跨信任域的联邦k-支配Skyline查询算法

史烨轩[1,2]，童咏昕[1,2,3]，周昊[1,2]，许可[1,2,3]，吕卫锋[1,2]
1. 北京航空航天大学软件开发环境国家重点实验室，北京 100191；
2. 北京航空航天大学计算机学院，北京 100191；
3. 北京航空航天大学未来区块链与隐私计算高精尖创新中心，北京 100191

摘要
k-支配Skyline查询是一种主流的Skyline查询变种，其在多目标决策与推荐领域有着广泛的应用。随着这些应用规模不断扩大，在由多个参与方组成的数据联邦中进行跨域k-支配Skyline查询的需求日益旺盛。然而，由于数据联邦中的参与方之间彼此不互信，进行跨信任域的查询计算需引入大量安全操作，效率较低。为此提出了一种基于跨域隐私向量聚合的算法，从而实现高效的联邦k-支配Skyline查询，并运用一种密文压缩技术进一步优化查询效率，最后通过充分的实验验证了所提方案的优越性。

关键词
k-支配Skyline查询；数据联邦；安全多方计算；同态加密

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023047

Cross trust domain federated k-dominant skyline query processing

SHI Yexuan[1,2], TONG Yongxin[1,2,3], ZHOU Hao[1,2], XU Ke[1,2,3], LYU Weifeng[1,2]
1. State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
2. School of Computer Science and Engineering, Beihang University, Beijing 100191, China
3. Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, Beihang University, Beijing 100191, China

Abstract
k-dominant skyline is a prevailing skyline query which has widespread applications in multi-criteria decision making and recommendation. As these applications continuously scale up, there is an increasing demand to support k-dominant skyline over a data federation which consists of multiple data silos, each holding disjoint columns of the entire dataset. Yet it is challenging to support k-dominant skyline over a data federation. This is because strict security constraints are often imposed to query processing over data federations, whereas naively adopting security techniques leads to unacceptably inefficient queries. In this paper, we presented an efficient and secure k-dominant skyline for a data federation. Specifically, we devised a novel private vector aggregation-based solution with ciphertext compression-based optimization for efficient k-dominant skyline query processing while providing security guarantees. Extensive evaluations on both synthetic and real datasets showed the superiority of our method.

Key words
k-dominant skyline, data federation, secure multi-party computation, homomorphic encryption

0 引言

Skyline查询是大数据管理领域中的一种多维数据检索手段，其可根据每维数据的偏序关系检索出所有在某一维度优于其他数据的元组[1]。这一查询已被广泛用于基于位置的服务、多目标决策、兴趣点推荐等应用之中[2]。

近年来，越来越多的应用需联合多个参与方的私有数据进行联合推荐从而提升服务质量，同时保证各参与方的私有数据不泄露，这种跨信任域的计算模式被称为数据联邦[3-4]。具体而言，数据联邦是指由多个参与方组成的数据库系统，每个参与方拥有全体数据库的一部分（相同数据对象的多个不相交列），这些参与方共同提供数据分析服务（如Skyline查询），而不向其他参与方透露其敏感数据。

以银行和保险公司的跨域联合推荐为例。银行通常会记录客户的资金状况，包括存款、贷款等情况。而保险公司则持有客户的投保记录。双方可利用Skyline查询对客户的资金状况和投保记录进行联合分析，从而发现那些尚未投保但存款较多的客户，以达到潜在客户挖掘的目的。在这一推荐过程中，由于客户的资金状况、投保记录等信息均较为敏感，银行与保险公司需在不暴露其私有数据的前提下完成。然而，随着进行跨域推荐的参与方数量增加，数据的维度会越来越大。而Skyline查询往往无法对高维数据进行有效筛选，从而导致推荐结果的可用性下降。作为Skyline查询的一种常见变种，k-支配Skyline查询[5]将筛选的约束条件限制在k维，从而有效地对高维数据进行处理，在跨信任域的大规模用户推荐中可起到关键作用。

然而，现有工作中尚未见在跨信任域的数据联邦中进行k-支配Skyline查询的相关研究。一方面，大量现有研究[5-7]的查询过程中未考虑数据隐私；另一方面，现有的安全Skyline查询方法[8-10]主要面向经典的Skyline查询，该方法可以应用Skyline查询中支配关系的传递性对数据集进行剪枝。然而，这种剪枝手段并不适用于k-支配Skyline查询。尽管通用的安全多方计算（secure multi-party computation，SMC）技术可用于支持查询，但直接应用该技术的算法效率较低。例如，使用当前最先进的SMC库MPDZ[11]实现的基线算法仅处理500个元组的联邦k-支配Skyline（federated k-dominant skyline，FKDS）查询即需花费12 h以上。因此，如何针对联邦k-支配Skyline查询设计安全且高效的算法仍是一个亟待解决的问题。

为了解决上述挑战，本文研究跨信任域的联邦k-支配Skyline查询，旨在设计具有安全保障的高效解决方案。具体而言，本文首先将k-支配关系的验证过程重新表述为参与方之间的隐私向量聚合，并设计了一种跨域隐私向量聚合方法用于解决该问题。此外，本文提出了一种密文压缩技术，可进一步优化算法的执行效率。本文的主要贡献总结如下。

