数据上执行包括加法和乘法在内的算术运算，而无需解密过程，这可作为计算复杂函数的基础 [13]。两种同态加密受到了更多关注：部分同态加密和全同态加密密码系统。前者仅允许执行一种运算，即加法或乘法；而后者支持在密文（CTs）上执行加法和乘法运算，旨在获得相应明文（PTs）上的计算结果。

最新发明的密码系统之一是 Paillier同态加密（PHE），它仅支持加法运算。

密码系统 [16]。下面将介绍 Paillier 系统的工作流程。

Paillier 算法首先随机选取两个互不相同的大素数 j 和 k，其中 $gcd(jk,(j-1)(k-1)) = 1$。接下来计算 $n = jk$、$\lambda = lcm(j-1, k-1)$ 并选取随机整数 g，其中 $g \in \mathbb{Z}^*_{n^2}$。必须重复上述步骤，直到确认 n整除 g 的阶。为此，需要检查是否存在 $\nu = (L(g^\lambda \bmod n^2))^{-1} \bmod n$，其中 $L(u) = \frac{u-1}{n}$。在这种情况下，(n, g) 和 (λ, ν) 分别称为公钥和私钥。

如果 $x \in \mathbb{Z}_n$ 和 $r \in \mathbb{Z}^*_n$ 分别是明文消息和一个随机数，那么 $E(x) = g^x.r^n \bmod n^2$ 是密文，其中 $E(x) \in \mathbb{Z}^*_{n^2}$。将使用以下公式对明文进行解密：$x = L(E(x)^\lambda \bmod n^2) \cdot \nu \bmod n$。由于 Paillier 密码系统只是一个加法同态密码系统，因此两个密文的乘积解密后等于其对应明文之和，如下所示：

$$D(E(x_1, r_1) \cdot E(x_2, r_2) \mod n^2) = x_1 + x_2 \mod n, \quad (9)$$

其中 $D(E(x)) = x$。如果 g、r、n 以及特别是明文x 都是较大的值，那么硬件和软件都会产生巨大的计算复杂度和延迟。因此，具有上述结构且包含大量乘法运算的同态加密不适用于当前的高速框架。PHE 的总结如算法1所述。

算法1 Paillier同态加密过程

输入：$\mathbf{x} = [x_1, ..., x_N]^T$个 N数字。

对于 $i = 1, ..., N$ **do**
  重复
    生成 j和 k当 $gcd(jk, (j-1)(k-1)) = 1$,
    设置 $n \leftarrow jk$,
    设置 $\lambda \leftarrow lcm(j-1, k-1)$、
    select random g where $g \in \mathbb{Z}^*_{n^2}$,
  直到 $gcd(L(g^\lambda \bmod n^2)\ n) \neq 1$,
  随机选择 r_i 其中 $r_i \in \mathbb{Z}^*_n$,
  计算密文 $E(x_i) = g^{x_i}.r_i^{n_i} \bmod n_i^2$,
结束循环

输出：$\mathbf{E(x)} = [Ex_1, ..., Ex_N]$，$\mathbf{r} = [r_1, ..., r_N]$。

III. 系统模型

超越工业4.0利用多个组织在串行和并行情形下，甚至在不同地理位置之间的协作，其中与不可信环境的数据传输是一个严重问题 [17]。对此，我们考虑了一个由 $M \in \mathbb{N}$个组织和一个流程挖掘节点组成的数字业务网络，如图1所示。事件日志在 1^{st} 组织（Org.1）处创建，并将在后续的 $M-1$ 个组织处通过添加新数据进行更新。流程挖掘节点应用流程挖掘算法，并将结果发送给第一个组织。最后，Org. 1 将改进指令发送给后续组织。在此场景中，Org. 1 和

