支持分销渠道绩效评价的个体化差分隐私 Skyline 查询研究

兰秋军，马佳琪

（湖南大学工商管理学院，湖南省长沙市，410000）

摘要： Skyline 查询在分销渠道绩效评价中可以有效应用，但是其潜在的隐私风险是一个值得关注的现实问题。差分隐私(DP)是一种严格的隐私保护方法，因其鲁棒性和可靠性而成为近年来的研究热点。考虑到差分隐私会降低数据的可用性性，本文提出了基于谱聚类的个体化差分隐私保护(iDP-SC)来解决 Skyline 查询中的隐私泄露问题。该方法将局部敏感度的计算基础从原始数据集转移到经过谱聚类处理的数据集上，在降低敏感度的同时降低了噪声。因此，它在不牺牲差分隐私所提供的隐私保护的前提下保持了更高的数据效用。此外，与现有的隐私保护 Skyline 查询方法相比，该方法避免了关键信息的泄露，同时提供了隐私保护水平的定量分析。通过在真实数据集和合成数据集上与差分隐私(DP)和个体化差分隐私(iDP)进行比较，验证了所提出的 iDP-SC 的性能。实验结果表明了该方法的优越性。

关键词： 隐私保护，差分隐私，Skyline 查询，谱聚类

中图分类号： C93　　文献标识码：A

1　引言

Skyline 查询在多准则决策中得到了广泛的应用[28]。在多维数据集中，Skyline 查询返回不受任何其他对象支配的 Skyline 对象集[7]。Skyline 查询的一个有趣的应用是供分销渠道绩效评价。例如，在一个以供应商为主导的两级分销渠道中，为了制定产品的定价策略，供应商需要对分销商的业绩进行综合评价。对于分销商来说，他们还需要从不同的角度了解自己在所有竞争对手中的相对表现，这可以帮助他们调整管理策略以提高绩效。作为一个强大的数据分析工具，Skyline 查询从来自不同分销商的大量绩效数据中检索出一组最佳绩效数据。然而，分销商提交的绩效数据可能包含敏感信息，Skyline 查询过程可能泄露隐私信息，从而导致商业机密的泄露。因此，如何保证 Skyline 查询的隐私性是需要我们关注的现实问题。

为了防止 Skyline 查询中隐私信息的泄露，一个可靠的隐私保护方案是必不可少的。常用的隐私保护技术大致可分为三类：匿名化、加密和数据失真。其中，匿名化方法，如 k -匿名、l-多样和 t-近邻，会匿名化个人信息的显式标识符；但当对手有特定的背景知识时，隐私保护将失效。加密技术被广泛用于隐私保护的 Skyline 查询。但是，如果加密密钥被泄漏或被暴力破解，则加密同样会失效。此外，基于加密数据的计算可能会导致一些关键信息的泄露，隐私保护的程度也无法定量分析。因此，我们需要一种更强大、更可

靠的隐私保护技术。差分隐私等数据失真技术利用校正过的噪声对原始数据进行干扰，从而保护个人隐私。差分隐私作为一种严格的隐私模型，即使攻击者拥有最大程度的背景知识也能抵抗各种形式的隐私攻击。

据我们所知，在隐私保护下的 Skyline 查询的相关研究中，只有 Qaosar 等人近年来的文献采用了差分隐私技术，他们提出的方案生成随机向量并随机打乱对象对列表，但这并不严格遵循随机响应机制。尽管差分隐私由于其健壮性和可靠性在隐私保护方面很有前景，但会较大地损害数据效用。为了解决这一问题，一些研究试图提高差分隐私的数据效用(Dwork &Rothblum, 2016;Dwork 等，2006;Machanavajjhala 等，2008)。总体而言，现有的研究主要分为两类:第一类文献提出了一些新的机制，通过设置合理的隐私预算(Li 等，2017)，或通过降低敏感度(Soria-Comas 等，2014;S anchez 等，2016)来提高数据效用。第二类文献的研究重点是视图找到差分隐私的放松。这一类文献通常会在数据效用和隐私之间做出权衡。然而，没有任何一篇文献能够比较理想地维护数据效用，尤其是在隐私预算较小的情况下。

最近，个体化差分隐私(Individual Differential Privacy, iDP) (Soria-Comas 等，2017)通过采用直观的隐私保障为解决上述文献的局限性提供了灵感。具体来说，它将差分隐私要求的查询结果的不可区分性从任意一对相邻数据集转换为实际数据集及其相邻数据集。因此，iDP 为数据集中的每个个体提供了与差分隐私相同的隐私保障，同时增加了差分隐私的实用性。本文试图在不牺牲 Skyline 查询的隐私保障的前提下尽可能地保留数据效用。为了实现这一目标，我们利用了 iDP 和谱聚类技术，并提出了基于谱聚类的个体化差分隐私。在该算法中，局部敏感度的计算基础从原始数据集转移到经过谱聚类处理的数据集。因此，敏感度可以进一步降低，进而降低噪声。本文的主要贡献可以总结如下：

- 针对 Skyline 查询中存在的隐私泄露问题，提出了一种基于差分隐私的新算法，以弥补传统研究的不足。
- 所提出的 iDP-SC 算法在不降低差分隐私保护强度的情况下,在很大程度上保持了数据的实用性，并且在很小的隐私预算下，噪声查询结果可以接近非隐私查询结果。
- 我们从隐私性和效用两个方面对所提出的 iDP-SC 进行了理论分析。
- 在多个数据集上进行的实验表明，该算法的性能优于 DP 和 iDP 算法。
- 同时，我们研究了 iDP-SC 中关键参数对数据效用的影响，并在此基础上给出了算法的推荐参数。

本文的其余部分组织如下。第二节回顾了相关理论及文献综述，总结了研究现状。第三节阐述了基于差分隐私的 Skyline 查询隐私保护问题，确定了算法的期望目标，并给出了建议的解决方案和理论分析。第 4 节我们评估了所提算法的性能，并与其他算法进行了比较。最后，本文在第五部分进行了总结，并对未来的工作提出了展望。

2 相关理论与文献综述

2.1 Skyline 查询

支配：给定 m 维空间中的数据集 $D=\{d_1,\Lambda\ ,d_n\}$，d_a 和 d_b 是数据集 D 中的两个不同的点，如果满足以下条件，则称 d_a 支配 d_b：

a. $\forall 1\le k\le m$，$d_a[k]\le d_b[k]$

b.至少存在一个 k，使得 $d_a[k]<d_b[k]$，其中 $d_i[k]$ 是 d_i 的第 k 维，且 $1\le k\le m$。

严格支配：给定 m 维空间中的数据集 $D=\{d_1,\Lambda\ ,d_n\}$，d_a 和 d_b 是数据集 D 中的两个不同的点，如果对于 $\forall 1\le k\le m$，$d_a[k]<d_b[k]$，则称 d_a 严格支配 d_b。

Skyline 点：给定 m 维空间中的数据集 $D=\{d_1,\Lambda\ ,d_n\}$，对于数据集 D 中的任意一点 d_i，如果 d_i 不被数据集中的其他任意点支配，则称 d_i 为集合 D 的 Skyline 点，其中集合 D 所有 Skyline 点构成的集合记为 Skyline 查询的结果，记作 $Sky(D)$。

Skyline 查询：给定 m 维空间中的数据集 $D=\{d_1,\Lambda\ ,d_n\}$，Skyline 查询是筛选出集合 D 所有 Skyline 点并构成集合 $Sky(D)$ 的过程。

巧妙运用 Skyline 查询的某些执行，可以大大提高 Skyline 查询的计算速度，其中较为常用的就是 Skyline 查询的可加性。

Skyline 查询的可加性：假设数据集 D 由 n 个数据集的并集组成，则有：

$$Sky(D)=Sky(Sky(D_1)\Upsilon\Lambda\ \Upsilon\Lambda\ Sky(D_n)) \tag{2.1}$$

这基于 D 的所有子集的 Skyline 查询结果，再次执行 Skyline 查询的结果，跟直接在 D 上计算的 Skyline 查询结果是一样的。

在 Skyline 查询中，局部 Skyline 对象（LSO）表示子集中非支配对象的集合，即 $Sky(D_i)$，全局 Skyline 对象（GSO）表示子集的并集中非支配对象的集合，即 $Sky(D)$。

2.2 差分隐私

差分隐私是 Dwork2006 年提出的一种隐私保护模型。在这个模型下，查询结果不会因为数据集中的单个记录的改变而发生显著变化，这使得攻击者很难通过观察查询结果来获得准确的个体信息。差分隐私法几乎不假设对手的背景知识，同时还为隐私保护程度提供了严格的数学证明。这两个特点使其迅速成为研究热点。

