arXiv:2305.00960v1 [cs.DB] 1 May 2023

PMDG：通过数据泛化保护多视角流程挖掘的隐私

Ryan Hildebrant[1], Stephan A. Fahrenkrog-Petersen[2,3],Matthias Weidlich[2],和 Shangping Ren[4]

[1] 加州大学欧文分校，rhildebr@uci.edu[2] 柏林洪堡大学¨， {fahrenks、weidlima}@hu-berlin.de[3] 魏岑鲍姆网络社会研究所[4] 圣迭戈州立大学，sren@sdsu.edu

摘要。 事件日志的匿名化有助于开展流程挖掘，同时保护流程利益相关者的敏感信息。然而，现有技术主要关注控制流的隐私化。其他流程视角，如角色、资源和对象，往往被忽略或进行随机化处理，从而破坏各视角之间的依赖关系。因此，现有技术不适用于高级流程挖掘任务，例如社交网络挖掘或预测性监控。为弥补这一空白，我们提出 PMDG，这是一种通过数据泛化确保多视角流程挖掘隐私的框架。它为事件日志提供基于群组的隐私保证，同时保留控制流与其他流程视角之间的特征依赖关系。不同于依赖数据抑制或噪声插入的现有隐私化技术，PMDG 采用数据泛化：将事件中引用的活动和属性值泛化为更抽象的形式，以获得从隐私角度看足够大的等价类。我们通过实验证明，在挖掘交接和预测结果时，PMDG 的表现优于最先进的匿名化技术。

关键词： 隐私化 · K-匿名性 · 属性泛化

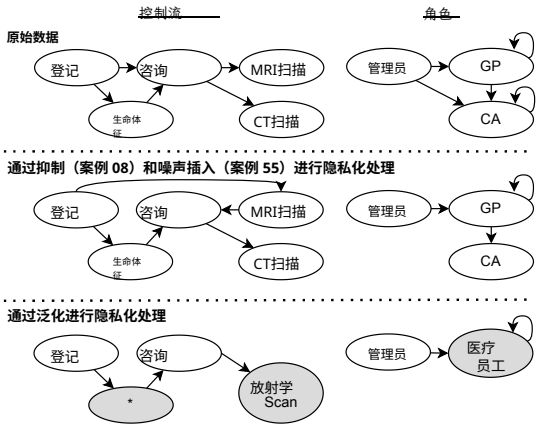
1 引言

隐私保护流程挖掘 [1]能够在保护流程执行相关人员敏感数据的同时，对业务流程进行数据驱动分析。为此，现有技术依赖于对事件日志进行匿名化处理。事件日志通常建模为一组轨迹，每条轨迹都是由表示活动执行的事件构成的序列。为获得可证明的隐私保证，需要对事件日志中的轨迹进行转换。现有技术在采用的隐私保证以及这些转换所保留的属性方面各不相同。事件日志匿名化可以保证差分隐私 [2]，也可以依赖基于群组的概念，例如 k-匿名性及其衍生形式 [3]。此外，相关转换可能只会抑制日志中的行为，也可能引入新的、带噪声的行为，例如此前未出现过的活动执行序列。

大多数用于隐私保护流程挖掘 [4–6] 的技术都侧重于从事件日志 [7]构建流程模型。因此，它们针对的是控制流 per-

病例活动		位置	Role
07	登记	日间诊所行政	
07	生命体征	日间诊所全科医生	
07	日间诊所全科医生咨询		
07	CT扫描	医院	CA
...	...	...	...
08	登记	医院	管理员
08	医院咨询		CA
08	MRI扫描	医院	CA
...	...	...	...
隐私化添加的噪声轨迹:			
55	登记	医院	管理员
55	MRI扫描	日间诊所全科医生	
55	日间诊所行政咨询		
...	...	...	...

(a)



(b)

图 1: （a）包含案例 07 和案例 08 轨迹的临床路径示例日志；（b）考虑原始日志、通过抑制（案例 08 被抑制）和噪声插入（添加案例 55）进行隐私化的事件日志，以及通过活动和角色泛化进行隐私化的事件日志时的控制流和角色视角。

流程视角，旨在确保匿名事件日志包含语义正确的流程活动执行序列。然而，事件日志还包含其他流程视角的信息，例如角色、资源和案例对象。高级流程挖掘任务会利用控制流与这些附加视角之间的关系，例如，在流程执行期间提取交接 [8] 或构建模型来预测正在运行的流程实例的结果 [9]。

截至目前，确保控制流之外的隐私保护仍存在显著的研究空白。迄今为止，与事件相关联的数据要么被忽略，要么在控制流匿名化后被随机分配 [10]。后一种做法往往会破坏各个视角之间的依赖关系，导致事件日志不适用于多视角流程挖掘，如图 1所示。这里，图 1a展示了一个包含两名患者（案例 07 和案例 08）的临床路径事件日志示例。图 1b （上）则重点展示了控制流依赖关系以及相关角色之间的交接。然而，现有通过抑制（案例 08）和噪声插入（案例 55）实现控制流隐私化的技术，不仅会扰乱控制流，还会破坏流程视角之间的依赖关系，如图 1b （中）所示，其中展示了角色之间的交接。这引出了一个问题：在匿名化事件日志时，如何保留流程视角之间的特征依赖关系。

在本文中，我们通过 PMDG 回答这一问题。PMDG 是一个通过数据泛化为多视角流程挖掘保护隐私的框架。在构建满足 k-匿名性（业界常采用的一种隐私保证）的事件日志时，它保留了流程视角之间的依赖关系 [11]。为此，我们采用数据泛化，而非数据抑制和噪声插入。图 1b（下方）展示了这一方法的直观思路：通过对活动进行泛化（*CT*扫描 和 *MRI*扫描 变为 放射学扫描）以及对角色信息进行泛化（全科医生 和 临床助理 变为 医务人员），可以得到规模足够大的轨迹等价类。虽然泛化会造成一定的信息损失，但可以说它保留了诸如以下特征：