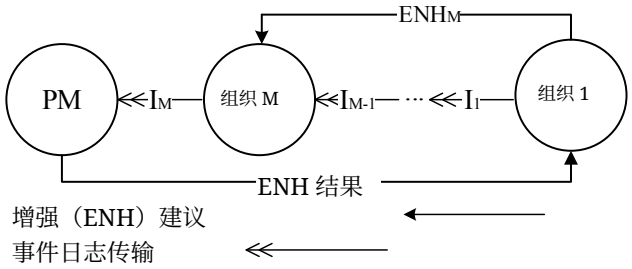


图1. 为 M 组织设计的系统模型

PM点是可信的，而 I_m、$m \in \{1, ..., M\}$（它们是将事件日志传输给下一个组织的任何类型的接口）是不可信环境。需要注意的是，除第一个组织外，任何组织对于其他组织的数据而言都被视为不可信环境。换句话说，每个组织的数据都使用个人私钥进行加密。

IV. 所提出的方法

令 $\mathbf{y} = [y_1, ..., y_N]^T$表示 $\mathbf{x} = [x_1, ..., x_N]^T$ 的密文向量，则利用（4）可得：

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N,} \end{bmatrix} = \zeta_p \frac{1}{\sqrt{N}} \mathbf{H}_{2^l} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}, \quad (10)$$

其中 $\zeta_p \in \mathbb{R}$ 是加密和解密过程中的任意随机私钥，其中 $p \in \{m, s\}$。个人私钥 ζ_m 使 m^{th}组织的数据无法从不可信环境中访问，而 ζ_s 指所有组织之间的共享密钥。

由于 H_{2^l} 是一个 2^l [14]阶方阵，如果由 $\tilde{N}$ 表示的实际数据数量小于 2^l，则在数据向量中补入 $\tilde{N} = 2^l - \tilde{N}$个零。例如，对于包含五个数字的小型数据日志，我们需要补入三个零，以适配一个 8×8 Haar矩阵。最后，利用 Haar矩阵 $\mathrm{H}_{2^l}^{-1} = \mathrm{H}_{2^l}^T$这一事实，可以从 y 中提取向量 $[x_1, ..., x_N]^T$，如下所示：

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \frac{1}{\zeta_m} \frac{1}{\sqrt{N}} \mathbf{H}_{2^l}^T \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}. \quad (11)$$

任何事件日志都由三种类型的数据属性组成：数字、时间戳和字符。后两种类型必须映射为唯一数字，如下所述：

时间戳根据时间原点计算，分钟或秒的差值是数值表达式。

字符根据其字母顺序映射为数字。

我们将所提出方案的设计步骤概括为算法 2 中的伪代码。为了更深入地了解该算法的工作原理，算法 2 的第 1 行至第 3 行负责将字符和时间戳映射为整数。

表 I 加密事件-日志在医疗保健领域中使用所提出的方法在 组织.1 和 组织.2

案例 ID		时间戳		活动		资源		心率	
PT	密文（ζ_s）	**PT**	密文（ζ_s）	**PT**	密文（ζ_s）	**PT**	密文（ζ_1）	**PT**	密文（ζ_2）
1	12	10 : 20	6812	$A \to 1$	23	Tom	(48, 22, 7, 18.2)	72	293
2	0	11 : 31	-1286	$A \to 1$	-11	Tom	(48, 22, 7, 18.2)	78	-8
1	0	11 : 42	-199.40	$B \to 2$	-2.82	Tom	(48, 22, 7, 18.2)	60	10.6
2	-2.82	12 : 30	-535.98	$B \to 2$	-1.41	Tom	(48, 22, 7, 18.2)	75	-45.96
1	-2	14 : 10	-142	$C \to 3$	0	$John$	(47, 2, -7, -8.4)	58	-6
1	-2	16 : 25	-96	$E \to 5$	0	$John$	(47, 2, -7, -8.4)	60	-15
2	0	17 : 30	-270	$D \to 4$	-4	$Anna$	(30, 0, -18.2, 18.2)	90	-2
2	0	19 : 24	-228	$E \to 5$	-2	$Anna$	(30, 0, -18.2, 18.2)	93	-3