- 首次定义了联邦k-支配Skyline查询问题，其在跨信任域的多目标决策和兴趣点推荐中具有十分广泛的应用前景。
- 本文提出了一种基于隐私向量聚合的联邦k-支配Skyline查询方案，该方案可在保证数据隐私的前提下实现低运行/

通信成本，并提出了一种密文压缩技术，以进一步提升查询效率。

● 本文通过在真实与合成数据集上进行充分的实验，验证了所提解决方案的优越性。实验表明，所提解决方案比最先进的通用SMC技术实现的基线算法查询效率快739.3倍，并可将通信成本降低至少两个数量级。

1 问题定义

本节介绍跨信任域的联邦k-支配Skyline查询相关基础概念与其形式化定义。

定义1（参与方）：参与方（由 s 表示）是一个拥有自己数据集 D_s 的自治机构。数据集的每一行对应一个元组，为一个 d_s 维向量。换言之，参与方 s 存储了每个元组的 d_s 个属性。

参与方的属性集可被视为数据集 D_s 的关系模式，由 $\mathcal{A}_s=\{A_1,\cdots,A_{d_s}\}$ 表示。参与方 s 持有的每个元组 p 都是数据模式 $\mathcal{A}_s$ 的一个关系实例。

定义2（数据联邦）：数据联邦由 m 个参与方组成，即 $S=\{s_1,\cdots,s_m\}$ 。其中，每个参与方 s_i 都持有自己的私有数据集 D_{s_i} ，即相同的 n 个元组中不同的 d_{s_i} 个属性。其中，每个参与方持有的属性集 $\mathcal{A}_{s_i}$ 在各个参与方中互不相同，即 m 个参与方持有的属性不重叠。

记 D 为纵向数据联邦 S 中全体参与方的总数据集，其中包含 D 上的 n 个元组的 $d=\sum_{i=1}^{m}d_{s_i}$ 个属性，D 的模式可记为 $\mathcal{A}=\bigcup_{i=1}^{m}\mathcal{A}_{s_i}$ 。本文中的数据联邦面向跨信任域的兴趣点推荐[2]等应用。方便起见，本文假设其中存在一个协调者负责接收查询，分发给其他参与方执行，并将查询结果汇总返回给查询方。在实际应用中，协调者可以是参与方之一，也可是不可信的第三方服务器。

定义3（k-支配）：在数据模式为 $\mathcal{A}$ 的数据联邦数据集 D 中，两条元组 p_a 和 p_b 之间的k-支配关系可记作 $\prec_k$ 。例如元组 p_a 可k-支配元组 p_b 可记作 $p_a\prec_k p_b$ 。这是指：①在所有属性中，至少存在个属性 $A_j\in\mathcal{A}$ ，满足 $p_a[A_j]\leqslant p_b[A_j]$ ；②至少存在1个属性 $A_j\in\mathcal{A}$ ，满足 $p_a[A_j]<p_b[A_j]$ 。其中，$p_a[A_j]$ 指的是元组 p_a 中第 j 个属性的取值（$1\leqslant j\leqslant d$ ）。

这种支配关系在推荐系统等应用中的推荐剪枝中可起到很大作用。例如，若 p_a 可k-支配 p_b ，则意味着至少存在 k 个属性使得 p_a 在这些属性上优于 p_b 。因此，在推荐中即可忽略 p_b ，只考虑 p_a 进行后续处理。

定义4（半诚实模型）：半诚实（又称诚实但好奇）模型，假设每个参与方诚实地执行所需的查询，但可能会在查询期间尝试推断其他参与方的数据。该模型还允许多达 $m-1$ 个参与方协作推断剩余参与方的敏感数据。

半诚实模型广泛用于数据联邦研究之中[3,4,10]。本文假设每个参与方均为半诚实的，即其不会与其他参与方或协调者共享自己的数据集 D_{s_i} ，且可能试图推断其他参与方的数据集，但会诚实地执行协调者分配的查询操作。

定义5（联邦k-支配Skyline查询）：给定数据联邦 S，其全体参与方总数据集为 D，数据模式为 $\mathcal{A}$ 。联邦k-支配Skyline查询旨在找出所有不被其他元组k-支配的元组ID，可形式化表述为：

$$\mathrm{FKDS}(D)=\{b\mid p_b\in D,\nexists p_a\in D\ \ \mathrm{s.t.}\,p_a\prec_k p_b\} \tag{1}$$

该查询需保证半诚实模型下的安全性，即每个参与方 s_i 均只能获知查询结果的

元组ID，但不应获知这些元组在其他参与方中的属性信息。

跨信任域的联邦k-支配Skyline查询可找出在所有属性中优于其他属性的元组，这有助于参与方提供更准确和有效的推荐，而不会泄露有关其自身私有数据的信息。本文旨在设计安全且高效的联邦k-支配Skyline查询算法，从而可在半诚实模型的安全保证下进行跨信任域的大规模用户推荐。

2 基于跨域隐私向量聚合的解决方案

本节基于k-支配关系的隐私向量形式表述，提出了一种高效的联邦k-支配Skyline查询算法。其关键思想是将k-支配关系的计算过程从元组之间的运算转换为向量之间的运算，并通过跨域隐私向量的安全聚合操作进行求解。本节首先解释了如何将k-支配关系表述为隐私向量运算，随后提出了一种跨域隐私向量的安全聚合操作，最后描述了整个基于隐私向量聚合的联邦k-支配Skyline查询算法。