ε-差分隐私：若随机机制 M 对任意一对相邻数据集 D 和 D'，都满足:

$$\Pr[M(D)\in S]\le e^{\varepsilon}\times\Pr[M(D')\in S], \tag{2.2}$$

则称算法 M 满足 ε-差分隐私。其中，$S\subset Range(M)$，表示算法 M 所有可能的输

出结果；ε 表示隐私预算，它用来衡量隐私保护的强度，越小的 ε 值说明算法的隐私保护强度越高。

我们可以通过在查询结果中添加适当数量的噪声来实现差异隐私。敏感度是决定加噪量的关键参数。

全局敏感度：对于一个函数 f ： $D \to R$，以及在一条记录中不同的所有数据集 D 和 $D^{'}$，函数 f 的全局敏感度为：

$$GS_f = \max_{D,D^{'}} \left\| f(D) - f(D^{'}) \right\|_1 \tag{2.3}$$

全局敏感度只与查询函数 f 有关，全局敏感度越大，表示 $f(D)$ 与 $f(D^{'})$ 的查询结果差异越大，需要添加更多的噪声来隐藏差异。

局部敏感度：对于函数 f ： $D \to R$，以及所有在一条记录中不同的数据集 D 和 $D^{'}$，函数 f 的局部敏感度为：

$$LS_f(D) = \max_{D^{'}} \left\| f(D) - f(D^{'}) \right\|_1 \tag{2.4}$$

对于给定的数据集 D，局部敏感度显示了函数在 D 和邻近数据集 $D^{'}$ 之间的可变性。与全局敏感度不同，局部敏感度是由函数 f 和给定的数据集 D 决定的。由于使用了数据集的数据分布特征，局部敏感度通常比全局敏感度小得多。更直观地说，全局敏感度是局部敏感度的上界。

拉普拉斯机制是一种常见的差分隐私机制。它通过在真实的查询结果中加入适量的噪声来满足差分隐私性，其中噪声服从拉普拉斯分布。

拉普拉斯分布：均值为 0、标准差为 b 的拉普拉斯分布的概率密度函数为：

$$Lap(x|b) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right) \tag{2.5}$$

可将上述概率密度简记为 $\mathrm{Lap}(b)$。

拉普拉斯机制：对于数据集 D 上的函数 f : $D \to R$，其敏感度为 Δf，如果机制 $M(D)$ 的输出满足等式 2.10，则机制 M 满足差分隐私。

$$M(D) = f(D) + Lap\left(\frac{\Delta f}{\varepsilon}\right) \tag{2.6}$$

2.3 文献综述

2.3.1 隐私保护下的 Skyline 查询

信息技术的发展和提高，给人们的生活带来了很多变化。数据库技术已经渗透到社会生活的各个领域，和人们的日常生活有着密切的联系。大到国防军事，小到出行购物，每

时每刻都有大量的数据产生。面对海量的数据，人们希望从中找到自己感兴趣的信息。然而，有效地对数据进行分析，从这些高维、庞大的数据中准确地获取自己所需要的信息是相当困难的，特别是在人们的决策目标并非单一的情况下。针对以上问题， S.Borzsony 等人在 2001 年的数据工程会议（ICDE）上第一次提出 Skyline 查询的概念[7]，并将 Skyline 计算应用到数据库领域。从此，Skyline 查询便得到数据库领域众多学者的关注，成为数据库查询领域的研究重点与难点之一。

随着人们隐私意识的增强，数据隐私成为近十年来的研究热点。在传统的 Skyline 查询中，如果不将原始数据发布给其他方，就不可能得到多方的 Skyline 查询结果。但这样的披露可能被潜在的攻击者利用，从而获取更多隐私信息。因此，保证 Skyline 查询的安全性和隐私性至关重要。然而，只有为数不多的研究关注隐私保护下的 Skyline 查询。现有的隐私保护 Skyline 查询方案主要采取加密技术，可分为两类方法：一类是根据每个属性上对象的顺序来比较支配关系;另一类方法是对加密值使用排列和比较等技术来比较支配关系。

（1）依据属性顺序比较支配关系

在使用第一类方法的相关研究中，Zenginler 展示了保留顺序的加密方案（OPE）[40]。Zaman A 等（2016）提出的方法被设计用于 MapReduce/Hadoop 框架[31]。在这个框架中，各方与协调器协作，在每个属性上构造数据库对象的顺序。经过几轮排序后，生成每个属性上对象的顺序，以计算 Skyline 查询结果，这确保了不会暴露对象的单个属性值。最近，Qaosar 等人（2019）提出了一种基于同态加密的安全多方排序协议[32]。通过引入 paillier 加密系统，在不改变对象的值在每个属性上的顺序的情况下，对对象属性值进行转换。因此，可以在不公开每个属性上对象的原始值的情况下计算出 Skyline 查询结果。

（2）使用排列和比较加密值比较支配关系

第二类方法的研究重点是使用排列和比较加密值的技术来比较支配关系。Liu X 等人（2016）提出的 EPSC 框架以一种保护隐私的方式实现了多域查询结果的交互式计算[26]。这项工作通过向量比较、整数比较和对加密数据的排列决定支配关系。这三种技术包含在一个名为 ESVC 的协议中，ESVC 在比较两个整数向量时采用了 0 编码和 1 编码方案，因此它取决于二进制位中属性值的长度。Qaosar M 等人（2019）对 ESVC 中的方案进行了改进，采用了基于本地差分隐私(LDP)的数据匿名化、扰动和随机化技术来代替 ESVC 的方案[32]。Bothe 等人（2014）提出了 eSkyline，一种处理加密数据的 Skyline 查询的新系统。在他们的系统中，支配关系被转化为加密元组之间的标量积计算，并通过随机矩阵来保证每个元组的安全性，这样就可以在不暴露实际值的情况下输出 Skyline 查询结果[8]。Liu 等人（2017）提出了一种利用置换和扰动技术计算两个加密元组间支配关系的安全 Skyline 协议，它确保了云服务器对数据的不可知性[25]。Chen 等人（2016）构建了一个基于位置的 Skyline 查询框架，在将数据集外包给第三方云服务提供商之前，服务提供商可以使用加密

技术将每个兴趣点记录下来，并与其所有 Skyline 邻居和相应的查询范围绑定在一起[10]。鉴于用户对不同数据维度的偏好不同，Liu 等人（2018）提出了基于多个加密域的以用户为中心的 Skyline 计算框架（PUSC）[27]。它通过利用一种安全的用户定义的向量支配协议来比较两个加密向量之间的支配关系，该协议允许用户根据自己的喜好选择一组 Skyline 对象，并且不会泄露隐私。Hua 等人（2019）提出了在线医疗诊断系统 CINEMA[20]。通过安全置换和比较技术，可以在加密查询的基础上准确计算出 Skyline 查询结果，同时为 Skyline 诊断模型和用户医疗数据提供隐私保护。

（3）现有工作的局限性

然而，现有的隐私保护下的 Skyline 查询工作存在一些局限性:

首先，依据属性顺序比较支配关系的研究工作无法提供足够的隐私保护。因为每个属性上对象的顺序包含大量的信息[47]。攻击者可以通过学习这些信息来获取更多的隐私。其次，现有的工作大多集中在如何在不披露原始值的情况下获得 Skyline 查询结果。但是，一个重要的隐私隐患仍然存在:不同用户的每对比较对象之间的支配关系会被相互揭示。根据这一知识，潜在的攻击者可能推断出更多的隐私信息。第三，在上述工作中，隐私保护的程度是无法量化的。用户无法得到隐私保护水平的反馈，也无法根据自己的隐私偏好定制隐私保护的强度。

综上所述，如何保证基于 Skyline 查询的分销渠道绩效评价的隐私性是我们需要关注的现实问题。但目前只有少数工作涉及到隐私保护下的 Skyline 查询，而且大多数工作都是基于加密技术进行的，存在较多局限性。针对以上问题和挑战，本文利用差分隐私技术研究如何解决基于 Skyline 查询的分销渠道绩效评价中的隐私隐患。

2.3.2 提高差分隐私的数据效用

近年来，差分隐私已经成为一种强大的隐私保护模型，能够在最大程度上利用攻击者的背景知识来抵御各种形式的攻击。虽然差分隐私因其在隐私保护方面的鲁棒性和可靠性而十分有发展潜力，但它存在着数据效用较低的弊端，影响了后续数据分析的准确性。为了解决这一问题，已经有一系列的研究开始关注提高差分隐私的数据效用[15]。总的来说，现有的研究主要可以分为两大类:第一类是提出一种新的差分隐私机制，第二类则是尝试放松差分隐私。

（1）提出新的差分隐私机制

使用第一类方法的研究中，一些机制通过设置合理的隐私预算来提高差分隐私的数据效用，如个性化差分隐私（PDP）[22]，以及个性化差分隐私下基于分区的机制[24]，这些机制根据个人隐私偏好来确定隐私预算。而其他机制则通过降低敏感度来提高查询结果的准确性，其中 David Sa ́nchez 等人近年来发表的一系列关于差分隐私的文章是典型代表，这些研究展示了差分隐私与 K-匿名之间的协同作用，表明如果查询在数据集的 K-匿名版本