表 II提供给ORG.3 的事件日志可用部分

案例 ID		时间戳		活动		资源		成本（$）	
PT	CT（ζ_s）	**PT**	CT（ζ_s）	**PT**	CT（ζ_s）	**PT**	CT (ζ_1)	**PT**	CT (ζ_3)
1	12	10 : 20	6812	$A \to 1$	23	Tom	(48, 22, 7, 18.2)	3	126
2	0	11 : 31	-1286	$A \to 1$	-11	Tom	(48, 22, 7, 18.2)	3	-54
1	0	11 : 42	-199.40	$B \to 2$	-2.82	Tom	(48, 22, 7, 18.2)	3	0
2	-2.82	12 : 30	-535.98	$B \to 2$	-1.41	Tom	(48, 22, 7, 18.2)	3	-42
1	-2	14 : 10	-142	$C \to 3$	0	$John$	(47, 2, -7, -8.4)	5	0
1	-2	16 : 25	-96	$E \to 5$	0	$John$	(47, 2, -7, -8.4)	5	0
2	0	17 : 30	-270	$D \to 4$	-4	$Anna$	(30, 0, -18.2, 18.2)	10	0
2	0	19 : 24	-228	$E \to 5$	-2	$Anna$	(30, 0, -18.2, 18.2)	10	0

算法 2 用于 m^{th} 组织的 HTE 方案

x $= [x_1, ..., x_N]^T$用于 $\bar{N}$数字。

1: 如果 $\mathbf{x}_i \notin \mathbb{R}$ 那么

2: 映射为数字

3: 结束 if

4: 如果 $\bar{N} < 2^l$ 则

5: 执行 $\tilde{N} = 2^l - \bar{N}$ 零填充过程，

6: 设置 $N \leftarrow N + N_\circ$

7: 结束 if

8: 计算 y $= \zeta_m \quad \frac{1}{\sqrt{N}} \mathbf{H}_{2^l} [x_1, ..., x_N]^T$，

输出：y，$\zeta_{m\circ}$

如果数据向量成员的实际数量（记为 N）小于 2^l，则需要执行零填充过程（参见算法 2 的第 4 行至第 6 行）。加密向量 y 只需通过第 8 行即可计算得到。最后，在应用过程挖掘工具之前，使用 y 和 ζ_p 进行解密过程。

V. 结果与讨论

本节将使用 $M = 3$ 组织，在包含典型医疗保健数据属性的事件日志上研究所提出方法的表现。此外，我们还将比较 HTE、WH 加密（WHE）[15]和 PHE 方案的计算复杂度及密码结构的脆弱性。在本节中，ζ_s、ζ_1、ζ_2、ζ_3、j、k 和 g 分别取值为 $\sqrt{8}$、2、$\sqrt{2}$、$3\sqrt{8}$、3、5 和 22。

A. 使用 HTE 加密

表 I 描述了一个小型事件日志及其通过 HTE 方法加密的匿名化版本和表 II。该日志由八个事件、两个不同的案例 ID：$\{1, 2\}$、五种不同的活动：$\{A, B, C, D, E\}$ 组成，这些活动由三位用户：{Tom、John、Anna}按时间戳升序执行。组织 1 使用 ζ_s 加密案例 ID、时间戳和活动列，因为这些数据是所有组织之间共享的数据。此外，它使用 ζ_1 匿名化资源列，以保护患者的隐私。更新过程可以向事件日志添加行和列。在本例中，组织 2 在不了解资源列的情况下，仅使用 ζ_2向日志添加加密后的心率信息列。在特定情况下，如有需要，ζ_1 可供组织 3 使用，并可使用 ζ_3将加密后的成本列添加到事件日志中。因此，只要拥有相应的密钥，任何组织都能够使用其他组织的数据。表 II 描述了组织 3 可用的信息。最后，可信过程挖掘点使用解密后的事件日志运行流程挖掘算法，并将改进建议发送给 1^{st} 组织。未来，在数字商业网络中，成千上万的组织协作处理大型事件日志是可以想象的。