2.1 k-支配关系的隐私向量表述

在联邦k-支配Skyline查询中，为判断元组 p_b 是否被元组 p_a k-支配，需统计 p_a 中小于等于 p_b 的属性个数 cle 与严格小于 p_b 的属性个数 csl 。二者存在k-支配关系即可转化为判断 $\mathrm{cle} \geqslant k$ 且csl>0。然而，在数据联邦中，逐条进行跨域k-支配关系的判断会引入大量的安全比较操作，计算效率较低。因此，本文将k-支配Skyline查询的条件重新表述为跨域隐私向量聚合的形式，通过向量式的安全计算求解，提升查询效率。

具体而言，若元组 p_b 出现在k-支配Skyline查询结果集中，则说明不存在 p_a 可k-支配元组 p_b 。因此，可将所有元组与 p_b 各属性的大小关系进行计数，并汇总成两向量：

$$\mathbf{CLE}=\mathrm{cle}_{p_1},\cdots,\mathrm{cle}_{p_n},\mathbf{CSL}=\mathrm{csl}_{p_1},\cdots,\mathrm{csl}_{p_n} \tag{2}$$

其中， cle_{p_i} 表示元组 p_i 的各属性中小于等于元组 p_b 的数量， csl_{p_i} 则表示元组 p_i 的各属性中严格小于元组 p_b 的数量。判断元组 p_b 是否属于k-支配Skyline的查询结果即可表述为如下向量运算：

$$\mathbf{res}=(\mathbf{CLE}-(k-1)\mathbf{1}_n)\cdot\mathbf{CSL} \tag{3}$$

其中，$\mathbf{1}_n$表示一个n维的全1向量。在计算出向量**res**后，若其中存在某一维大于0，则说明元组p_b不在查询结果之中；否则，元组p_b属于联邦k-支配Skyline的查询结果。然而，向量**CLE**和**SCL**分散在各参与方之中，由于隐私约束无法直接对其进行汇总与统计。第2.2节将详细介绍如何通过同态加密技术来高效地进行各参与方之间跨域隐私向量的安全聚合。

2.2 跨域隐私向量的安全聚合操作

(1) 安全原语

本文设计的跨域隐私向量安全聚合操作使用Paillier同态加密作为主要的构建模块。Paillier是一种部分同态加密方案，该方案允许直接在加密后的数据上进行计算。对给定的明文u，将对其使用Paillier同态加密方案加密后的密文记作$\varepsilon(u)$。本文仅简要介绍这一加密方案的同态运算性质，其实现方法可参考文献[12]。

- Paillier同态加密方案支持对两个密文$\varepsilon(u_1)$与$\varepsilon(u_2)$进行加法运算，即$\varepsilon(u_1+u_2)=\varepsilon(u_1)\oplus\varepsilon(u_2)$。其中，$\oplus$表示密文的同态加法运算符。

● Paillier同态加密方案支持对密文$\varepsilon(u_1)$与明文u_2进行数乘运算，即$\varepsilon(u_1 \cdot u_2) = u_2 \otimes \varepsilon(u_1) = \varepsilon(u_1)^{u_2}$。

Paillier加密方案已被证明具有语义安全性[12]，即对于给定的密文，攻击者无法推断出其对应明文的任何信息。本文采用的是Paillier同态加密方案的多参与方版本，即私钥分散在各参与方之中，只有各参与方联合起来才能对密文进行解密。

（2）算法实现

在数据联邦中，每个参与方s_i仅能根据其本地拥有的属性集$\mathcal{A}_{s_i}$统计出在这一属性集中小于等于或严格小于p_b的向量，记作$\mathbf{CLE}^{s_i}$与$\mathbf{CSL}^{s_i}$。联邦全局的向量则需将各参与方的隐私向量进行加法运算，这一过程可以使用Paillier的同态加法运算完成。在计算出全局的向量**CLE**和**CSL**后，计算的过程则需先对向量**CSL**解密，随后运用Paillier的同态数乘运算完成。这一过程并不会造成额外的隐私泄露，因为若出现在查询结果中，则表示向量**CSL**中必定有大于0的元素出现，因此向量**CSL**可直接解密参与运算。这一过程的伪代码如算法1所示。

算法1：跨域隐私向量聚合操作伪代码

Input：参与方s_1，s_2，…，s_m的隐私向量$\{\mathbf{CLE}^{s_i}\}$与$\{\mathbf{CSL}^{s_i}\}$

Output：隐私向量聚合结果**res**

1：for 数据拥有方s_i do

2：　　对隐私向量加密得到$\varepsilon(\mathbf{CLE}^{s_i})$与$\varepsilon(\mathbf{CSL}^{s_i})$

3：　　将密文发送给参与方s_1

4：参与方s_1进行同态加法运算：

$$\varepsilon(\mathrm{CLE}) \leftarrow \oplus_{i=1}^{m}\varepsilon(\mathrm{CLE}^{s_i}),$$
$$\varepsilon(\mathrm{CSL}) \leftarrow \oplus_{i=1}^{m}\varepsilon(\mathrm{CSL}^{s_i})$$

5：所有参与方联合对密文$\varepsilon(\mathbf{CSL})$解密得到向量**CSL**

6：参与方s_1计算：

$$\varepsilon(\mathbf{res}) \leftarrow \mathbf{CSL} \otimes \left(\varepsilon(\mathbf{CLE}) \oplus (1-k) \otimes \varepsilon(1_n)\right)$$