上运行，敏感度就可以降低，因此，差分隐私所需要的噪声就会大大降低。

（2）放松差分隐私

在第二种研究中，Cynthia Dwork（2006）中提出的δ-近似 ε-不可区分机制为标准差分隐私增加了额外放松边界δ[17]；2016 年 Cynthia Dwork 又提出了(τ,μ)-集中差分隐私，它允许小概率不满足ε-差分隐私的情况出现，这个概率是由参数μ和τ控制的[15]。类似地，Ashwin Machanavajjhala[29]提出的(δ,ε)-概率差分隐私将这个概率设置为δ。换句话说，实现查询结果的不可分辨性的概率至少为$1-\delta$。

（3）现有研究的限制

现有的工作改进了差分隐私的数据效用，但在应用于隐私保护的 Skyline 查询时仍然存在一些严重的问题。更具体地说，有两个重要的限制：首先，这些研究中的机制削弱了差分隐私的隐私保护力度。因为在某些情况下，它们允许查询结果在一对相邻的数据集上存在显著差异。其次，当隐私保护需求较大，即隐私预算较小时，这些工作仍然会对查询结果产生扭曲，降低后续数据分析的准确性。

3 改进个体化差分隐私下的 Skyline 查询

3.1 问题定义

在本节中，我们将形式化场景，并定义算法应该实现的对象。

3.1.1 问题的公式化定义

本节主要研究如何在保证隐私的前提下，基于分销渠道绩效评估响应分销商的 Skyline 查询。具体来说，我们的场景包括两个部分：供应商和分销商，如图 1 所示。

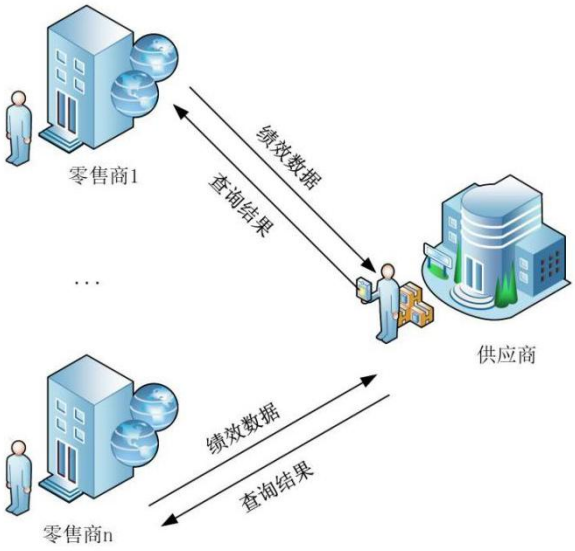

图 3.1 场景描述图

对于供应商来说，分销商的绩效通常会影响其产品的批发价格。在以供应商为主导的两级供应链中，一个供应商对应多个分销商。所以供应商需要从多个角度综合评价分销商的绩效，以便更好地制定下一个周期的产品定价策略。为此，数据管理员的角色由该场景中的供应商承担，他们收集每个分销商的绩效数据，一旦 Skyline 查询响应完成，他们将把经过隐私保护处理的结果发送回请求查询的分销商。

对于分销商来说，他们不仅是查询请求者，还是数据提供者。一方面，对于特定的分销商，我们场景中的其他分销商会被他视为其竞争对手。他们需要通过 Skyline 查询了解自己在所有竞争对手中的绩效表现，从而帮助自己调整管理策略以提高绩效表现。因此，分销商需要向供应商发送 Skyline 查询请求。另一方面，Skyline 查询的结果是基于所有分销商的绩效数据计算的，因此为了实现绩效评估，每个分销商都需要将自己的绩效数据作为查询发送给供应商。

3.1.1 隐私和效用的平衡

根据上述场景和隐私需求，本文旨在设计一种隐私保护性强且数据准确度高的支持分销渠道绩效评价的 Skyline 查询算法。具体地说，应该实现以下目标。

首先，对用于分销渠道绩效评价的 Skyline 查询，本研究提出的算法要具有较高的数据效用。为了提供高质量的隐私保护，本文所设计的隐私保护算法不应牺牲查询结果的数据效用。

此外，该算法还应满足隐私保护的要求。在该场景中，我们假设供应商和所有分销商都是诚实但好奇的[9,12]。具体而言，供应商在情景中严格执行规范，以保证基于 Skyline 查询的绩效评估的准确性，但他却打算从分销商那里获取大量的原始数据，寻找数据背后潜在的规律；分销商诚实地执行操作以获得最终的 Skyline 查询结果，但是他们可能会尝试

获取其他分销商的绩效数据，并获取关于数据之间的支配关系的消息。因此，在我们的场景中，为了保证 Skyline 查询的隐私性，需要满足以下隐私要求。

（1）在响应查询期间，任何分销商都不会向其他分销商披露其绩效数据。

（2）每个分销商只能知道自己的 Skyline 查询结果，任何一方都无法识别其他分销商的查询结果。

（3）每个分销商都不知道别的分销商的特定绩效数据是否支配其他绩效数据。

（4）对于任何分销商来说,它和其他分销商的每一对数据之间的支配关系都是不能相互揭露的。

（5）每个分销商都不应该将他们所有的原始数据发送给供应商。

在不考虑隐私的情况下，某个分销商的大量敏感信息可能被泄露给其竞争对手或供应商。这些人将利用隐私信息试图获利，这将阻止分销商们持续提供其数据进行查询计算。因此，供应商和分销商都无法通过 Skyline 查询得到绩效评价结果。因此，所有人都应满足遵循上述隐私要求。

3.2 个体化差分隐私保护下的 Skyline 查询

3.2.1 个体化差分隐私

由于其在隐私保护方面的健壮性和可靠性，差分隐私很有前途，但它也存在争议，因为它牺牲了数据的效用。在本节中，我们通过使用个体差分隐私(*iDP-SC*)[36]来解决这个问题，在该方法下，数据集中的每个个体都得到与标准差分隐私相同的隐私保证，同时尽可能保留差分隐私的数据效用。本节首先针对我们的场景提出了一种直接实现 *iDP-SC* 的方法，即基于个体差分隐私的差异化 Skyline 查询。然后利用聚类技术对其进行进一步改进，以提高差分隐私的数据利用率。

现有的研究成果在应用于隐私保护的 Skyline 查询时存在局限性:在某些情况下，它们允许查询结果在一对相邻数据集上存在显著差异，这削弱了差分隐私的隐私保护能力[36]。此外，当隐私预算很小时，它们可能会扭曲查询结果。相反，*iDP-SC* 保留了 DP 向个人提供的严格的隐私保证。特别地，*iDP-SC* 认为实际数据集与其相邻数据集之间的不可分辨性就足够了。

（1）个体化差分隐私。给定一个数据集 X，对于 X 的任意一个相邻数据集 $X^{'}$ 和任意一个 $S \subset Range(M)$，如果响应机制 $M(\cdot)$ 满足 ε -个体化差分隐私(或 ε -*iDP*)，那么有：

$$\exp(-\varepsilon)\Pr(M(X^{'}) \in S) \le \Pr(M(X) \in S) \le \exp(-\varepsilon)\Pr(M(X^{'}) \in S) \quad (3.1)$$

正如[36]中证明的，*iDP* 和标准 DP 是有共同点的，即实现查询结果不可区分的概率至少为 $1-\exp(\varepsilon)$。然而，*iDP* 和标准 DP 之间也存在着显著的差异:：*iDP* 只要求实际数据集与

该数据集的任意一个相邻数据集不可区分，而差分隐私则要求任何一对相邻数据集间都是不可区分的。因此，在 *iDP* 中实际数据集 X 与其相邻的数据集 X' 是不可互换的。同时，实现 *iDP* 需要的是以局部敏感度为标准添加噪声，而局部敏感度通常远低于标准 DP 所要求的全局敏感度，这直接导致了 *iDP* 所需要的的噪声量要远小于标准 DP 所需要的的噪声量。

标准 DP 要求的直观隐私保证是，不能从查询结果推断数据集中某个个体的存在与否[13]。*iDP* 保持了标准 DP 这种直观的隐私保证，同时提高了查询结果的准确性。*iDP* 可以达到这样的效果本质上是因为它利用了这样一个事实，即数据管理员在响应查询时其实是可以学习到关于实际数据集的信息的。

3.2.2 个体化差分隐私下的 Skyline 查询

在分销渠道绩效评价的场景中，数据管理员的角色由供应商承担，供应商负责接受每个分销商发送的数据，并响应来自分销商的 Skyline 查询，然后对查询结果进行隐私保护处理，确保做出的响应是不会泄露隐私的。具体来说，实际情况与 *iDP* 核心思想是一致的：供应商并不是在任意一对相邻数据集上去响应 Skyline 查询，而是在他本身持有的实际数据集上进行响应的，因为在响应 Skyline 查询之前，供应商已经知道了当前实际数据集的所有信息。同时，由于供应商不会被分销商视为竞争对手，也不是潜在的攻击者，因此是受到分销商的信任的。基于上述事实，供应商是被允许利用其掌握的实际数据集的信息来调整隐私保护级别。