参照 10 和算法 2，将数值数据向量 x 乘以 ζ_p 后进行哈尔变换，得到加密向量 y。时间戳是根据时间原点计算得出的，并生成数值向量 x（例如 10 : 20 - 00 : $00 = 620$ 分钟）。请注意，表 I 和表 II 中的每一列都分别且仅包含其明文向量 x 及对应的密文向量 y。对于活动和资源等非数值数据属性，x 是通过根据字母顺序将每个字符映射为一个数字而构成的。然而，如果明文向量中的元素包含多个字符，则每个单词都会对应一个单独的 y（例如 Tom $\to \mathbf{x} = [20, 15, 13, 0]$ 和 y $= [48, 22, 7, 18.2]$ with$N = 4$）。最后，表 I 和表 II 中的密文列将被视为加密数据日志。

表 III 基于表 I 中两个数据向量的 HTE、WHE 和 PHE 在所需乘法次数和加法次数方面的复杂度比较

数据向量	HTE		WHE		PHE	
x	Mul.	Add.	Mul.	Add.	Mul.	Add.
心率	32	24	64	56	$(22^{72}+22^{78}+22^{60}+22^{75}+22^{58}+22^{60}+22^{90}+22^{93})-8$	0
案例 ID	32	24	64	56	$(22^{1}+22^{2}+22^{1}+22^{2}+22^{1}+22^{1}+22^{2}+22^{2})-8$	0

B. 计算复杂度分析

在本小节中，我们将考察 Paillier 密码系统和所提出方案的计算复杂度，并根据乘法运算次数展示后者的优越性。我们证明，算法 1 中所述 PHE 方案的总体阶为 $\mathcal{O}\left(\sum_{i=1}^{N} g^{x_i}\right)$，而算法 2 中所述 HTE 的总体阶为 $\mathcal{O}\left(N^2\right)$。

证明1：算法1的计算复杂度分析中有三处瓶颈代码行，即第1–10行、第2–7行和第9行。第一个瓶颈包含一个长度为 N的“for”循环，其阶为 $\mathcal{O}(N)$。采用使第7行条件在第一次循环时成立的最简单假设，包含“repeat”循环的第二个瓶颈的计算复杂度受限于 gcd 和 lcm 算法，其阶分别为 $\mathcal{O}(\log\ jk.((j-1)(k-1))) \equiv \mathcal{O}\left(\log\ n^2\right)$ 和$\mathcal{O}$(1)，其中 $lcm = (|(j-1)_{\circ}\ (k-1)|)\ gcd(j-1,k-1)_{\circ}$ 考虑到$N \ll x_i$和 $n \ll x_i$，其中 $i \in \{1,...,N\}$，只有包含幂运算、阶为$\mathcal{O}$ (g^{x_i})的第三个瓶颈构成主导项。因此，最简单情况下PHE的总体阶为 $\mathcal{O}\left(\sum_{i=1}^{N} g^{x_i}\right)$。事实上，PHE的复杂度会随着数据值而变化，这会在硬件和编程架构中导致巨大的计算复杂度。

证明 2：参见算法 2，唯一的瓶颈出现在第 8 行，其中包含所提出方案的主要部分。根据公式 (5) 和 (6)，我们通过以下公式，将 $z(\mathrm{H}_{2^l})$ 提取为 H_{2^l} 中零的数量：

$$z(\mathbf{H}_{2^l}) = 2(2^{2(l-1)} - 2^{l-1}) + 2z(\mathbf{H}_{2^{l-1}}), \quad z(\mathbf{H}_2) = 0. \tag{12}$$

使用 HTE 方案时，长度为 $N = 2^l$ 的数据向量实际执行的乘法和加法运算次数如下：

$$Mul_N = N^2 - z(\mathbf{H}_{2^l}), \tag{13}$$

$$Add_N = N^2 - N - z(\mathbf{H}_{2^l}). \tag{14}$$

因此，由于哈尔矩阵中存在零，乘法次数显著减少，尤其是在 x_i 和 N 值较大时。最后，HTE 方案的总体阶数计算为 $\mathcal{O}\left(N^2\right)$。