7：所有参与方联合对密文$\varepsilon(\mathbf{res})$解密得到向量**res**

8：return

（3）密文压缩优化

这一跨域隐私向量聚合操作可进一步通过密文压缩技术进行优化。为保证Paillier加密方案的安全性，其安全参数通常需至少设置为512 bit，这意味着可将向量中的多个元素拼接为一个大整数参与运算。例如，若将每个元素视作一个50 bit的整数，则可将向量中的每10个元素压缩为一个整数，从而减少Paillier加密和解密的次数与密文的传输成本。

（4）复杂度分析

在算法1中，各隐私向量的初始化加密步骤（第1~3行）可以在$O(n)$时间内完成，其中，n为各参与方所持有的元组数量。随后的同态加法与同态数乘运算（第4~6行）需要花费$O(nm)$的时间。最后对密文$\varepsilon(\mathrm{res})$的解密操作可在$O(m)$的时间内完成。因此，该跨域隐私向量聚合操作的复杂度为$O(nm)$。

2.3 基于跨域隐私向量聚合的查询算法

本节将基于跨域隐私向量聚合操作给出完整的联邦k-支配Skyline查询算法，其核心思想为对每个元组p_b通过参与方本地的统计与参与方之间的跨域隐私向量聚合操作得到向量**res**。最后根据向量**res**中是否包含大于0的元素来判断元组p_b是否为联邦k-支配Skyline查询的结果。

（1）算法实现

算法2详细阐述了本文基于跨域隐私向量聚合的联邦k-支配Skyline查询算法。

首先将k-支配Skyline查询结果初始化为空集（第1行）；对于每个元组p_b，每个参与方首先根据本地模式$\mathcal{A}_{s_i}$计算小于等于p_b的属性个数向量$\mathbf{CLE}^{s_i}$与严格小于p_b的属性个数向量$\mathbf{CSL}^{s_i}$（第4~5行）；然后，各参与方联合起来执行跨域隐私向量聚合协议。并对聚合结果**res**中的元素逐一进行检查判断p_b是否属于k-支配Skyline查询结果（第6~8行）。在此检查过程中，若发现**res**中的某一元素大于0，即可立即中止该检查过程，因为已可得知存在某一元组可-支配元组p_b。只有当**res**中所有元素均为0时，才将其添加至查询结果之中。

算法2：联邦k-支配Skyline查询伪代码

Input: 数据联邦S，全体数据集D与数据模式$\mathcal{A}$

Output: 联邦k-支配Skyline查询结果ans

1: $\text{ans} \leftarrow \varnothing$

2:for $p_b \in D$ do

3:for $s_i \in S$ do

4: $\mathbf{CLE}^{s_i} \leftarrow$ 在模式$\mathcal{A}_{s_i}$中小于等于p_b的属性个数

5: $\mathbf{CLS}^{s_i} \leftarrow$ 在模式$\mathcal{A}_{s_i}$中严格等于p_b的属性个数

6:所有参与方进行跨域隐私向量聚合得到**res**

7:if **res**中不存在大于0的元素 then

8: $\text{ans} \leftarrow \text{ans} \bigcup \{b\}$

9:return ans

（2）复杂度分析

算法2中第3~5行在各参与方本地执行，复杂度为$O(nd_{s_i})$。第6行对应算法1中的跨域隐私向量聚合操作，其时间复杂度和通信开销均为$O(nm)$。第7~8行则可在$O(n)$时间内完成。由于对每个元组均需执行上述流程，故算法2的时间复杂度和通信开销为$O(n2m)$。

3 实验评估

本节在真实数据集与合成数据集上分别对联邦k-支配Skyline查询算法进行实验评估。

3.1 实验设置

（1）数据集设置

参照现有Skyline查询的研究[1,5,10]，本文选取两个具有不同分布的合成数据集与一个关于NBA球员信息的真实数据集进行实验。

- 合成数据集：本文生成了具有独立（IND）与正相关（COR）分布的元组集。具体而言，IND数据集按照均匀分布独立地生成所有元组值。而COR数据集的生成方式为：首先使用正态分布选择一个垂直于从(0,…,0)到(1,…,1)的直线的平面；然后使用正态分布来选择各个平面，以便中间的点多于两端的点；在平面内，各属性值再次使用正态分布生成，从而确保大多数点靠近从(0,…,0)到(1,…,1)的直线。
- 真实数据集：为了在实际应用中评估本文设计的联邦k-支配Skyline查询算法，本文收集了2002年至2022年包括4 000名NBA球员在内的统计数据，其中每个球员信息都拥有20个属性，包括出场数（GP）、得分（PTS）、投篮命中率（FG）、篮板数（REB）、抢断数（STL）等。

为了模拟真实世界中的Skyline查询，本文将数据集的元组数的大小n的变化范围设置为100~4 000。为了分析参与方数量m对算法效果的影响，本文参照之前的研究方法，将元组等量划分来形成多个参与方[3,4,10]。默认情形下，每方属性数为1。为了进一步评估算法效率，本文将每方持