但是另一方面，如果供应商可以获得过多的原始数据，就有推断出与分销商除绩效数据外其他隐私信息的潜在风险。根据公式（2.1），根据每个分销商的原始数据计算得到的 *GSO* 与根据每个分销商的 *LSO* 计算得到的 *GSO* 是相同的。因此，为了防止这方面的隐私隐患，本文设置了如图 2 所示的加噪机制：首先每个分销商先在其原始数据上进行局部 Skyline 查询，计算得到各自的 *LSO*。然后，分销商仅将其 *LSO* 发送给供应商以计算 *GSO*，而不是所有的原始数据，这个改变避免了分销商的过多原始数据的泄漏，同时不会影响 Skyline 查询结果的准确性。需要指出的是，在本文的场景中，*GSO* 指的是所有分销商中最佳绩效数据的集合，其中的每个数据在任何维度上都不差于任何其他数据。与此相似，对于特定的分销商，*LSO* 表示其该分销商最好的绩效数据集。

如前所述，*LSO* 和 *GSO* 代表两个不同范围内的最佳绩效数据集。对于某个分销商而言，当某些绩效数据只存在于其 *LSO* 中而不存在于 *GSO* 中，意味着其他分销商的绩效数据在某些维度上表现得更好。当该分销商的某些绩效数据同时存在于 *LSO* 和 *GSO* 中时，表明该数据在所有分销商中都是最佳的。同时存在于 *LSO* 和 *GSO* 的绩效数据越多，表明该分销商的绩效在所有分销商中表现越好。也就是说，*LSO* 和 *GSO* 之间的相似度反映了分销商的绩效，这里我们引入 Jaccard 相似度来度量 *LSO* 和 GSO 之间的相似度。它定义为两个集

合的交集大小与并集大小的比值。具体来说，我们使用 *LSO* 和 *GSO* 之间的 Jaccard 相似度来衡量分销商的绩效，即 $J(LSO.GSO)$。

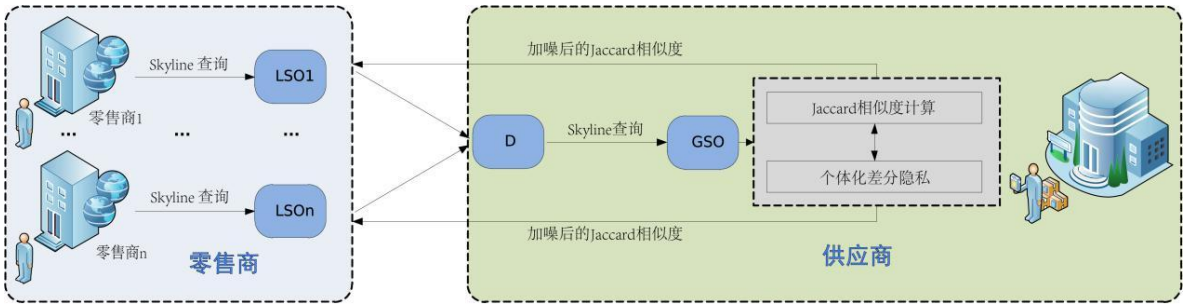

图 3.2　个体化差分隐私保护下的 Skyline 查询

（1）Jaccard 相似度。给定两个集合 S_A 和 S_B，定义它们之间的 Jaccard 相似度如下：

$$j = J(S_A, S_B) = \frac{|S_A \text{ I } S_B|}{|S_A \text{ Y } S_B|} \tag{3.2}$$

表 3.1 基于个体化差分隐私的 Skyline 查询

算法 1：基于个体化差分隐私的 Skyline 查询

输入：$D = \{LSO_1, LSO_2, \Lambda\ LSO_n\}$；$\varepsilon$

输出：满足差分隐私的加噪 Jaccard 相似度 $\{\tilde{j}_1, \Lambda\ \tilde{j}_n\}$

1.进行全局 Skyline 查询得到 *GSO*;

2.对于 $i = 1,2,\Lambda\ ,n$:

3. 计算真实的 Jaccard 相似度：$j_i = J(LSO_i, GSO)$;

4. 计算局部敏感度 $LS_J(D)$;

5. 使得噪声 η 服从拉普拉斯分布 $Lap(0, \frac{LS_J(D)}{\varepsilon})$

6. 计算满足差分隐私的加噪 Jaccard 相似度 $\tilde{j}_i = j_i + \eta$

7.结束

供应商一旦收到来自各分销商的数据 D，即 $\{LSO_1, LSO_2, \Lambda\ , LSO_n\}$，就会对 D 执行全局 Skyline 查询，得到 *GSO*，然后计算 *GSO* 与各分销商 *LSO* 之间的 Jaccard 相似 j。出

于隐私考虑，分销商不希望其绩效数据泄漏给其他分销商，同时希望其他分销商无法识别其查询结果。这些要求可以通过供应商的查询响应操作来实现。然而，如果供应商直接将每个分销商的 Jaccard 相似度发送给他们，会存在一个重要的隐私隐患：假设供应商两次响应了零售商$_i$的查询请求，零售商$_i$知道零售商$_j$的某个特定绩效数据在第二次查询的时候没有提供给供应商做第二次全局 Skyline 查询。然后，如果这两次查询中的零售商$_i$的 Jaccard 相似度不同，则可以确定来自零售商$_j$的该绩效数据不在 *GSO* 中。相反，如果两次查询中分销商的 Jaccard 相似性相同，则可以推断来自零售商$_j$的该绩效数据在 *GSO* 中。与辅助信息结合后，零售商$_i$可以推断出更多关于零售商$_j$的隐私信息，这将成为零售商$_j$隐私信息暴露的导火索。

针对以上对隐私性和查询结果准确性的要求，供应商在计算得到各分销商的 *LSO* 和 *GSO* 之间的 Jaccard 相似度后，根据 *iDP* 执行隐私保护操作。正如前面所讨论的，供应商可以使用信息的实际数据集 $D=\{LSO_1, LSO_2, \Lambda\ LSO_n\}$ 来计算需要添加的噪声量，从而大大提高了查询结果的准确性，这是由通过校准噪声到实际数数据集 D 的局部敏感度实现的。具体来说，当响应分销商的查询时，供应商已经获得了关于数据集 D 的信息。通过删除实际数据集 D 中的任意一条记录，将生成 D 的相邻数据集 D'。然后供应商执行 Skyline 查询并计算 D' 上的相似度，直到找到查询结果的最大变化，即局部敏感度 $LS_J(D)$。之后，根据拉普拉斯机制的定义，对于一个给定的隐私预算 ε，供应商只需要对真实的 Jaccard 相似度 j 添加服从拉普拉斯分布 $Lap(0, LS_J(D)/\varepsilon)$ 的噪声就可以得到满足差分隐私的加噪 Jaccard 相似度 $\tilde{j}$。

3.3 基于谱聚类的改进个体化差分隐私下的 Skyline 查询

为了确保隐私保护力度，差分隐私所设置的隐私预算 ε 通常很小，比如 0.1。在本文中，基于 *iDP* 的 Skyline 查询实际上需要很高的噪声级别，而噪声过多会限制查询结果的可用性。为了进一步提高查询结果的可用性，我们提出了基于谱聚类的个体化差分隐私（Individual Differential Privacy via Spectral Clustering (*iDP-SC*)）。该算法将 *iDP* 中计算的局部敏感度的基础从原始数据集转移到谱聚类处理后的数据集上，形成了基于聚类的局部敏感度。最终，*iDP-SC* 在不牺牲差分隐私保护的前提下进一步降低了敏感度，从而使得噪声量进一步减少。

3.3.1 基于谱聚类的改进个体化差分隐私

近期，Sanchez D 等人（2016）已经证明了差分隐私和 K-匿名之间的协同作用[33-35]。他们表明，如果查询在 K-匿名版本的数据集上运行，敏感度可以进一步降低，相应地，为满足差分隐私所添加的噪声就可以进一步减少。这些研究为本文提供了灵感。但是在这些工作中，算法是独立应用在数据集的每个属性上，这并不适用于需要对数据集的多个属

性进行综合分析的查询，如 Skyline 查询。

为了对数据集的多个属性进行综合分析，同时进一步降低敏感度，我们需要对数据集的版本进行转换，以便在多维数据空间中实现数据查询。由此我们引入谱聚类实现这一目的。谱聚类是一种基于图谱理论[37]的聚类算法。该算法将数据点表示为顶点，将两两相似度表示为对应边的权值，从而得到无向图的权值，最后将聚类重新公式化为图的划分问题。与传统聚类算法相比，谱聚类具有收敛到全局最优解并能在任意形状的数据空间上聚类的能力，适用于多维数据聚类。