只有在咨询之后才进行放射学扫描，以及行政人员与医务人员之间的交接。

总而言之，我们的贡献包括：定义 PMDG，将其作为首个实现多视角流程挖掘隐私保护的框架；以及通过具体技术对其进行实例化，包括：(i) 对轨迹进行向量化，以促进控制流泛化；(ii) 选择用于抽象的层次结构；以及 (iii) 应用所选层次结构，对控制流和分配给事件的数据进行泛化。

我们通过在三个公开事件日志上进行实验，展示了 PMDG 在多视角流程挖掘中的有效性。在挖掘交接和预测流程结果时，我们发现，在分别保持特征性交接和分类准确率方面，PMDG 显著优于最先进的匿名化技术。下文中，第 2 节回顾了隐私保护流程挖掘的相关工作。第 3 节随后介绍背景信息。我们在第 4 节概述并实例化 PMDG 框架，然后在第 5 节介绍评估实验。我们从定性角度讨论我们的方法，第 6 节然后在第 7 节作出结论。

2 相关工作

近年来，隐私保护流程挖掘受到了广泛关注 [1]。已有多种方法被提出，用于确保 k-匿名性及其他基于群组的隐私保证，例如合并相似轨迹 [4,12] 或过滤数据 [5]。由于这些方法侧重于控制流，因此并不适用于多视角流程挖掘。

与在轨迹群组中隐藏敏感数据不同，一些方法通过向事件日志中注入噪声来实现差分隐私 [6,13,14]。在此，隐私保证限制了任何单个个体对匿名化数据可能产生的影响。然而，这些方法会过滤日志中的行为，或引入新的、此前未出现过的行为。

文献 [15], 强调了对其他流程视角进行隐私化的重要性，并提出了一种针对特定视角（即资源分配）的技术。一般而言，上述控制流匿名化方法可以与丰富化步骤结合；该步骤要么随机分配值 [10]，要么统一这些值在事件日志中的分布 [16]。无论采用哪种方式，流程视角之间的特征依赖关系都会受到破坏，而引入新的依赖关系则可能导致分析得出错误结论。

机密流程挖掘采用了另一种思路，旨在通过加密保护事件日志 [17]。其中可能包含多个流程视角，但无法对数据集中个体的隐私提供任何形式化保证。

我们的方法依赖于数据泛化来实现多个流程视角的隐私保护。泛化是一种成熟的方法，可用于对关系型数据、层次型数据和简单序列数据进行隐私保护，参见 [18–20]。然而，PMDG 首次将数据泛化应用于事件日志，即应用于多变量序列，在这些序列中，必须保留不同维度之间的依赖关系。

3 背景

下面，我们总结事件日志和基于群组的隐私保障中的常见概念，这些概念是定义 PMDG 所必需的。
事件日志。流程挖掘基于事件，每个事件都表示一次活动执行的记录，例如图 1 中的登记或 MRI扫描。我们分别用 $\mathcal{A}$ 和 $\mathcal{E}$表示活动和事件的全集。事件 $e \in \mathcal{E}$所指示的执行活动记作 $e.a \in \mathcal{A}$。事件具有一个模式，由属性集合 $\mathcal{D} = \{D_1, \ldots, D_n\}$定义；我们用$\mathcal{V}_D$表示属性 D 的取值域。对于事件 e，我们用 $e.D \in \mathcal{V}_D$ 表示属性 D的相应取值。在图 1 中，我们有 $\mathcal{D} = \{$位置、角色$\}$。这里，角色属性的取值包括管理员、临床助理和全科医生，而相应的取值域还可能包含其他值。具体而言，它可以包含更抽象的角色，例如医务人员或员工。

与同一次流程执行相关的事件会归为一条轨迹。每条轨迹 σ 都是由 $|\sigma| = n$ 个事件 $\langle e_1, \ldots, e_n \rangle \in \mathcal{E}^*$ 组成的有限序列。我们用 $\sigma.A$ 表示轨迹 σ的控制流，即轨迹中各事件所指示的活动执行顺序。例如，在我们的示例中，案例 07 的轨迹为 $\sigma.A = \langle$登记、生命体征、咨询、*CT*扫描$\rangle$。所有控制流相同的轨迹都称为同一轨迹变体，并由其中一条轨迹标识。也就是说，$[\![\sigma]\!]^A$ 是与 σ控制流相同的轨迹多重集。轨迹多重集称为事件日志，$L = [\sigma_1, \ldots, \sigma_n]$。它是许多流程挖掘算法的输入。

k-匿名性。 保护个人隐私的一种广为人知的方法，是将个人隐藏在一个群体中，这正是 k-匿名性隐私保证 [3]的目标。其理念是，在一个数据集（在我们的场景中即事件日志）中，某个人应当无法与至少 $k-1$ 其他人区分开来。因此，识别某个个体的概率——即所谓的 *身份披露*问题——可以被限制在 $1/k$以内。为了使 k 个人彼此无法区分，需要对齐准标识符。一般而言，准标识符是能够用于识别个人的属性，例如邮政编码或出生日期。

在我们的场景中，我们假设所有事件的所有属性以及轨迹的控制流都可以作为准标识符。因此，我们将具有相同控制流和所选属性值序列的所有轨迹视为属于同一个等价类。所选属性是根据高级流程挖掘任务所需的特定视角生成的。与轨迹变体的记号方式一致，我们用等价类中的一个成员来标识该等价类，即 $[\![\sigma \mid \mathcal{D}']\!]$ 表示一个等价类，其中包含所有控制流和指定属性的属性值均与 σ 相同的轨迹 $\mathcal{D}' \subseteq \mathcal{D}$。据此，我们将 k-匿名性定义如下：设 $L = [\sigma_1, \ldots, \sigma_n]$ 为一个事件日志。那么，如果对于由轨迹 $\sigma \in L$所诱导的每个等价类 $[\![\sigma \mid \mathcal{D}']\!]$ in L，均有 $|[\![\sigma \mid \mathcal{D}']\!]| \geq k$，则日志 L 满足关于给定视角的*k*-匿名性。

我们指出，上述对等价类的定义，以及由此得出的 k-匿名性，体现了可能最严格的概念。它假设了最强的攻击者，在这种情况下，任何属性和控制流都可能被用作身份披露攻击中的准标识符。因此，该定义涵盖了攻击者仅掌握某一类