有的属性数(参与方数据的维度)变化范围设置为1~5。

(2)评估指标

本文选取以下两个指标来对查询处理效率进行评估。

● 运行时间:表示从接收到FKDS查询请求到返回查询结果所用的时间。

● 通信成本:表示FKDS查询过程中跨各参与方之间的网络通信之和。

(3)对比算法

本文对下列算法进行详细的实验对比。

● MP-SPDZ:一种基于最优通用SMC库MP-SPDZ[13]实现的基线FKDS查询算法。

● PVA:本文所提基于跨域隐私向量聚合的FKDS算法(算法2)。

● PVA+:使用密文压缩优化的基于跨域隐私向量聚合的FKDS算法。

(4)实验环境

本实验在至多10台服务器的集群上进行评估,每台服务器使用4个3.10 GHz Intel(R) Xeon(R) Platinum 8269CY 处理器,其内存容量为32 GB,操作系统为Ubuntu 18.04 LTS。机器之间的网络带宽最高为6 Gbit/s。所有算法均使用GNU C++实现,其中大整数运算均采用GMP库实现。Paillier同态加密方案的安全参数为$\kappa = 512$ bit。

3.2 实验结果与分析

(1)参与方数量对算法性能的影响

图1展示了在参与方数量变化时各对比算法的实验表现。可以发现,随着参与方数量不断增加,所有联邦k-支配Skyline查询算法的运行时间均逐渐增加。这是由于参与方数量增加会导致涉及k-支配计算的元组维度增加,从而增加了隐私向量聚合阶段中隐私向量的计算成本。在所有对比算法中,PVA+算法的性能最优,其次是PVA算法。MP-SPDZ算法的性能最差,如当参与方数量大于等于5时,该算法已无法在12 h之内得出查询结果。相较于PVA算法和MP-SPDZ算法,PVA+的执行效率

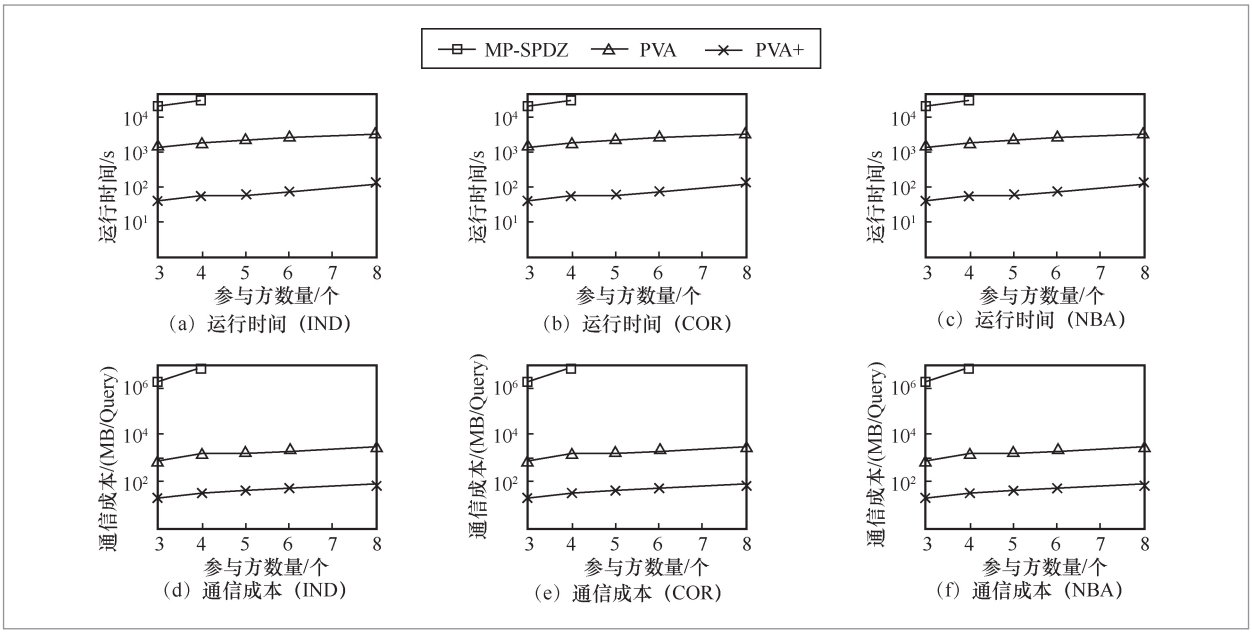

图1 参与方数量变化时各算法性能比较

分别比这两种算法快486.1倍和48.3倍。就通信成本而言，随着参与方数量的增加，PVA与PVA+的通信成本均缓慢增加。实验结果表明，与MP-SPDZ算法相比，基于隐私向量聚合的FKDS算法可将通信成本降低2个数量级。此外，本文所提密文压缩技术可进一步减少通信成本。

（2）数量总量对算法性能的影响

图2展示了在数据联邦的数量总量变化时各对比算法的实验表现。类似地，所有联邦k-支配Skyline查询算法的运行时间随着数量总量的增大均逐渐增加，且MP-SPDZ算法的性能表现依旧是最差的。PVA+的运行效率最高，如处理2 000条元组的联邦k-支配Skyline查询仅需不到15 min，而MP-SPDZ算法在处理500条元组时，查询时间已超过12 h。PVA+算法的执行效率相较于MP-SPDZ算法和PVA算法，分别加速了739.3倍与53.6倍。所有对比算法的通信成本随数据总量的增大均呈近线性增长。与MP-SPDZ算法相比，基于隐私向量聚合的FKDS算法（VPA和VPA+）可将通信成本降低2个数量级。