为了进一步降低 *iDP-SC* 下的局部敏感度，本文提出了基于谱聚类的质心替换算法（centroid replacement via spectral clustering （CRSC）），如表 3.2 所示。需要指出的是，CRSC 不仅适用于需要对多个维度进行综合分析的查询，也适用于其他的普通查询。

表 3.2 基于谱聚类的质心替换算法

算法 2：基于谱聚类的质心替换算法

输入：原始数据集 $X=\{d_1,\Lambda\ d_n\}$；聚类数 K

输出：谱聚类替换质心后的数据集 $\overline{X}=\{d_{c1},\Lambda\ ,d_{cn}\}$

1.通过谱聚类将数据集 X 聚为 K 簇；

2.对于 $k=1,2,\Lambda\ ,K$：

3. 计算质心;

4. 用质心代替原始数据得到 $\overline{X}$;

5. **结束**

对于一个函数 f 和和有 n 条记录的数据集 X $\{d_1,\Lambda\ d_n\}$，X 的局部敏感度测量了函数 f 的最大变化，该变化是由于 X 的单个记录被移除导致的。在实践中，当数据集中包含异常值时，往往会产生较大的敏感度。这里我们用最常见的最大值查询来说明这个问题。设 $d_{\max}$ 和 $d'_{\max}$ 分别为所有记录中的最大值和第二大值。假设 $d_{\max}$ 是一个离群值，即 $d_{\max}$ 比其他记录大得多。对于最大值查询，移除 $d_{\max}$ 时查询结果变化最大，即 $LS_{\max imum}(X)=d_{\max}$。因此，删除 $d_{\max}$ 将导致最大值查询结果发生很大变化。如果离群值 $d_{\max}$ 不在 X 中，则有 $LS_{\max imum}(X)=d'_{\max}$，这远小于 $d_{\max}$ 在数据集中的情况。因此，当数据集中包含异常值时，我们会得到比实际数据更大的局部敏感度，这直接导致了噪声量非常多，从而使得查询结果的准确性很差。在 3.2 节中提出的 *iDP* 数据集生成

方法不能避免这个问题。CRSC 通过谱聚类解决了这一问题。它将 *iDP* 计算的局部敏感度的基础从原始数据集转移到谱聚类处理后的数据集上，形成了基于聚类的局部敏感度 $LS_{(f\cdot CRSC)}(X)$。通过将数据集划分为 K 簇，在每一簇中都用质心替换原始值，使得替换后的数据集只包含 K 条不同的记录。本质上讲，CRSC 其实是减少了可能导致查询结果发生重大变化的数据量，把数据从从 n 条记录减少到 K 条记录。直接效果就是削弱了异常值对查询结果的影响。我们将这一结论形成了下文的命题：

命题 1：设 f 为一个函数，X 为有 n 条记录 $\{d_1,\Lambda\ d_n\}$ 的数据集，$LS_{(f\cdot CRSC)}(X)$ 为根据算法 2 在数据集 X 上计算得到的基于聚类的局部敏感度。那么，我们有 $LS_{(f\cdot CRSC)}(X)\le LS_f(X)$。

证明：对于函数 f，X 的局部敏感度为原始值的最大变化量，用公式表示即 $\max_{X'}\left\|f(X)-f(X')\right\|_1$。$LS_{(f\cdot CRSC)}(X)$ 实际上是 X 根据 CRSC 通过谱聚类产生的 $\overline{X}$ 的局部敏感度，相当 $\max_{\overline{X}'}\left\|f(\overline{X})-f(\overline{X}')\right\|_1$。它表示移除类中某条记录所导致的质心值的最大变化。将原始值替换为 X 中的质心，将使得 X 的每个聚类中的记录均为质心值。但是对于 X 来说，数据集中记录的多样性远远大于 $\overline{X}$，因为它没有经过 CRSC 的处理，每条记录都保留了原始的值。换句话说，X 中会显著改变查询结果的记录数量要远多于 $\overline{X}$ 中的。因此，删除数据集中的任何一条记录，$\overline{X}$ 会发生的最大变化永远不大于 X 的最大变化，用不等式可以表示为 $\max_{\overline{X}'}\left\|f(\overline{X})-f(\overline{X}')\right\|_1\le\max_{X'}\left\|f(X)-f(X')\right\|_1$，即 $LS_{(f\cdot CRSC)}(X)\le LS_f(X)$。

根据以上讨论，我们将 *iDP* 与 CRSC 结合，形成基于谱聚类的个体化差分隐私（*iDP-SC*），该隐私模型将 *iDP* 计算的局部敏感度从原始数据集转移到谱聚类处理后的数据集。总的来说，它不仅利用了数据管理员在响应查询时可以了解到实际数据集的信息，而且通过在实际数据集上运行 CRSC，将 *iDP* 中的局部敏感度 $LS_f(X)$ 转换为基于聚类的局部敏感度 $LS_{(f\cdot CRSC)}(X)$。

基于上述结论，本文提出一种实现 *iDP-SC* 的机制。拉普拉斯机制是一种常见的差分隐私机制。由公式.(2.8)可以看出，它以 $M(D)=f(D)+\eta$ 的形式满足差分隐私，其中 η 代表服从拉普拉斯分布的随机噪声。若想 $M(D)=f(D)+\eta$ 满足 *iDP-SC*，可以将噪声以基于聚类的局部敏感度 $LS_{(f\cdot CRSC)}(X)$ 为基准来调整。3.4 节将给出详细的证明。在由拉普拉斯分布生成随机噪声的情况下，*iDP-SC* 可以重写为：

命题 2：设 f 为在 R 中取值的查询函数，$LS_{(f\cdot CRSC)}(X)$ 为 X 根据算法 2 生成的基于聚类的局部敏感度。机制 $M(X)=f(X)+(\eta_1,\Lambda\ \eta_n)$ 满足 ε-*iDP*，其中 η_i 为独立同分布的随机噪声，满足拉普拉斯分布 $Lap(0,LS_{(f\cdot CRSC)}(X)/\varepsilon)$

3.3.2 基于谱聚类的改进个体化差分隐私下的 Skyline 查询

为了提高效用，本文基于 Skyline 查询的分销渠道绩效评估中引入了 *iDP-SC*。在图 3.3

中我们描述了整个过程。与 3.3.1 节相比，供应商在得到 *GSO* 后需要运行 *CRSC*（表 3.2），并且计算 *GSO* 与各分销商的 *LSO* 之间的 Jaccard 相似度 j 。这样做的结果是，它将数据集 $D=\{LSO_1,\Lambda\ ,LSO_n\}$ 转化为谱聚类处理后的数据集 $\overline{D}=\{LSO_{c1},\Lambda\ ,LSO_{cn}\}$。

表 3.3 基于谱聚类的改进个体化差分隐私下的 Skyline 查询

算法 3：基于谱聚类的改进个体化差分隐私保护下的 Skyline 查询

输入：$D=\{LSO_1,\Lambda\ ,LSO_n\}$；$\varepsilon$；聚类数 K

输出：满足差分隐私的加噪 Jaccard 相似度 $\{\widetilde{j}_1,\Lambda\ \widetilde{j}_n\}$

1.进行全局 Skyline 查询得到 GSO；

2.对于 $i=1,2,\Lambda\ ,n$:

3. 计算真实的 Jaccard 相似度：$j_i=J(LSO_i,GSO)$；

4. 根据算法 2 计算 LSO_{ci}；

5.结束

6. $\overline{D}=\{LSO_{c1},\Lambda\ ,LSO_{cn}\}$;

7.在 $\overline{D}$ 上执行全局 Skyline 查询得到 GSO_c

8.对于 $i=1,2,\Lambda\ ,n$:

9. 计算聚类后的 jaccard 相似度：$j_{ci}=J(LSO_{ci},GSO_c)$;