背景知识 [5], 的攻击，例如知道某个人执行了哪些活动，但不知道这些活动的执行顺序。为避免对攻击者的背景知识作出假设，我们采用上述模型来表示最坏情况。

4 事件日志的泛化

本节首先概述我们工作的设计原则（第4.1节）。随后，我们概述用于弥补已识别研究空白的 PMDG框架（第4.2节）。虽然其中一些步骤依赖现有技术，但为确保匿名事件日志具有较高效用，某些方面需要采用新技术。具体而言，我们介绍轨迹向量化策略（第4.3节）和层级选择策略（第4.4节）。

4.1 框架的设计原则

我们采用设计科学方法论构建了 PMDG框架 [21]。问题观察的出发点是，现有的事件日志匿名化技术大多依赖噪声插入和聚合，却未采用任何泛化策略。为弥补这一研究空白，我们为该制品推导出以下设计目标：给定一个包含 m 条轨迹的事件日志以及隐私参数$k \leq m$，该制品应当：（i）通过泛化转换事件日志，使其满足 k-匿名性；（ii）尽可能减少应用于日志的泛化总量。在本节余下部分，我们将介绍用于实现这些设计目标的 PMDG框架。该制品的评估步骤将在后续章节中介绍，而本文则代表设计科学方法论的最终沟通步骤。

4.2 PMDG框架

如图 2 所示，我们的框架应用于包含多个流程视角信息的事件日志，这些信息通过分配给事件的属性值表示（见第 3 节）。此外，该框架依赖于泛化层次结构。也就是说，活动层次结构可建模为函数 $\rho_A : \mathcal{A} \to \mathcal{A}' \cup \{\star\}$，它将某个活动映射为更抽象的活动或通配符 $\star$。对于表示额外视角的属性 D，值层次结构 $\rho_D : \mathcal{V}_D \to \mathcal{V}_D \cup \{\star\}$ 将某个属性值映射为更抽象的值或通配符。无论是控制流还是视角，其层次结构都以通配符 $\star$为根，这意味着对于任意活动 $a \in \mathcal{A}$ 和属性值$v \in \mathcal{V}_D$，均有 $\star \in \rho_A^*(a)$ 和 $\star \in \rho_D^*(v)$，其中 ρ^* 表示泛化层次结构 ρ的传递应用。此外，对于所有流程视角，都可以使用多个层次结构来泛化相应信息。借助这些层次结构，PMDG框架将作为输入给定的事件日志转换为另一事件日志，使转换后的日志能够保证 k-匿名性。

数据泛化的常见策略基于对数据集元素的单个值进行更改的操作 [22]。因此，为了实现全面泛化，即确保任意两个元素最终都可能属于同一个等价

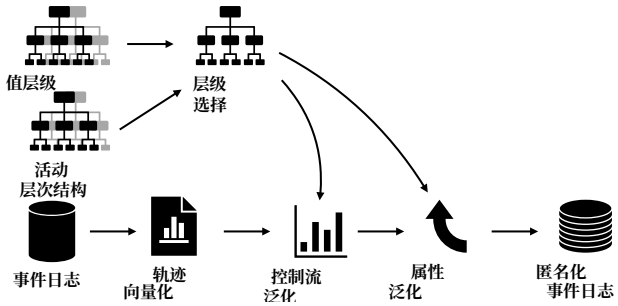


图 2：PMDG框架概览。

类，必须确保所有元素都具有相同的结构。应用到我们的场景中，这意味着所有轨迹的长度都必须相同，且所有事件都必须具有相同的模式。后一项要求在实践中通常不会带来任何挑战，并且已纳入我们对模型的定义；而轨迹长度的差异则需要处理。为此，我们的框架将轨迹向量化作为第一步，并将在第4.3节中作进一步说明。

接下来，PMDG 框架会先泛化控制流，再考虑其他过程视角。原因在于，行为信息是高级流程挖掘任务的基础，因此始终应受到 k-匿名性保护。其他视角会丰富控制流，也可以对隐私保证进行更细粒度的控制，例如采用 t-接近性 [23] 而非 k-匿名性。

然而，要通过泛化获得隐私保证，可能存在多个层级，既适用于控制流，也适用于其他流程视角。例如，可以根据所处理的对象对活动进行泛化（例如，与某项*CT*扫描 相关的所有活动都泛化为单个活动），也可以根据所执行操作的类型进行泛化（例如，将所有开具不同药物的活动合并为单个活动）。类似地，可以依据某种组织结构（例如，医院中的病房）或某种能力（例如，医生的专科）对角色进行泛化。每种层级对所得事件日志效用的影响都不同。因此，PMDG框架包含一个步骤，用于指导选择一个层级进行控制流泛化，并为代表其他流程视角的每个属性选择一个层级，详见第 4.4 节。

选定泛化层次结构后，便使用现有算法将其应用于事件日志，以实现 k-匿名性 [22, 24]。本质上，这些算法采用逐步泛化的方式，直至生成的等价类足够大。具体而言，在我们的上下文中，给定一个活动层次结构 ρ_A 以及每个属性对应的值层次结构 ρ_D， D泛化步骤的结果可以表示为

轨迹向量化后的轨迹：

案例 07	
活动	Role
登记	管理员
生命体征	GP
全科医生咨询	
CT扫描	CA

案例 08	
活动	Role
登记	管理员
⋆	⋆
CA咨询	
MRI扫描	CA

泛化后的轨迹：

案例 07	
活动	Role
登记	管理员
⋆	⋆
咨询	医务人员
放射扫描	CA

案例 08	
活动	Role
登记	管理员
⋆	⋆
咨询	医务人员
放射扫描	CA

图 3：泛化示例。

可表征如下：对于原始日志 L 中的每条轨迹 $\sigma = \langle e_1, \ldots, e_n \rangle$，生成的日志 L' 将包含一条轨迹 $\sigma' = \langle e'_1, \ldots, e'_n \rangle$，使得对于每个事件 e_j、$1 \leq j \leq n$，活动和每个属性值要么保持不变，要么已被泛化，即 $e'_j.a = e_j.a$ 或 $e_j.a = \rho^*_A(e_j.a)$，以及 $e'_j.D = e_j.D$ 或 $e_j.D = \rho^*_D(e_j.D)$。