（3）属性个数对算法性能的影响

图3展示了在各参与方持有属性个数变化时各对比算法的实验表现，此时所有联邦k-支配Skyline查询算法的运行时间均较为稳定。在所有对比算法中，PVA+的运行效率依旧最高，其执行效率相较于MP-SPDZ算法和PVA算法，分别加速了684.8倍与35.3倍。就通信成本而言，MP-SPDZ算法的通信成本依旧比基于隐私向量聚合的FKDS算法（VPA和VPA+）高2个数量级。

本文的实验结果总结如下。

- 本文所提基于隐私向量聚合的联邦k-支配Skyline查询算法的执行效率显著优于使用通用SMC库实现的基线算法。特别地，PVA+算法的执行效率相较于MP-SPDZ和PVA算法提升了739.3倍与53.6倍。
- 本文所提基于隐私向量聚合的联邦k-支配Skyline查询算法具有较低的通信

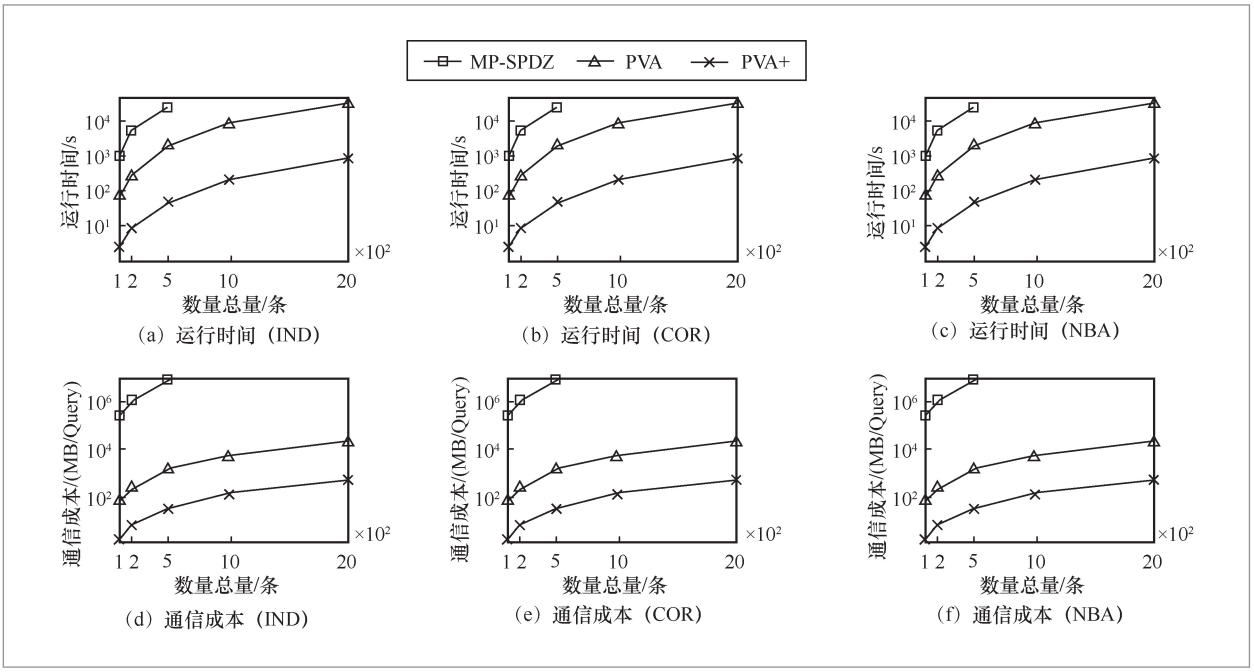

图 2 数量总量变化时各算法性能比较

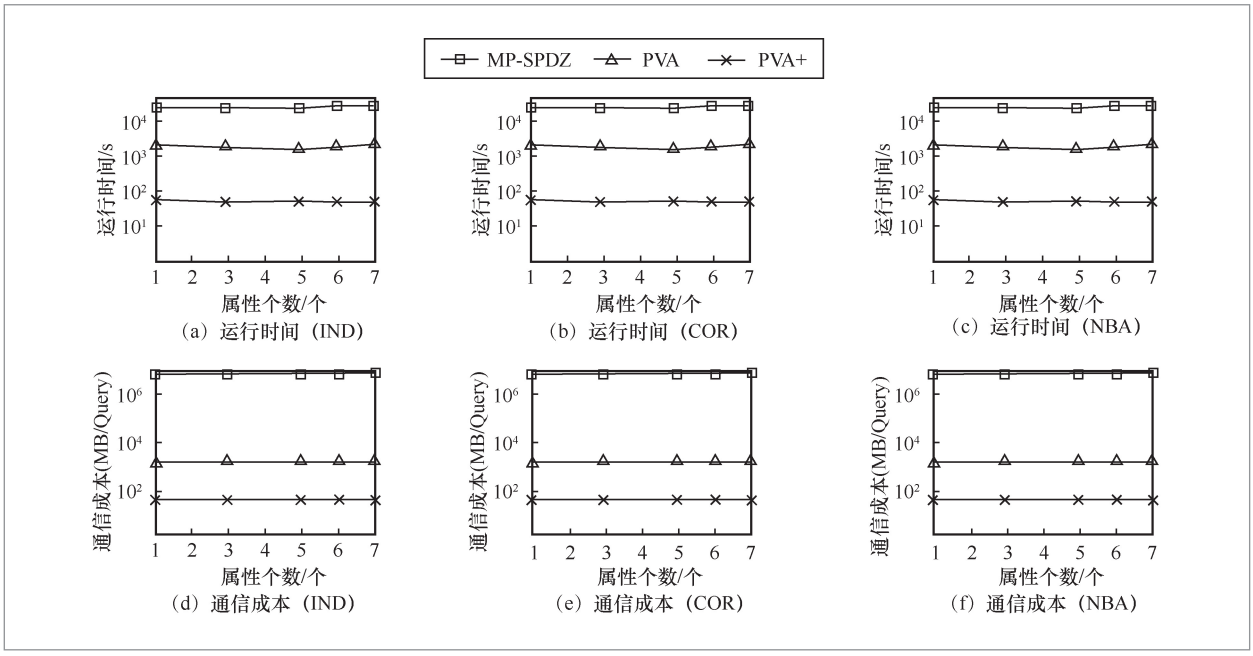

图 3 各参与方持有属性个数变化时各算法性能比较

成本。例如，PVA与PVA+算法的平均通信成本约为70.1 GB，而MP-SPDZ算法处理一条联邦k-支配Skyline查询的通信成本可达40 TB。本文所提基于隐私向量聚合的联邦k-支配Skyline查询算法可将通信成本降低至少2个数量级。

4 相关工作

本节将从k-支配Skyline查询处理和数据联邦查询处理两个方面介绍本文的相关工作。

4.1 k-支配Skyline查询处理

Skyline查询操作最早由Borzsony等人[1]于2001年提出，其提出了一种基于分治策略的Skyline查询处理框架。然而，在数据维度较高时，Skyline查询的结果集合往往较大，难以反映数据之间的支配关系。为了解决这一问题，许多Skyline问题的变种被提出[5,13-15]。其中，k-支配Skyline查询[5]是一种常见的变种问题，其将支配关系的属性约束放松至k条，从而提升高维数据中的查询效用。基于这一概念，Lee等人[16]提出可利用Z-曲线实现快速k-支配Skyline查询，Siddique等人[17]则提出了一种排序-过滤的方法。Miao等人[7]进一步研究了在不完全数据中进行k-支配Skyline查询的方法。

近年来，已有一些工作研究在不可信的云平台上安全执行Skyline查询的方法。Chen等人[18]研究了如何验证云平台返回Skyline查询结果的正确性，Liu等人[8]则提出了一种基于同态加密的安全Skyline查询算法，Ding等人[9]与Liu等人[19]进一步将安全Skyline查询扩展到多方场景下。这些工作均利用Skyline查询中支配关系的传递性进行算法优化。然而，在k-支配Skyline查询中，k-支配关系不具有传递性，因此无法直接将现有研

究扩展到安全k-支配Skyline查询中。

4.2 数据联邦查询处理