10. 计算基于聚类的局部敏感度 $LS_{(J\cdot CRSC)}(D)$；

11. 使得 η 服从拉普拉斯分布 $Lap(0,\frac{LS_{(J\cdot CRSC)}(D)}{\varepsilon})$

12. 计算加噪后的 jaccard 相似度 $\widetilde{j}_i=j_{ci}+\eta$

13.**结束**

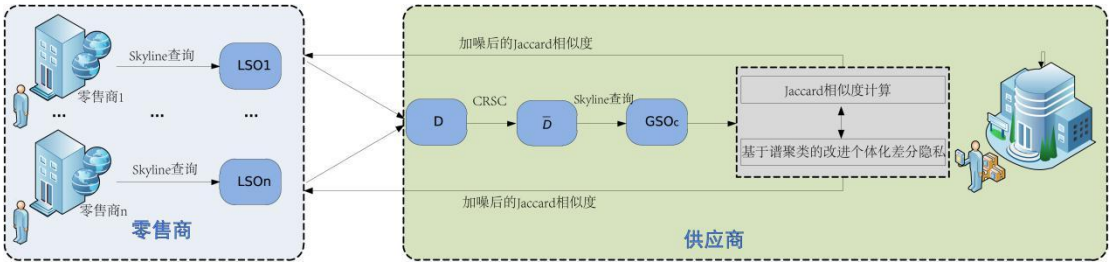

图 3.3 基于谱聚类的改进个体化差分隐私保护下的 Skyline 查询

表 3.3 用伪代码明确了图 3.3 中描述的过程。具体来说，将每个分销商的 LSO 发送给供应商后，通过谱聚类将其划分为 K 类。在每一类中,供应商需要计算质心并用质心代替原始值，这样就可以得到聚类替换后的数据集 $\overline{D}$ 。在这之后，供应商将在 $\overline{D}$ 上执行全局 Skyline 查询来获得 GSO_c 。出于与 3.2 节相同的隐私考虑,供应商在计算得到 GSO_c 和每个分销商的 LSO_c 之间的 Jaccard 相似度 j_c 后，仍然需要执行一些隐私保护操作以确保查询不会泄露隐私。很明显，每个分销商的 j_c 和 j 并不相等，因为 *iDP-SC* 的加入会带来一些信息失真。但更值得注意的是，由于 *CRSC* 的运行，基于聚类的局部敏感度 $LS_{(J\cdot CRSC)}(D)$ 将远小于低于 *iDP-SC* 下计算出来的局部敏感度，这已经在 3.3.1 节得到了证明。最后，根据提案 2，对于一个给定的隐私预算 ε ，为了得到加噪后的 jaccard 相似度，供应商只需要为 Jaccard 相似度 j_c 添加服从拉普拉斯分布 $Lap(0, LS_{(J\cdot CRSC)}(D)/\varepsilon)$ 的随机噪声即可。

3.4 理论证明与分析

本文提出的 *iDP-SC* 旨在在隐私预算 ε 不变的情况下进一步提高查询结果的数据效用。在本节中，我们将证明 *iDP-SC* 是满足 ε-iDP 的，然后分析为什么 *iDP-SC* 下的数据效用可以得到提高。

3.4.1 隐私性证明

为了证明 *iDP-SC* 所提供的隐私保证，我们首先证明它是满足 ε-DP 的，然后再将此结论拓展到是满足 ε-iDP 的。如果我们能证明查询结果在任意一对相邻数据集上的被分辨出来的概率不大于 $\exp(\varepsilon)$ ，那么我们可以说在命题 2 中提出的用于实现 *iDP-SC* 的机制满足 ε-DP 。

证明：设 X 和 X' 为一对相邻数据集， p_X 和 $p_{X'}$ 分别表示 $M(X)$ 和 $M(X')$ 的概率密度函数。在任意点 $z \in R$ ：

$$\begin{aligned}\frac{p_X(z)}{p_{X'}(z)} &= \prod_{i=1}^{n}\left(\frac{\exp\left(\frac{-\varepsilon\left|f(X)_i - z_i\right|}{LS_{(f\cdot CRSC)}(X)}\right)}{\exp\left(\frac{-\varepsilon\left|f(X')_i - z_i\right|}{LS_{(f\cdot CRSC)}(X)}\right)}\right)\\ &= \prod_{i=1}^{n}\exp\left(\frac{\varepsilon\left(\left|f(X')_i - z_i\right| - \left|f(X)_i - z_i\right|\right)}{LS_{(f\cdot CRSC)}(X)}\right)\\ &\le \prod_{i=1}^{n}\exp\left(\frac{\varepsilon\left|f(X)_i - f(X')_i\right|}{LS_{(f\cdot CRSC)}(X)}\right)\\ &= \exp\left(\frac{\varepsilon\left\|f(X) - f(X')\right\|_1}{LS_{(f\cdot CRSC)}(X)}\right)\\ &\le \exp(\varepsilon)\end{aligned} \tag{3.3}$$

所以机制 $M(X)$ 满足 ε-DP.

命题 3：对于任意一个实际数据集 X，任意满足 ε-DP 的机制也满足 ε-iDP。

由上述命题，我们便可以证明机制 $M(X)$ 满足 ε-iDP。

3.4.2 效用分析

一般来说，对于差分隐私机制，只有一个错误来源，即为了符合公式（2.3）而添加的噪声误差。但是对于 *iDP-SC*，我们将查询 f 函数 f 近似为 $f\cdot CRSC$，这直接导致了第二个误差来源：f 近似为 $f\cdot CRSC$ 所造成的误差。也就是说在原始数据集 X 上计算查询 f 与在谱聚类处理后的数据集 $\overline{X}$ 上计算查询 f 所造成的误差是不可忽略的。与直观相反的是，$f\cdot CRSC$ 计算中的这两种误差之和小于 f 的噪声误差。

毫无疑问，f 与 $f\cdot CRSC$ 的近似在一定程度上扭曲了数据。但对于差分隐私保护下的查询结果，由于差分隐私严格的隐私保证，所需要的噪声量通常较大，这使得信息失真的误差在总误差中所占的比例要小于噪声误差所占的比例。如 3.3.1 节所示，聚类后质心的敏感度要低于原始数据集的单个记录，这有效地降低了敏感度。当隐私预算不变的时候，敏感度的降低意味着满足差分隐私所需的噪声将更少。因此，引入 *CRSC* 后，噪声将大大降低。另一方面，也存在一定的信息失真，但总的来说，*CRSC* 降低噪声误差带来的准确性的提升补偿了其自身信息失真误差导致的准确性的降低，因此总体上引入 *CRSC* 后总误差将进一步降低。

与原始记录相比，只注意到 *CRSC* 处理过的质心降低了敏感度是不够的。如前文所述，*CRSC* 减少了可能导致查询结果发生重大变化的数据量，从 n 条记录减少到 K 条记录。随

着聚类数的增加，K 会影响敏感度的降低程度。一方面，K 越大，聚类数据集越接近原始数据集，即 $f \cdot CRSC$ 越接近 f，信息失真误差减小。但另一方面，随着 K 的增加，质心的数量也增加，所以记录减少带来的查询结果发生重大变化的可能性也增强了，这带来的结果就是 CRSC 降低敏感度带来的好处将逐渐减小。然而，正如我们前面所讨论的，噪声误差在差分隐私查询结果的总误差中占很大比例，所以对于 K 的增加，敏感度增大带来的影响将大于信息失真误差减小带来的影响。总体上，K 越大，数据效用越低，这将在下一节内容中得到验证。

4 实验结果及分析

在本节中，我们通过回答以下问题来评估 *iDP-SC* 的性能：

（1）*iDP-SC* 如何保证数据效用？

提出的 *iDP-SC* 旨在进一步提高 *iDP-SC* 的数据效用。在 6.2 节中，我们将利用真实 Skyline 查询结果和加噪 Skyline 查询结果的均方根误差（*RMSE*）和平均绝对误差（*MAE*）评估其性能，同时也会将 *iDP-SC* 与 DP 和 *iDP* 进行比较。

（2）*iDP-SC* 中的主要参数如何影响其性能的？

iDP-SC 有两个主要参数，决定 iDP-SC 聚类数的 K 和控制 iDP-SC 隐私保护级别的 ε。在第 4.2 节，我们会评估两个参数对 iDP-SC 的具体影响。

4.1 实验设置

4.1.1 对比模型

我们将 DP 和 *iDP* 作为 *iDP-SC* 的对比模型。

差分隐私（DP）：我们使用拉普拉斯机制来实现差分隐私，它根据全局敏感度校准噪声，并要求查询结果在任意一对相邻数据集上都是不可区分的。如 2.1.2 节所述，它是一种原始且广泛使用的机制，可以用来生成具有强大隐私保证的数据集。

个体化差分隐私（*iDP*）：与 DP 不同的是，*iDP* 只要求查询结果在实际数据集与其任意一个相邻数据集上不可区分，而不是像标准 DP 那样要求任意一对相邻数据集上不可区分。本文利用拉普拉斯机制实现了 ε-iDP，其中噪声是根据局部敏感度校准的，另外数据集也没有经过聚类处理。

4.1.2 数据集介绍

我们通过在几个数据集上进行 Skyline 查询来评估 iDP-SC，这些数据集包括两个虚拟数据集和一个从 Kaggle 上获得的真实的 Automobile 数据集。为了更好地模拟本文中的分销渠道绩效评价，我们在每个数据集中都设置了 3 个分销商。

用虚拟数据集进行实验，不仅可以方便地对方法进行评价，而且可以准确地模拟本文的分销渠道绩效评价场景。我们生成了两个虚拟数据集，这两个数据集都包含四个属性，每个属性都有不同的取值范围。在虚拟数据集 1 中，每个分销商有 400 条记录，对于虚拟数据集 2，每个分销商有 100 条记录。对于这两个数据集，Skyline 查询建立在以下支配规则之上：如果一条数据的所有属性都不大于另一条数据，并且至少存在一个维度上小于另一条数据，则该数据就会支配另一条数据。