例如，考虑图 1 中的案例 07 和 08 的轨迹，并仅关注控制流和角色视角。轨迹向量化会通过插入一个包含通配符值的活动（详见第4.3节）来规范化两条轨迹的长度，见图 3 （上）。通过活动层次结构将*MRI扫描* 和 *CT扫描* 泛化为放射学扫描，并通过值层次结构将全科医生 和 临床助理 泛化为医务人员，这些轨迹便可被泛化到同一个等价类中，见图 3 （下）。由此得到的事件日志将满足2-匿名性。

4.3 轨迹向量化

如上所述，全面泛化要求数据集中的所有元素都采用相同的结构。此外，我们希望确保对于每个 k，只要其数目不超过日志中的轨迹数，就能生成一个匿名事件日志。然而，只有长度相同的轨迹才能合并到同一个等价类中。因此，我们需要统一事件日志中轨迹的长度。为此，我们采用轨迹向量化，这与预测性流程监控中的序列编码思路类似 [25]。具体而言，给定一个事件日志 $L = [\sigma_1, \ldots, \sigma_n]$，轨迹向量化会生成一个日志 $L' = [\sigma'_1, \ldots, \sigma'_n]$，使得：

- 所有轨迹的长度都相同，即对于 σ'_i，σ'_j 中的所有 L'，均有 $|\sigma'_i| = |\sigma'_j|$。
- 对于原始日志 L中的每条轨迹 $\sigma = \langle e_1, \ldots, e_m \rangle$，$L'$中都存在一条对应轨迹$\sigma' = \langle e'_1, \ldots, e'_k \rangle$，使得将 σ' 投影到事件 $\{e_1, \ldots, e_m\}$上可得到轨迹 σ，并且 σ' 中不属于该投影的所有事件 e 都是通配元素，也就是说，$e.A = \star$ 且对于所有 $D \in \mathcal{D}$，均有 $e.D = \star$。

一种朴素方法是，在所有较短轨迹的末尾添加通配符事件，将其扩展至事件日志中轨迹的最大长度。然而，不能指望这种方法保留轨迹对流程挖掘的效用，尤其是从控制流视角来看。例如，对于图 3中案例 08 的轨迹，在末尾添加通配符事件会对后续泛化造成严重影响：这会导致两条轨迹中除第一个活动以外的所有活动都被泛化为根元素（$\star$），而无法保留其中咨询和放射学扫描活动的信息。

因此，在 PMDG 中，我们采用一种基于多序列比对（MSA）的策略 [26]。本质上，MSA 会确定如何在符号序列中插入间隙，使所有序列中相同的符号都被分配到相同的索引位置，同时使间隙索引的数量最少。在我们的场景中，我们将 MSA 用于控制流视角，因为它是执行过程挖掘任务的基础。也就是说，给定一个事件日志 $L = [\sigma_1, \ldots, \sigma_n]$，将 MSA 应用于集合 $\{\sigma_1.A, \ldots, \sigma_n.A\}$，以确定应在何处插入通配符事件。

4.4 层级选择

如上所述，可以采用多个层次结构来泛化控制流，或泛化代表其他流程视角的数据。下文将先详细介绍层次结构的类型及其来源，以及它们对匿名事件日志效用的影响；随后，我们将提出一种启发式方案，以指导将泛化层次结构选作 PMDG 的一部分。

层次结构类型。一般而言，可以区分两种层次结构：

(i) 句法层次结构通过省略活动标签或属性值的一部分得到。句法层次结构的常见示例包括数值（例如，邮政编码*'12489'* 和*'12555'* 泛化为*'12—'*）或日期（*'10/2022'*和*'12/2022'* 泛化为*'–/2022'*）。不过，也可以对活动进行泛化，例如，将*CT*扫描和*MRI*扫描泛化为扫描，方法是省略标签的第一个词元。

(ii) 语义层级对活动或属性值的含义进行泛化。例如，可以将表示城市的属性（'柏林'）泛化为国家（'德国'）、更大的区域（'欧盟'）或洲（'欧洲'）。创建语义层级需要领域知识，这些层级通常由用户定义或从知识库中提取。例如，对于传统业务流程中的活动，《MIT流程手册》 [27]定义了活动的泛化层次结构。

即使只考虑一种类型的层次结构，层次结构的选择也会影响最终事件日志的效用。以日期的句法泛化为例，*'11.2022'* 和*'12.2022'* 不仅可以泛化为*'–/2022'*，也可以分别泛化为*'11/—'* 和*'12/—'*，具体取决于要抑制哪些部分。每种泛化都会提供不同类型的信息，从而影响我们能通过流程挖掘从匿名化日志中回答的问题类型。

选择层次结构。 鉴于为数据泛化选择特定层次结构会产生重大影响，理想情况下，应针对控制流和所有属性测试所有可用的层次结构。根据选定的效用度量衡量最终事件日志的质量，即可确定最佳的层次结构组合。然而，由于层次结构组合数量呈指数级增长，这种蛮力方法通常难以实施。因此，在 PMDG 中，我们采用一种启发式策略，分别为控制流和每个属性选择泛化层次结构。该启发式策略基于效用这一概念，我们考虑以下几种具体形式：

- 效用由匿名事件日志中的等价类数量决定。这里的直观理解是，匿名日志中的等价类数量越多，对原始日志中方差的呈现就越好。
- 效用与等价类大小的差异（即其中包含的轨迹数量）成反比。其动机是，相较于罕见行为，更精确地保留有关常见行为的信息。

基于特定的效用概念，可以依据估算的效用，为每个流程视角选择层次结构，方法如下。设 $\{\rho_D^1, \ldots, \rho_D^n\}$ 是属性 D 的一组层次结构（或类似地，是活动的一组层次结构）。然后，对于每个层次结构，我们在仅考虑属性 D