作为一种数据共享的新范式，数据联邦旨在联合多个参与方的私有数据，在保证原始数据不出域的前提下进行联合查询操作，以达到数据可用不可见的效果[3,4,20]。其在联邦数据库的基础上，采用多方安全计算等安全技术保护在联邦查询过程中各参与方的数据隐私。SMCQL[3]是首个数据联邦系统，其利用通用安全多方计算库支持在两个参与方之间的数据库中安全查询。Conclave[21]则将安全查询扩展至大数据引擎上，并可支持至多3个参与方的安全运算。然而，受限于通用安全多方计算库的效率瓶颈，这些系统的查询效率仍明显慢于明文查询效率。为了解决这一问题，SAQE[22]选择在数据联邦中进行近似查询，虎符系统则通过设计查询重写机制与专用安全多方计算协议的方法提升联邦查询效率。这种数据联邦查询的通用框架同样可解决联邦k-支配Skyline查询，如利用通用安全多方计算库MP-SPDZ[13]进行计算求解。然而，根据本文的实验结果，这种通用方法的计算与通信效率较低，难以适用于大规模的数据联邦推荐场景中。

5 结束语

本文聚焦于联邦k-支配Skyline查询这一新问题，旨在进行跨信任域的联邦信息推荐，具有广泛的应用前景。为了实现高效的联邦k-支配Skyline查询，本文提出了一种基于隐私向量聚合的算法，其可将k-支配关系转化为隐私向量之间的聚合运算，从而利用专用的安全多方计算协议提升查询效率。此外，本文还提出了一种密文压缩技术进一步优化查询效率。在合成数据集与真实数据集上的实验表明，本文所提算法相较于基于通用安全多方计算库的方法而言，查询效率提升了739.3倍，并可将通信成本降低至少两个数量级。

参考文献:

[1] BORZSONY S, KOSSMANN D, STOCKER K. The skyline operator[C]//Proceedings of Proceedings 17th International Conference on Data Engineering. Piscataway: IEEE Press, 2002: 421-430.

[2] SHARIFZADEH M, SHAHABI C. The spatial skyline queries[C]//Proceedings of the 32nd International Conference on Very Large Data Bases. New York: ACM Press, 2006: 751-762.

[3] BATER J, ELLIOTT G, EGGEN C, et al. SMCQL: secure query processing for private data networks[J]. Proceedings of the VLDB Endowment, 2017, 10(6): 673-684.

[4] TONG Y X, PAN X C, ZENG Y X, et al. Hu-fu[J]. Proceedings of the VLDB Endowment, 2022, 15(6): 1159-1172.