根据表 I 中的数据，对 HTE、WHE 和 PHE 的乘法与加法运算次数进行比较，结果如表 III 所示。由于 PHE 的计算复杂度取决于 N 和 x_i 的值，因此，对于较大的数据值，需要大量乘法运算；而所提出的方案与明文值无关，可大幅降低计算量。

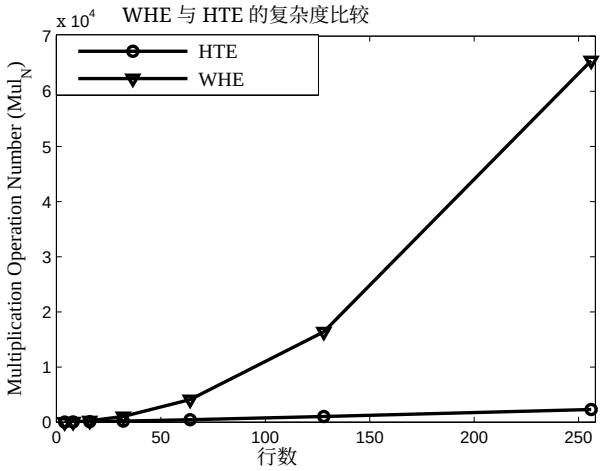


图 2. WHE 与 HTE 在 $N = 4$ 至 $N = 256$ 范围内的乘法次数计算复杂度对比

节省。Walsh-Hadamard 与哈尔矩阵之间的相似性促使我们在图 2 中比较 WHE 和 HTE。展示 $l \in \{2,...,8\}$ 下 WHE 和 HTE 的表现表明，与 WHE 相比，后者在较大的行数下具有乘法次数方面的显著优势。

C. 密码结构脆弱性

在本节中，我们研究破坏性因素对 PHE、WHE 和 HTE 方法结构的影响。

1）PHE的脆弱性：参照算法1第9行， g、 x、 r 和 n 是PHE方案中的密钥参数。假设 $n_1 = n_2$ 和 $g_1 = g_2$，两个受加性误差（AE）影响的密文 E（x_1）和 E（x_2）的乘法计算如下：

$$E(x_1 + x\acute{}_1) \cdot E(x_2 + x\acute{}_2) = g^{x_1+x\acute{}_1+x_2+x\acute{}_2} \cdot (r_1 \cdot r_2)^{n+n\acute{}} + \Re \mod (n+n\acute{})^2, \tag{15}$$