以及泛化层次结构中的一个层级（即，ρ_D^1，$\ldots \rho_D^n$ 仅应用一次，而非传递应用）时，确定等价类。令 u_1^i 为针对使用 ρ_D^i 得到的等价类所确定的效用；如上所述，该效用可以定义为类的数量。随后，对层次结构后续层级所得到的等价类进行迭代评估，从而得到层次结构 ρ_D^i 在纳入至第 j 层级时的效用值 u_j^i。对于每个层次结构 ρ_D^i，以加权方式对这些效用值求和，即 $u^i = \sum_{j=1}^{k} w_j \cdot u_j^i$，其中 k 为该层次结构的最大深度。使用权重 w_j 可以让我们偏好不同的泛化层级，例如，优先选择从城市泛化到国家，而不是从国家泛化到洲。最后，我们选择属性 D 的所有层次结构 $\{\rho_D^1, \ldots \rho_D^n\}$ 中估算效用 u^i 最大的层次结构 ρ_D^i。

5 评估

本节将探讨使用 PMDG 对事件日志进行匿名化会如何影响高级流程挖掘任务的效用。通过实证评估，我们展示了 PMDG 的可行性和有效性。首先，我们将在第 5.1 节概述实验中使用的数据集。接下来，我们将在第 5.2 节介绍实验设置、基线和评估指标。我们的实验结果见第 5.3 节。

5.1 数据集和实施

我们的实验使用了三个真实世界的事件日志：BPIC 2013 [28], 道路交通罚款 [29], 和 CoSeLoG [30]。对于每个日志，我们都排除了只出现一次的所有变体。这确保了匿名化设置较为合理（无论如何，唯一轨迹都会造成问题）。某些高级流程挖掘任务的实验要求所有事件都包含相同的属性；在这种情况下，我们仅使用 BPIC 2013 [28] 和 CoSeLoG [30] 事件日志进行这些实验，因为道路交通罚款数据集缺少此类属性。

对于我们所有的实验，我们都在 GitHub 上提供了开源实现。[5]轨迹泛化方法使用 Python 实现。对于属性的泛化，我们使用了 ARX 项目的 Java 库。[6] 对于交接挖掘实验，我们使用了 PM4Py 的组织挖掘功能。[7] 对于结果预测实验，我们使用了 scikit-learn。[8]

5.2 实验设置

参数设置。 在实验中，我们采用不同强度的 k-匿名化，数值范围从 k 到 $\{5, 10, 15, 20\}$。此外，我们使用手动创建的语义层级。我们也测试了句法层次结构，但这些方法

[5] https://github.com/Ryanhilde/PMDG_框架/[6] https://arx.deidentifier.org[7] https://pm4py.fit.fraunhofer.de[8] https://scikit-learn.org

表 1：控制流保持的比较

Log	轨迹向量	$k = 5$	$k = 10$	$k = 15$	$k = 20$
CoSeLoG	MSA	**17**	**13**	**9**	**8**
	朴素方法	**17**	**13**	8	**8**
BPIC 2013	MSA	**82**	31	23	**23**
	朴素方法	65	**34**	**27**	22
交通罚款	MSA	**75**	**53**	**43**	**37**
	朴素方法	12	12	12	12

在减少对匿名化日志所做更改的数量方面始终表现较差。我们根据控制流中保留的等价类以及属性的最小泛化，选择了语义层级。在预测性流程监控实验中，我们在 1,000 个随机生成的 20/80 训练测试划分上训练了决策树。

基线。 作为我们部分实验的基线，我们使用了 PRIPEL [10], ，这是一种转换事件日志以实现 ϵ-差分隐私的框架 [2]。所提供的隐私保证无法与 k-匿名性直接比较，即我们不能假定某个特定的 k-值会确保与 PRIPEL 中某种设置相同程度的隐私。不过，PRIPEL 是最佳的比较选择，因为它是唯一能够处理所有属性值的现有技术。一般而言，较低的 ϵ 值对应更强的隐私保证。在实验中，我们针对两个事件日志，考虑 PRIPEL 的两种设置：较弱的隐私保证（$\epsilon = 1.0$）和较强的隐私保证（$\epsilon = 0.1$）。此外，在较弱的设置中，我们将 PRIPEL 的剪枝参数设为 2 ，在较强的设置中设为 20 ；并始终将最大前缀长度设为轨迹变体的平均长度。由于底层采用了控制流匿名化技术，这两个参数是 PRIPEL 所必需的 [6]。此外，我们将基于 MSA 的轨迹向量化技术与一种朴素方法进行比较，该方法在所有轨迹末尾填充通配符。

评估指标。我们使用剩余变体的数量作为指标，衡量匿名化后的控制流保持情况，结果见表 1。为研究属性匿名化对高级流程挖掘任务效用的影响，我们考察了两种高级流程挖掘技术：交接发现 [8, 31] 和流程结果预测 [9]。

通过交接分析，分析人员可以调查同一条轨迹中两个事件之间哪些属性值直接相继出现。这类分析通常针对与资源相关的属性，例如资源角色或位置。为了量化匿名化的结果，我们测量泛化后的事件日志相较于原始事件日志保留的信息。我们使用信息保留度量来捕捉泛化造成的信息损失。我们的度量基于这样一种直觉：将交接从其原始关系（例如，在资源位置的情况下，从德国到中国）泛化为某种泛化关系（例如，从欧洲到中国），仍具有一定效用。此外，这种效用高于将交接泛化到更高层级（从欧洲到亚洲）或最高层级（从欧洲到世界）的情况。因此，我们将泛化交接的保留度定义为

p 如下:

$$p = \frac{[1 - \frac{\alpha(e_1'.D)}{\alpha(*)} + \frac{\alpha(e_1.D)}{|\alpha(*)}] + [1 - \frac{\alpha(e_2'.D)}{\alpha(*)} + \frac{\alpha(e_2.D)}{\alpha(*)}]}{2} \quad (1)$$

这里假设 α 是一个函数，用于返回可通过（泛化后的）属性值表示的潜在属性值数量。例如，值 “欧盟” 可以表示属性 D 中的 27 个国家，该属性用于编码国家。特殊情况 $\alpha(*)$ 返回属性 D 的细粒度值数量，即所有可能的国家。值 e_1 和 e_1' 分别表示交接左侧的原始值和泛化值，而 e_2 和 e_2' 表示右侧的值，即（中国）。我们的指标只衡量现有交接的信息损失，因为泛化无法插入新的交接关系。