真实数据集 Automobile 来自 Kaggle，有 206 条记录。该数据集包含了一辆汽车锁包含的所有属性，包括车门数量、燃油类型等 9 个属性。在该数据集中，Skyline 查询的支配原则是，如果一辆汽车相比另一辆汽车有相同或更大的马力、尺寸、门的数量、燃料类型、压缩比，而且它更便宜，则称一辆汽车支配另一辆汽车。

本文所有的实验都在 Python3 中实现，且都在 Intel Core i5-3210M 2.50GHz 的 6GB 内存 PC 上进行。每种方法在每次实验中都测试了 1000 次，然后计算相应的评价指标。

4.1.3 实验参数设置

我们考虑了影响 *iDP-SC* 性能的两个参数 K 和 ε：

聚类数量 K：我们在前面的章节从理论上阐述了 K 会影响查询结果的数据效用，在本节我们将通过实验验证这个理论。在后文的实验中，我们主要观察查询结果的数据效用是如何随着 K 改变的。

隐私预算 ε：在简单的拉普拉斯机制下，数据效用与隐私预算之间的关系是已知的，本节的实验将着重观察在不同的隐私预算下，所提出的 *iDP-SC* 与其他两种基准模型的性能差异。我们将 ε 的值设置在 0.1 到 1.0 之间，并且以 0.1 为步长进行递增，这个取值范围是 ε 能保证差分隐私可靠性的合理范围[10]。

在本文的实验中，我们将改变这两个参数，观察第前文提到的效用指标，以研究两个参数对三种隐私模型的影响。

4.1.4 效用指标

原始数据和加噪数据之间的差异显示了效用的变化。因此，为了衡量添加噪声后的数据效用，我们采用均方根误差（*RMSE*）和平均绝对误差（*MAE*）作为对 DP、*iDP* 和本文提出的 *iDP-SC* 性能的效用度量指标。

均方根误差（*RMSE*）：均方根误差（*RMSE*）是通过计算估计量的预测值与其真实值之间的误差来衡量模型性能的。*RMSE* 较低时，意味着加噪的查询结果更接近真实的查询结果。在后文的实验中，*RMSE* 可以计算为：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(j-\tilde{j})^2} \tag{4.1}$$

其中 j 和 $\tilde{j}$ 分别对应精确的 jaccard 相似度和加噪后的 jaccard 相似度。

平均绝对误差（*MAE*）：平均绝对误差（*MAE*）是一个非常容易解释的度量，它不太可能被少量的离群值所影响。*MAE* 值越小，加噪后的 jaccard 相似度越接近真实值，隐私模型的数据可用性越好。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|j - \tilde{j}\right| \tag{4.2}$$

4.2 实验结果及分析

通过与 DP 和 *iDP* 的比较，验证了所提出的 *iDP-SC* 的性能。对于这三个隐私模型中的每个模型、每个数据集、每个分销商和每个实验参数（聚类数 K 和隐私预算 ε），我们分别做了 1000 次试验，并记录了 *MAE* 和 *RMSE*。

4.2.1 隐私预算的影响

首先，我们将 *iDP-SC* 下每个数据集和每个分销商的 K 值固定为 2，因为它可以清楚地显示三种模型之间的差异。在该实验下隐私预算 ε 从 0.1 到 1.0 逐步递增，在每个 ε 下我们都记录了这三个模型的效用指标值。

如图 4.1-4.3 所示，我们观察到 DP 由于全局敏感度最大因而 *MAE* 和 *RMSE* 最高，其次是 *iDP*。在所有的数据集上，*iDP-SC* 都比 DP 和 *iDP* 的 *MAE* 和 *RMSE* 要小。以图 4.1 中的分销商 1 为例，当 $\varepsilon = 0.1$ 时，*iDP-SC* 的 *MAE* 为 0.0309，DP 的为 1.4937，*iDP* 的为 0.2762，相比 DP 和 *iDP* 分别提高了 97.94%和 88.81%。当 $\varepsilon = 1$ 时，*iDP-SC* 的 *MAE* 值为 0.0174，比 *DP* 和 *iDP* 的对应值分别高出 88.28%和 33.88%。这三种模型在 *RMSE* 上的表现与在 *MAE* 上的相似。*iDP-SC* 在数据效用上的显著提高也可以在图 4.2-4.3 中看到。因此，本文提出的 *iDP-SC* 具有更好的数据效用，因为它进一步降低了 *DP* 和 *iDP* 的敏感度。实验结果表明了 *iDP-SC* 的强大性能。

作为差分隐私中的一个关键参数，隐私预算 ε 能够决定模型的隐私保护水平。从图 4.1-4.3 中，我们可以看到，隐私预算 ε 对这三种隐私模型有不同的影响。出于隐私保护的目的，$\varepsilon = 1$ 或更少的值将是一个合理的取值范围。在我们的实验中，我们遵循这个规则。为了进行全面的调查，我们在所有数据集中将隐私预算从 0.1 以 0.1 的步长递增到 1.0，由此分别评估了 *iDP-SC* 和其他两个基准模型在不同隐私保护水平下的性能。非常容易观察到，DP 和 *iDP* 的 *MAE* 和 *RMSE* 随着 ε 的增加而显著降低。但是，*iDP-SC* 的这两个效用指标却不会随着 ε 的增加而发生显著变化。对于图 4.1 中的分销商 1，当 $\varepsilon = 0.1$ 时，DP 和 *iDP* 的 *MAE* 分别为 1.4937 和 0.2762。当 $\varepsilon = 0.5$ 时，两种模型的 *MAE* 分别降至 0.2916 和 0.0513。但是 *iDP-SC* 的指标在这种情况下是非常稳定的：*MAE* 一直保持在 0.017 左右，最大的降低范围只有 0.013。同样的趋势也可以在其他数据集中观察到。实验结果表明，在

DP 和 *iDP* 下，为了获得更高的数据效用，需要牺牲隐私保护水平。而对于 *iDP-SC* 而言，在考虑隐私与效用的权衡时，可以自由选择较小的隐私预算 ε 来提供更严格的隐私保障，而不用担心较小的 ε 会造成很大的效用损失。换句话说，*iDP-SC* 能够在满足严格隐私保护要求的同时保持较高的数据效用。

此外，我们还观察到，在 ε 较小时，DP、*iDP*、*iDP-SC* 三条曲线相隔较远。但是，*iDP-SC* 和 *iDP* 的曲线会随着 ε 的增大而逐渐接近甚至重叠。对于图 4.1 中的分销商 1，当 $\varepsilon = 0.1$ 时，*iDP-SC* 与 *iDP* 的 *MAE* 差值为 0.2453。当 $\varepsilon = 0.4$ 和 $\varepsilon = 1$ 时，差值分别缩小到 0.0372 和 0.0089。*RMSE* 的趋势与 *MAE* 相似，在其他数据集上也可以观察到同样的趋势。对于 *iDP-SC*，有两个误差源，为满足公式（2.3）而增加的噪声误差和引入 *CRSC* 造成的信息失真误差。在本文的实验中，由于 *CRSC* 中的 K 是固定的，因此引入的信息失真误差是确定的。但对于噪声误差来说，敏感度和隐私预算 ε 都会导致其变化。毫无疑问，在隐私预算相同的情况下，敏感度较小时噪声误差更小。但当敏感度的降低保持不变时，较的小隐私预算下的噪声误差降低量大于较大的隐私预算下的噪声误差降低量。换句话说，隐私预算越小，噪声误差降低量越明显。当 ε 较小时，噪声误差的降低量比引入的信息失真误差要大得多。在这种情况下，这三条曲线在同一 ε 下 就会相隔较远，因为通过噪声误差的降低可以补偿 CRSC 所增加的信息失真误差。但另一方面，随着 ε 的增大，噪声误差降低量逐渐减小，慢慢接近于引入的信息失真误差。因此，*iDP-SC* 的低敏感度带来的误差降低量将逐渐被信息失真增加的误差所抵消。因此，*iDP-SC* 与 *iDP* 的曲线逐渐接近。此外，值得注意的是，在某些情况下，*iDP* 比 *iDP-SC* 具有更低的 *MAE* 和 *RMSE*，图 4.1 中的分销商 3 就是这样的情况。这是因为降低敏感度带来的噪声误差降低量要小于高隐私预算下信息失真增加的误差。