其中，$\acute{}$和 $\Re$ 分别表示到（·）的 AE，以及以 AE 为底数的项。

证明3: $E(x_1 + x\acute{}_1) \cdot E(x_2 + x\acute{}_2) = (g+\acute{g})^{x_1+x\acute{}_1} \cdot (r_1+r\acute{}_1)^{n+n\acute{}} \cdot ($ $g+\acute{g})^{\ x_2+x\acute{}_2} \cdot r_2+r\acute{}_2{}^{\ n+n\acute{}} \quad n+n\acute{}^{\ 2} = ()\mathrm{mod}\ ()(g+\acute{g})^{x_1} \cdot (g+\acute{g})^{x_1} \cdot (r_1+r\acute{}_1)^n \cdot (r_1+r\acute{}_1)^n \cdot (g+\acute{g})^{x_2} \cdot ($ $g+\acute{g} \quad r_2+r\acute{}_2{}^{\ n} \cdot r_2+r\acute{}_2 \quad n+n\acute{}^{\ 2} = g^{x_1}+)^{\ x\acute{}_2} \cdot ()\ ()^{n\acute{}}$ $\acute{g}^{x_1}+\ddot{g} \cdot g^{x\acute{}_1}+\acute{g}^{x\acute{}_1}+\ddot{\acute{g}} \cdot r_1{}^n+r\acute{}_1{}^n+\ddot{r}_1 \cdot r_1{}^{n\acute{}}+r\acute{}_1{}^{n\acute{}}+\ddot{r}\acute{}_1 \cdot$ $\mathrm{mod}\ ()\ ()\ ()\ ()\ ()\ ($ $g^{x_2}+\acute{g}^{x_2}+\ddot{g} \cdot g^{x\acute{}_2}+\acute{g}^{x\acute{}_2}+\ddot{\acute{g}} \cdot r_2{}^n+r\acute{}_2{}^n+\ddot{r}_2 \cdot r_2{}^{n\acute{}}+)\ ()\ ()\ ()$ $r_2{}^{n\acute{}}+\ddot{r}\acute{}_2 \quad n+n\acute{}^{\ 2} = g^{x_1+x\acute{}_1+x_2+x\acute{}_2} \cdot r_1 \cdot r_2{}^{\ n+n\acute{}} +$ $\mathrm{mod}\ ()\ (()\ )(\acute{g}^{x_1+x\acute{}_1+x_2+x\acute{}_2} \cdot (r\acute{}_1 \cdot r\acute{}_2)^{n+n\acute{}} + \ldots \mathrm{mod}\ (n+n\acute{})^2 \simeq E(x_1 + .. x\acute{}_1+x_2+x\acute{}_2)$, 其中 $\{\cdot\}$和 $\{\cdot\}$是使用

二项式定理 [18] 其中指数分别为真实数据和 AE。

参照 15 和 $Proof\ 3$，向 PHE 的每一个可能元素添加误差，会产生各种额外计算，其中 AE 作为基数和/或指数。此外，受扰动的密文-密文乘法会导致明文-明文加法不准确，从而破坏流程挖掘方案。因此，在存在 AE 的情况下，随着明文数量增加（即 x_1、 x_2.....），PHE 会成为一种计算复杂度极高且计算不准确的方法，无法用于高速应用。

2）WHE 和 HTE 的脆弱性：参照 5 和 7，Walsh-Hadamard 矩阵的顶行和最左列，以及 Haar矩阵的顶行都只包含 $+1$。此外，忽略前者的顶行和最左列后，行会呈现循环移位性质。由于破坏性因素通常只会导致上述矩阵发生相移， [15], WH 和哈尔矩阵的确定性结构可以轻松恢复。因此，Haar矩阵的确定性低复杂度结构与私钥相结合，是高速应用中的高效解决方案。

VI. 结论

实时网络和虚拟网络在数字商业技术中的兴起，促使我们研究行业中的隐私保护理念。由于这些网络本质上高度复杂，本文提出了一种名为 HTE、计算复杂度极低的加密方法，用于在执行流程挖掘技术之前对数据进行匿名化处理。本文的主要目标是为包含大量组织的商业网络提供一种低复杂度且易于还原的加密结构。所提出的方法基于哈尔变换和一个私钥，适用于事件日志中的任何数据类型。复杂度分析表明，在总体计算复杂度方面，HTE 显著优于 WHE 和 PHE 密码系统。所提出的方法对相关企业大有裨益，因为消费者隐私、复杂度、成本和速度都是这些企业十分关注的问题。未来的工作目标是设计一种能够处理经 HTE 加密数据的流程挖掘算法。