作为第二项分析任务，我们考虑流程结果预测。此处，事件日志的效用由分类结果决定。我们使用广为人知的分类指标评估这些结果：精确率，即被标记为正类的实例中实际正确的比例；召回率，即实际正类中被正确标记的比例；以及*F1* 分数，即精确率和召回率的调和平均数。

5.3 结果

控制流保持。 在 表 1中，我们展示了基于 MSA 的轨迹向量化通常优于朴素方法。可以看出，在交通罚款事件日志中，该方法能带来显著收益。对于朴素方法表现更好的案例，其优势相对较小。总体而言，我们观察到，越高，控制流方差损失就越大。

交接。 在 图 4中，我们展示了根据属性 组织：角色和 $k = 5$，从匿名化后的 BPIC 2013 日志中生成的交接情况。此类分析有助于组织了解哪些资源角色通常会相互交互。我们可以清楚地看到，与原始交接图（左图）相比，通过 PMDG 进行匿名化后生成的交接图（中图）规模更小，包含的信息也更少。不过，从信息保留度量的结果中可以得出更详细的见解，如表 2所示。我们注意到，实际上大量交接信息得以保留。这表明，图 4 中所示的信息损失，主要是因为低粒度的交接关系被更高泛化层级的交接关系所取代。

相比之下，右侧图 图 4 展示了使用弱配置的 PRIPEL 得到的结果。在这里，几乎所有属性值都彼此相连。虽然这显然保留了所有现有的交接，但也引入了大量虚假的交接。可以说，这是信息的一大损失。不过，对于一种基于噪声插入、并对分配给事件的属性值采用随机化的匿名化技术而言，这一结果是可以预期的。

我们通过一个示例分析问题来说明两种匿名化策略之间的差异。Volvo IT（即提供 BPIC 2013 日志的公司）希望了解乒乓行为，也就是交接循环 [28]。若基于属性 组织：角色来探讨这个问题，原始日志显示角色 E10 和 V3 2 之间存在交接，而以下角色对之间则没有连接

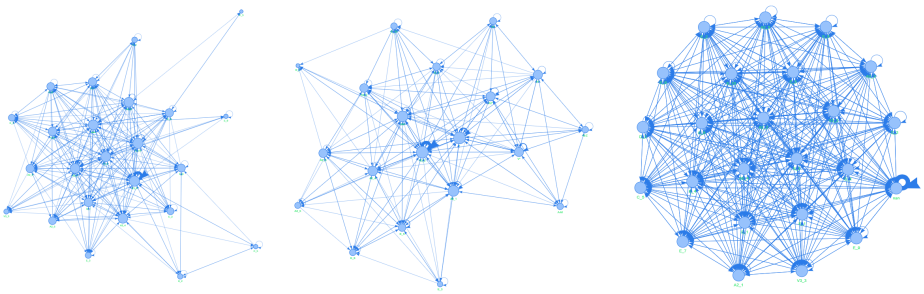


图 4: （左）原始事件日志，（中） $k = 5$，（右）PRIPEL 弱设置下 org:role 属性的交接图。

表 2：泛化交接的精确率

日志和属性	$k = 5$	$k = 10$	$k = 15$	$k = 20$
BPIC 2013 ”org:role”	85.2	80.0	79.3	79.4
BPIC 2013 ”涉及的组织”	100	100	100	100
BPIC 2013“资源所在国家”	89.7	89.7	89.7	90.2
BPIC 2013“组织所在国家”	89.2	88.1	88.1	87.6
CoSeLoG“org:resource”	73.1	75.4	75.3	75.3

：角色 {E9、V3 2}、 {A2 1、C 1}, 以及 {A2 2、C 1}。使用 PMDG 后，角色 E9 和 E10 会被泛化为单一角色 E*，并与角色 V3 2 相连。虽然这掩盖了 E9 与 V3 2 之间原本没有连接这一事实，但仍有助于评估 E 类角色与 V3 2 之间的交接情况。同时，PMDG 生成的图中不包含 A2 和 C 1 之间错误的交接，因此分析乒乓行为时不会考虑这些角色。相比之下，PRIPEL 得到的含噪结果不适合用于此项分析，因为它暗示所有角色都参与了循环交接。

流程结果预测。接下来，我们考察常见的流程结果预测任务。这里，我们使用决策树分类器来预测结束活动。在图5中，我们展示了分类实验的结果。左侧热图展示了不同隐私保证下的结果 k。我们观察到，对于 BPIC 2013 日志和 CoSeLoG 日志，较高的隐私保证会带来更好的预测指标。这一现象符合预期，因为更通用的日志包含的控制流差异更少，因此更容易进行预测。

另一方面，基于从 PRIPEL 获取的事件日志训练的分类器，其分类结果的精确率和召回率都极低。观察右侧热图即可看到这些结果。PRIPEL 在匿名化日志中加入的噪声显然对分类结果产生了强烈的负面影响，因此使匿名事件日志无法用于结果预测。