[5] CHAN C Y, JAGADISH H V, TAN K L, et al. Finding k-dominant skylines in high dimensional space[C]//Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data. New York: ACM Press, 2006: 503-514.

[6] AWASTHI A, BHATTACHARYA A, GUPTA S, et al. k-dominant skyline join queries: extending the join paradigm to K-dominant skylines[C]//Proceedings of 2017 IEEE 33rd International Conference on Data Engineering (ICDE). Piscataway: IEEE Press, 2017: 99-102.

[7] MIAO X Y, GAO Y J, CHEN G, et al. k-dominant skyline queries on incomplete data[J]. Information Sciences, 2016,

367/368: 990–1011.

[8] LIU J F, YANG J C, XIONG L, et al. Secure skyline queries on cloud platform[J]. Proceedings International Conference on Data Engineering, 2017, 2017: 633–644.

[9] DING X F, WANG Z, ZHOU P, et al. Efficient and privacy–preserving multi–party skyline queries over encrypted data[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4589–4604.

[10] ZHANG Y Y, SHI Y X, ZHOU Z M, et al. Efficient and secure skyline queries over vertical data federation[J]. IEEE Transactions on Knowledge and Data Engineering, 2022(99): 1–12.

[11] KELLER M. MP–SPDZ: a versatile framework for multi–party computation[C]// Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2020: 1575–1590.

[12] KATZ J, LINDELL Y. Introduction to modern cryptography[M]. [S.l.]: Chapman and Hall/CRC, 2020.

[13] GAO Y J, MIAO X Y, CUI H Y, et al. Processing k–skyband, constrained skyline, and group–by skyline queries on incomplete data[J]. Expert Systems With Applications, 2014, 41(10): 4959–4974.

[14] GAO Y J, LIU Q, ZHENG B H, et al. On efficient reverse skyline query processing[J]. Expert Systems With Applications, 2014, 41(7): 3237–3249.

[15] LIN X M, YUAN Y D, ZHANG Q, et al. Selecting stars: the k most representative skyline operator[C]//Proceedings of 2007 IEEE 23rd International Conference on Data Engineering. Piscataway: IEEE Press, 2007: 86–95.

[16] LEE K C K, ZHENG B H, LI H J, et al. Approaching the skyline in Z order[C]// Proceedings of the 33rd International Conference on Very Large Data Bases. New York: ACM Press, 2007: 279–290.

[17] SIDDIQUE M A, MORIMOTO Y. k–dominant skyline computation by using sort–filtering method[C]//Pacific–Asia Conference on Knowledge Discovery and Data Mining. Heidelberg: Springer, 2009: 839–848.

[18] CHEN W X, LIU M J, ZHANG R, et al. Secure outsourced skyline query processing via untrusted cloud service providers[C]//Proceedings of IEEE INFOCOM 2016 – The 35th Annual IEEE International Conference on Computer Communications. Piscataway: IEEE Press, 2016: 1–9.

[19] LIU X M, CHOO K K R, DENG R H, et al. PUSC: privacy–preserving user–centric skyline computation over multiple encrypted domains[C]//Proceedings of 2018 17th IEEE International Conference on Trust, Security and Privacy, In Computing and Communications/ 12th IEEE International Conference on Big Data Science and Engineering (TrustCom/ BigDataSE). Piscataway: IEEE Press, 2018: 958–963.

[20] 李书缘, 季与点, 史鼎元, 等. 面向多方安全的数据联邦系统[J]. 软件学报, 2022, 33(3): 1111–1127.
LI S Y, JI Y D, SHI D Y, et al. Data federation system for multi–party security[J]. Journal of Software, 2022, 33(3): 1111–1127.

[21] VOLGUSHEV N, SCHWARZKOPF M, GETCHELL B, et al. Conclave: secure multi–party computation on big data[C]// Proceedings of the 14th EuroSys Conference 2019. New York: ACM Press, 2019: 1–18.

[22] BATER J, PARK Y, HE X, et al. SAQE: Practical privacy–preserving approximate query processing for data federations[J]. Proceedings of the VLDB Endowment, 2020, 13(12): 2691–2705.

作者简介

史烨轩（1994– ），男，博士，北京航空航天大学计算机学院博士后，主要研究方向为大数据分析处理、联邦学习和隐私计算。

童咏昕（1982– ），男，北京航空航天大学计算机学院教授，主要研究方向为联邦学习、隐私计算、时空大数据分析、数据库技术和群体智能。

周昊（1999– ），男，北京航空航天大学计算机学院硕士生，主要研究方向为大数据分析处理、联邦学习和隐私计算。

许可（1971– ），男，北京航空航天大学计算机学院教授，主要研究方向为算法与复杂性、数据挖掘和群体智能等。

吕卫锋（1972– ），男，北京航空航天大学计算机学院教授，北京航空航天大学副校长，软件开发环境国家重点实验室副主任，主要研究方向为时空大数据分析、智慧城市和群体智能等。

收稿日期： 2023-02-28

基金项目： 国家自然科学基金资助项目（No.U21A20516，No.62076017）；北航基础研究建设基金资助项目（No.YWF-22-L-531）；微众学者计划

Foundation Items: The National Natural Science Foundation of China (No.U21A20516, No.62076017), Beihang University Basic Research Funding (No.YWF-22-L-531), WeBank Scholars Program