参考文献

[1] L. Banda、M. Mzyece 和 F. Mekuria，“面向移动网络运营商的 5G 商业模式——一项调查”，《IEEE Access》，第 10 卷，第 94 851–94 886 页，2022 年 9 月。[2] T.-Y. Lin、K.-R. Wu、Y.-S. Chen、W.-H. Huang 和 Y.-T. Chen，“疫情转型后的餐饮业中采用训练机器人实现外卖服务自动化”，《IEEE Robotics and Automation Letters》，第 6 卷，第 2 期，第 903–910 页，2021 年 11 月。[3] O. Loyola-Gonz´alez，“流程挖掘：软件比较、趋势与挑战”，《International Journal of Data Science and Analytics》，第 1–14 页，2022 年 12 月。[4] A. Hemmer、M. Abderrahim、R. Badonnel、J. Franc.ois 和 I. Chrisment，“用于支持物联网预测性安全的流程挖掘比较评估”，《IEEE Transactions on Network and Service Management》，第 18 卷，第 1 期，第 1092–1103 页，2021 年 3 月。[5] L. Pufahl、F. Zerbato、B. Weber 和 I. Weber，“医疗保健领域的 BPMN：挑战与最佳实践”，《Information Systems》，第 107 卷，文章编号 102013，2022 年 7 月。

[6] V. Bastidas、I. Reychav、A. Ofir、M.Bezbradica 和 M.Helfert，“智能城市建模概念”，《商业与信息系统工程》，第 1 卷，第1–15 页，2021 年 11月。[7] B. Chen、S. Qiao、J. Zhao、D. Liu、X. Shi、M.Lyu、H.Chen、H. Lu和 Y. Zhai，“基于零信任架构的5G 智能医疗安全意识与保护系统”，《IEEE物联网期刊》，第 8 卷，第 13期，第10 248–10 263 页，2021 年 7 月。[8] A. Pika、M. Wynn、S. Budiono、A. T. Hofstede、vanderWMP Aalst 和 H. Reijers，“医疗保健中的隐私保护流程挖掘”，《国际环境研究与公共卫生期刊》，第 17卷，第 5 期，第 1–12 页，2020 年 3 月。[9] A.Burattin、M. Conti和 D.Turato，“ 迈向匿名流程挖掘 ”，《 FiCloud 会议论文集》，第 69 卷，第 4 期，第 58–63 页，2015年 8 月。[10] E.Batista、A. Mart´ınez-Ballest´e 和 A. Solanas，“隐私保护流程挖掘：一种基于微聚合的方法”，《信息安全与应用期刊》，第68 卷，第 103235 页，2022 年 6 月。[11] J. L. Sebastian Saavedra 和J. Armas-Aguirre，“保障医疗保健领域个人数据隐私的流程挖掘模型”，载于《国际教育与科学技术大会》，2021 年 11月，第 1–10 页。[12] C. Liu、H. Duan、Q.Zeng、M. Zhou、F. Lu 和 J. Cheng，“为隐私保护跨组织业务流程挖掘提供全面支持”，《IEEE 服务计算汇刊》，第 12卷，第4期，第 639–653 页，2019年 8月。[13] D. Schoinianakis，“密码学中的剩余算术系统：现代安全应用综述”，《密码工程期刊》，第 10卷，第2期，第 249–267 页，2020 年 6 月。[14] N. A.Yalcin 和 F.Vatansever，“一种基于哈尔变换和Prony分析的信号估计新型混合方法”，《IEEE仪器与测量汇刊》，第 70 卷，第 1–9 页，2021 年。[15] M. Kazemian 、 J.Abouei 和 A.Anpalagan，“一种用于降低 5G 无线系统中用户间干扰、BER 和 PAPR 的低复杂度增强型 NOMA 方案”，《物理通信》，第 48卷，第 101412 页，2021 年 11 月。[16] Y. Yan、Z. Chen、V. Varadharajan、M. J. Hossain和G.E.Town，“利用 Paillier密码系统在电网中基于分布式共识进行经济调度”，《IEEE智能电网汇刊》，第12 卷，第 4期，第 3493–3502页，2021年7 月。[17] F. Zhu、F. Xu、X. Yang、X.Yi 和 A.Abuadbba，“IIoT环境下高效基于证书的代理签名方案的密码分析与改进”，《信息处理快报》，第 173 卷，第 106170页，2022年1月。[18] X.-S. Yang，“第 3 章——二项式定理与展开式”，载于《带示例与应用的工程数学》，X.-S. Yang 编，Academic Press，2017 年，第 31–35 页。