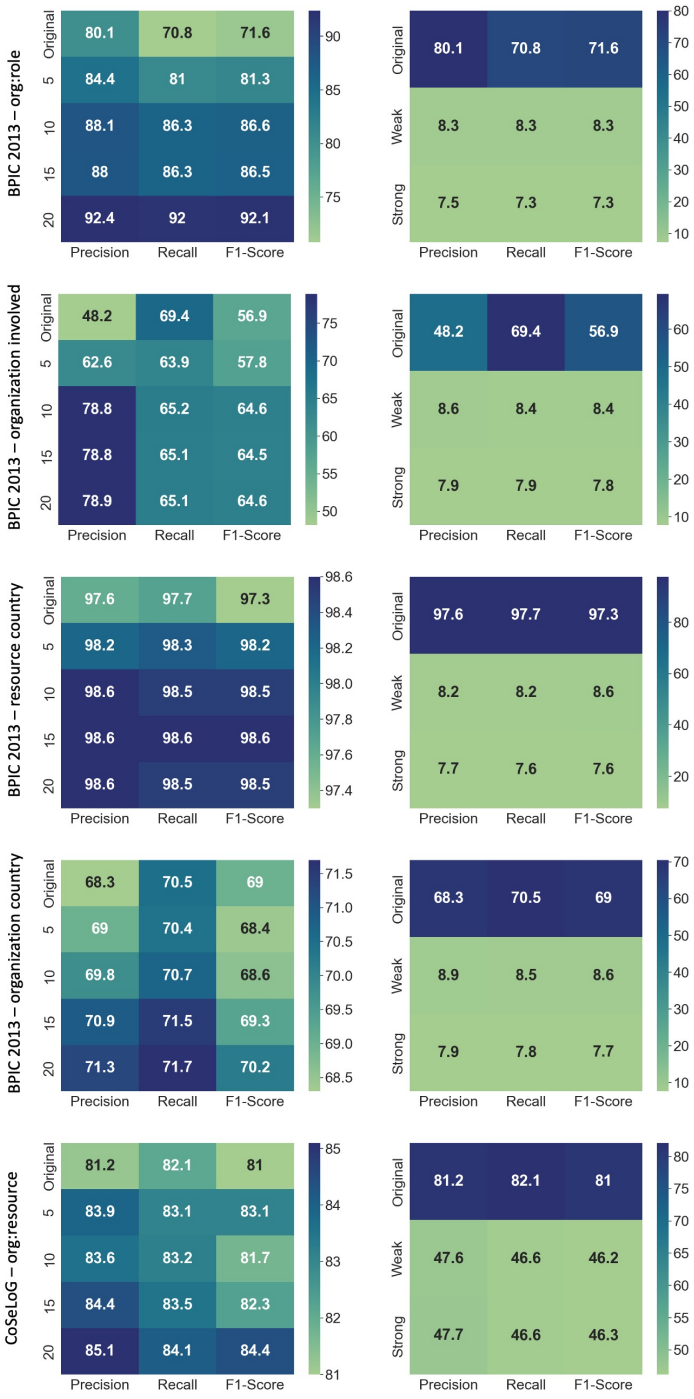


图5：**PMDG** （左）和**PRIPEL** （右）方法的决策树结果。

6 讨论

下面，我们将讨论我们的方法的若干方面，以帮助理解应用 PMDG 时需要考虑的机遇和局限性。

泛化的局限性。 泛化的主要缺点是匿名化数据中细节的丢失。因此，数据可能不再适用于某些分析。例如，如果将单个资源泛化到其所属部门，就无法再检查是否遵循了双人操作规则。再比如医疗领域，药物剂量可能会被泛化。在强泛化的情况下，可能无法回答与每日剂量上限有关的问题。

泛化层次结构的选择。 PMDG 的应用效果在很大程度上取决于可用的泛化层次结构。语义层级需要人工创建，这可能会限制其可用性。此外，对于某些属性，如何对其进行泛化可能并不明显。一个典型例子是活动，它们往往缺乏明确无歧义的泛化层次结构。若没有领域专家的知识，就难以判断泛化在多大程度上保留了特定于流程的信息。此外，PMDG 不保证抽象后的结果在所有情况下都具有实用性。结果是否实用，取决于所提供的泛化层次结构的质量，以及为提供隐私保证所需的抽象程度。

不同的抽象层级。根据所使用的泛化技术，匿名化日志可能包含抽象层级不同的属性值。例如，表示区域的属性可能包含代表某个国家或某个洲的值。在分析中混用这些不同的抽象层级可能颇具挑战，因为大多数技术都没有内置解决方案来处理这种异质的抽象层级。因此，使用 PMDG 匿名化的事件日志在应用于常见的流程挖掘解决方案之前，可能需要进行一些后处理。

完全抑制的风险。如果事件日志仅由少量轨迹的变体构成，且这些轨迹的属性值差异很大，那么这些属性值可能会被实质性地抑制，即将表示某个城市的区域值泛化为「世界」。在这种情况下，泛化的所有潜在优势都会丧失。提供包含大量泛化层级的层次结构，可以解决这一问题，使属性值能够收敛到仍具有一定效用的层级。然而，大量的泛化层级可能导致事件日志中的属性值方差很大。

维数灾难。实现 k-匿名时，一个众所周知的问题是维数灾难 [32], ，这意味着属性或事件的增加会使隐私保证更难实现。随着我们引入分配给案例和事件的其他属性，数据会被划分为更小的等价类。因此，匿名事件日志的效用可能会进一步降低。解决这一问题的一种潜在方法是采用混合隐私保证 [33]。这些技术允许对某些属性使用基于噪声的匿名化，对其他属性进行泛化，并根据特定分析场景的要求作出选择。

7 结论

在本研究中，我们介绍了 PMDG，这是一种匿名化框架，可转换事件日志，使其受到 k-匿名保护。我们方法的创新之处在于：（i）能够保留事件日志中记录的不同流程视角之间的依赖关系，即控制流以及分配给事件的属性值；（ii）利用数据泛化技术实现隐私保证。在真实世界事件日志的实验中，我们表明，与最先进的方法相比，PMDG 在先进流程挖掘技术的效用保留方面表现更优。未来，我们计划研究如何支持泛化层次结构的构建与应用，以优化匿名化日志的效用。

致谢

本研究得到美国国家科学基金会（NSF）资助，项目编号为 1952225；并得到德国联邦教育与研究部（BMBF）资助，项目编号为 16DII133。

参考文献

1. Gamal Elkoumy、Stephan A Fahrenkrog-Petersen、Mohammadreza Fani Sani、Agnes Koschmider、Felix Mannhardt、Saskia Nu˜nez Von Voigt、Majid Rafiei 和 Leopold Von Waldthausen。《流程挖掘中的隐私与保密性：威胁与研究挑战》。*ACM TMIS*，13(1):1–17，2021。2. 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。Springer，2008。3. Latanya Sweeney。K-匿名性：一种隐私保护模型。《国际不确定性、模糊性与基于知识的系统杂志》，10(5):557–570，2002年10月。4. Stephan A. Fahrenkrog-Petersen、Han van der Aa 和 Matthias Weidlich。Pretsa：用于隐私感知流程发现的事件日志净化。*ICPM*，第1–8页，2019。5. Majid Rafiei 和 Wil M.P. van der Aalst。流程挖掘中的基于群组的隐私保护技术。*DKE*，134:101908，2021。6. Felix Mannhardt、Agnes Koschmider、Nathalie Baracaldo、Matthias Weidlich 和 Judith Michael。隐私保护流程挖掘。*BISE*，61(5):595–614，2019。7. Adriano Augusto、Raffaele Conforti、Marlon Dumas、Marcello La Rosa、Fabrizio Maria Maggi、Andrea Marrella、Massimo Mecella 和 Allar Soo。基于事件日志的流程模型自动发现：综述与基准测试。*IEEE TKDE*，31(4):686–705，2018。8. Weidong Zhao 和 Xudong Zhao。从组织视角进行流程挖掘。载于《智能系统基础》，第701–708页。Springer，2014。9. Irene Teinemaa、Marlon Dumas、Marcello La Rosa 和 Fabrizio Maria Maggi。面向结果的预测性流程监控：综述与基准测试。*ACM TKDD*，13(2):17:1–17:57，2019。10. Stephan A. Fahrenkrog-Petersen、Han van der Aa 和 Matthias Weidlich。PRIPEL：包含上下文信息的隐私保护事件日志发布。载于*BPM*，收录于《计算机科学讲义》第12168卷，第111–128页。Springer，2020。11. Stephan Kessler、Jens Hoff 和 Johann-Christoph Freytag。SAP HANA走向隐私保护：从隐私研究到隐私感知企业分析。*VLDB*，12(12):1998–2009，2019。12. Edgar Batista、Antoni Mart´ınez-Ballest´e 和 Agusti Solanas。隐私保护流程挖掘：一种基于微聚合的方法。*JISA*，68:103235，2022。13. Stephan A Fahrenkog-Petersen、Martin Kabierski、Fabian R¨osel、Han van der Aa 和 Matthias Weidlich。Sacofa：面向流程挖掘的语义感知控制流匿名化。载于*ICPM*，第72–79页。IEEE，2021。

14. Gamal Elkoumy、Alisa Pankova 和 Marlon Dumas。挖掘我的数据，但别单独识别我：用于流程挖掘的差分隐私事件日志。载于*ICPM*，第80–87页。IEEE，2021。15. Majid Rafiei 和 Wil Aalst。在保护隐私的同时从事件日志中挖掘角色。2019年7月。16. Edgar Batista 和 Agusti Solanas。一种基于均匀化的方法，用于在流程挖掘分析期间保护个人隐私。《*Peer-to-Peer Networking and Applications*》，14(3):1500–1519，2021。17. Majid Rafiei、Leopold von Waldthausen 和 Wil MP van der Aalst。利用抽象和加密保障流程挖掘中的保密性。载于*SIMPDA*，第101–123页。Springer，2018。18. Florence Corpet。利用层次聚类进行多序列比对。《*Nucleic Acids Research*》，16(22):10881–10890，1988。19. Ke Wang、Philip S Yu 和 Sourav Chakraborty。自底向上泛化：一种用于隐私保护的数据挖掘解决方案。载于*ICDM*，第249–256页。IEEE，2004。20. Wai Kit Wong、Nikos Mamoulis 和 David Wai Lok Cheung。隐私保护数据发布中的非同质泛化。载于*SIGMOD*，第747–758页，美国纽约州纽约市，2010。ACM。21. Ken Peffers、Tuure Tuunanen、Marcus A Rothenberger 和 Samir Chatterjee。面向信息系统研究的设计科学研究方法论。《*Journal of Management Information Systems*》，24(3):45–77，2007。22. Kristen LeFevre、David J DeWitt 和 Raghu Ramakrishnan。Incognito：高效的全域 k-匿名。载于*SIGMOD*，第49–60页，2005。23. Ninghui Li、Tiancheng Li 和 Suresh Venkatasubramanian。t-接近性：超越 k-匿名性和 l-多样性的隐私保护。载于*ICDE 2007, 20, 2007*，*15*106 115 2007马尔马拉酒店，土耳其伊斯坦布尔，*4*月-页 – 。IEEE计算机学会，。24. Fabian Prasser、Florian Kohlmayer、Ronald Lautenschl¨ager 和 Klaus A Kuhn。ARX：用于生物医学数据匿名化的综合工具。载于*AMIA*年度研讨会论文集，第2014卷，第984页。美国医学信息学会，2014。25. Anna Leontjeva、Raffaele Conforti、Chiara Di Francescomarino、Marlon Dumas 和 Fab-rizio Maria Maggi。用于业务流程预测性监控的复杂符号序列编码。载于*BPM*，第297–313页。Springer，2016。26. RP Jagadeesh Chandra Bose 和 Wil MP van der Aalst。利用轨迹对齐进行流程诊断：机遇、问题与挑战。《*Information Systems*》，37(2):117–141，2012。27. Thomas W Malone、Kevin Crowston 和 George Arthur Herman。组织业务知识：《*MIT*流程手册》。MIT出版社，2003。28. Boudewijn F. van Dongen、Barbara Weber、Diogo R. Ferreira 和 Jochen De Weerdt，编者。第三届业务流程智能挑战赛论文集，与第九届国际业务流程智能研讨会（*BPI 2013*）同期举行，中国北京，*2013*年*8*月*26*日，收录于*CEUR*研讨会论文集第1052卷。CEUR-WS.org，2013。29. M de Leoni 和 F Mannhardt。道路交通罚款管理流程，2015。doi：10.4121/u-uid:270fd440-1057-4fb9-89a9-b699b47990f5。30. Joos Buijs。环境许可申请流程的受理阶段（“WABO”），CoSeLoG项目。*https://doi.org/106*，2014*4121073865734367953570977./uuid:aa-be---bcedbe*。31. Wil M. P. van der Aalst。《流程挖掘：数据科学实践》。Springer，海德堡，第2版，2016。32. Charu C Aggarwal。论 k-匿名性与维数灾难。载于*VLDB*，第5卷，第901–909页，2005。33. Naoise Holohan、Spiros Antonatos、Stefano Braghin 和 Pol Mac Aonghusa。(k，ϵ)-匿名性：具有 ϵ-差分隐私的 k-匿名性。*CoRR*，abs/1710.01615，2017。