除此之外，我们将 *iDP-SC* 的两个效用度量指标与没有经过隐私保护处理的查询结果，即真实的查询结果进行了比较。非常明显，在没有任何隐私保证的情况下响应查询，*MAE* 和 *RMSE* 为 0。从图 4.1-图 4.3 可以看出，当 ε 从 0.1 变化到 1.0 时，*iDP-SC* 的 *MAE* 和 *RMSE* 接近于 0。这是因为 *iDP-SC* 采用了 *CRSC*，大大降低了敏感度，从而降低了引入噪声误差。该实验结果证实了 *iDP-SC* 在严格保证隐私性的同时能够保持非常高的数据效用。

综上讨论，上述实验评估充分显示了本文提出的 iDP-SC 在以下几个方面的强大性能：（1）与 DP 和 iDP 相比，iDP-SC 可以保持更高的数据效用。（2）iDP-SC 的数据效用不会随着隐私预算 ε 的增加而发生明显的变化，因此，为了更严格的隐私保证我们可以自由地选择较小的 ε，而不用担心较小的 ε 会带来明显的数据效用损失。（3）在 iDP-SC 下，数据效用的损失量很小，在 ε 从 0.1 到 1.0 之间的范围内，加噪的查询结果都接近于真实的查询结果。

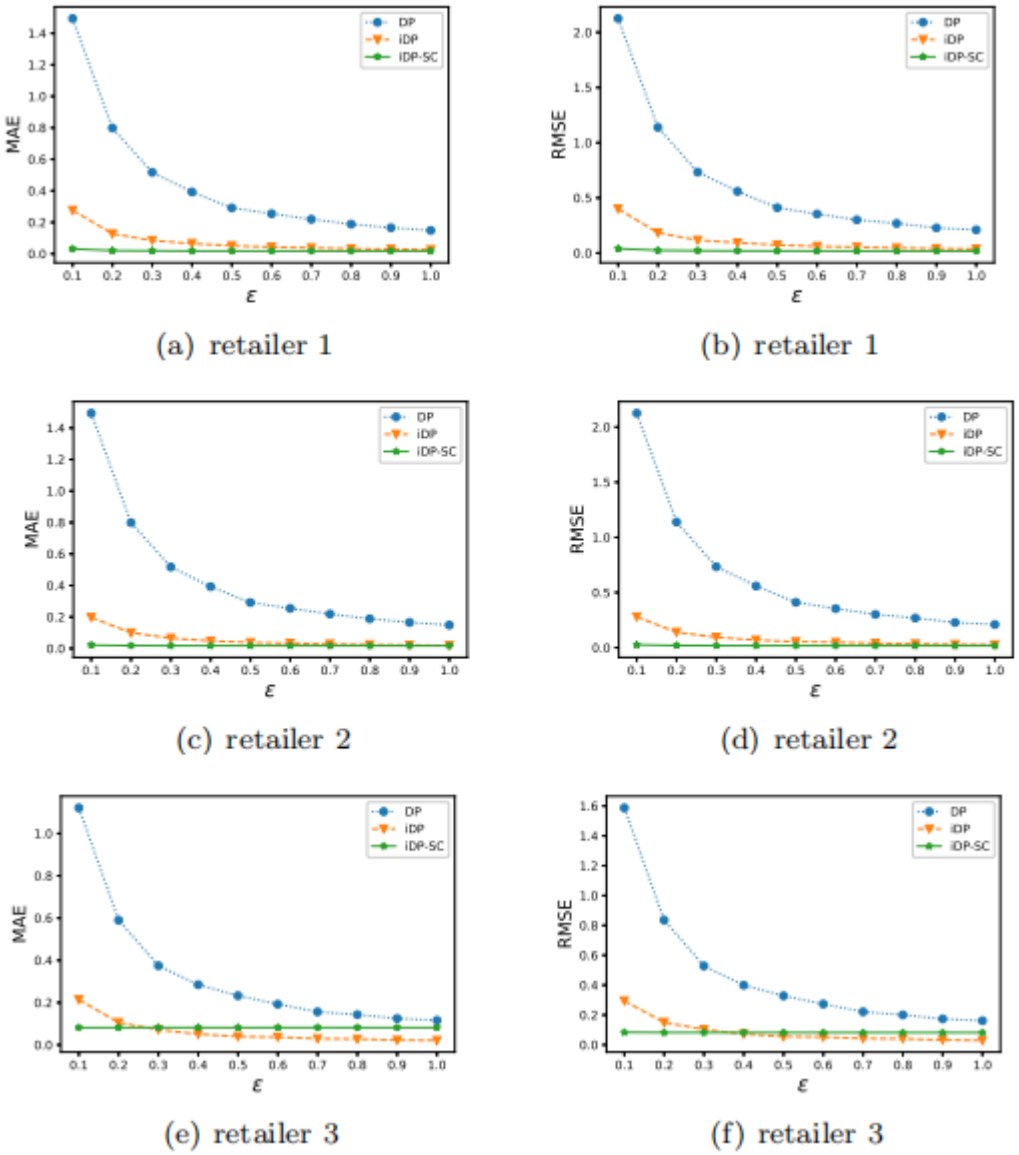

图 4.1　Automobile 数据集上各隐私模型在不同 ε 下的数据可用性对比图

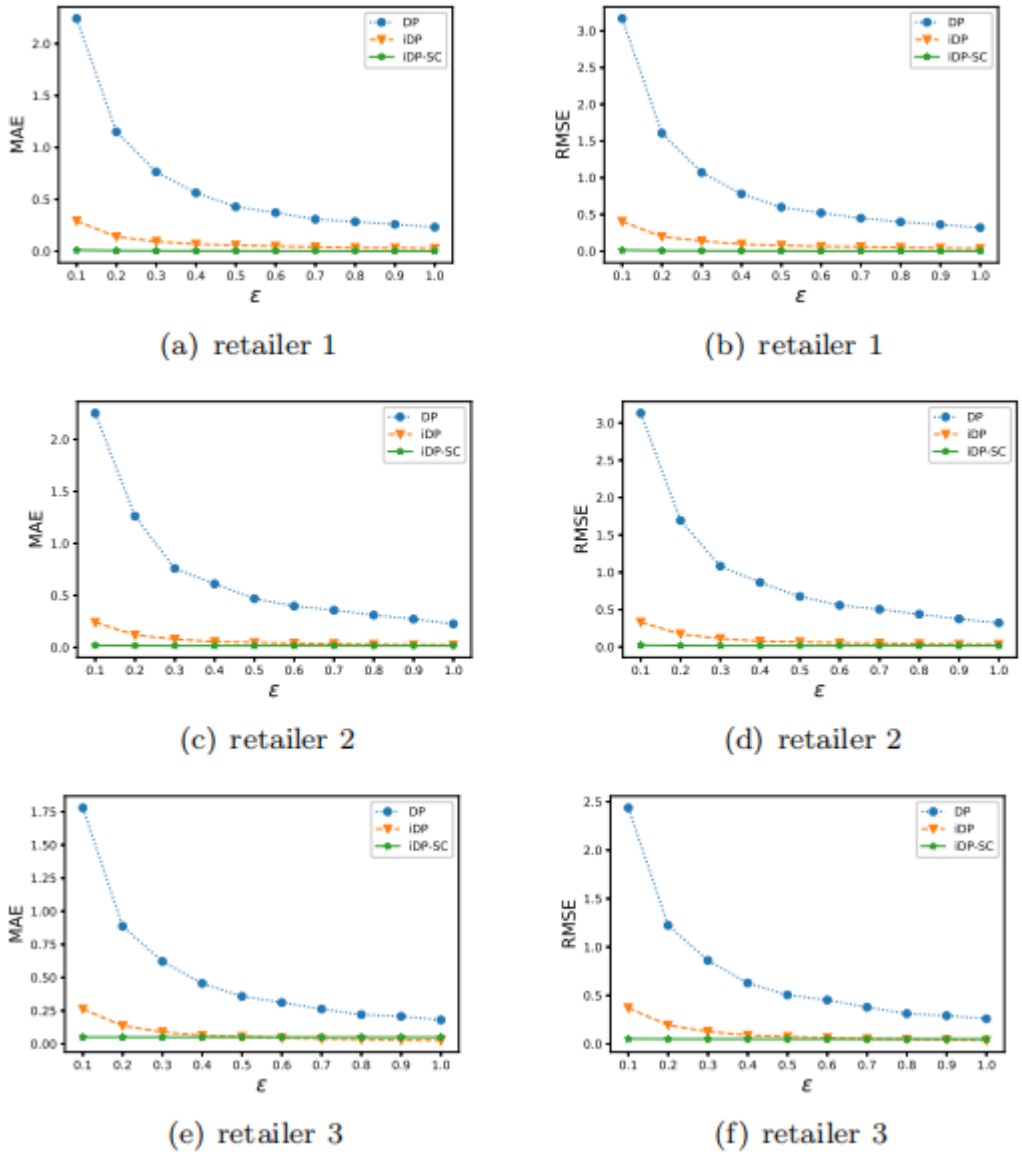

图 4.2　虚拟数据集 1 上各隐私模型在不同 ε 下的数据可用性对比图

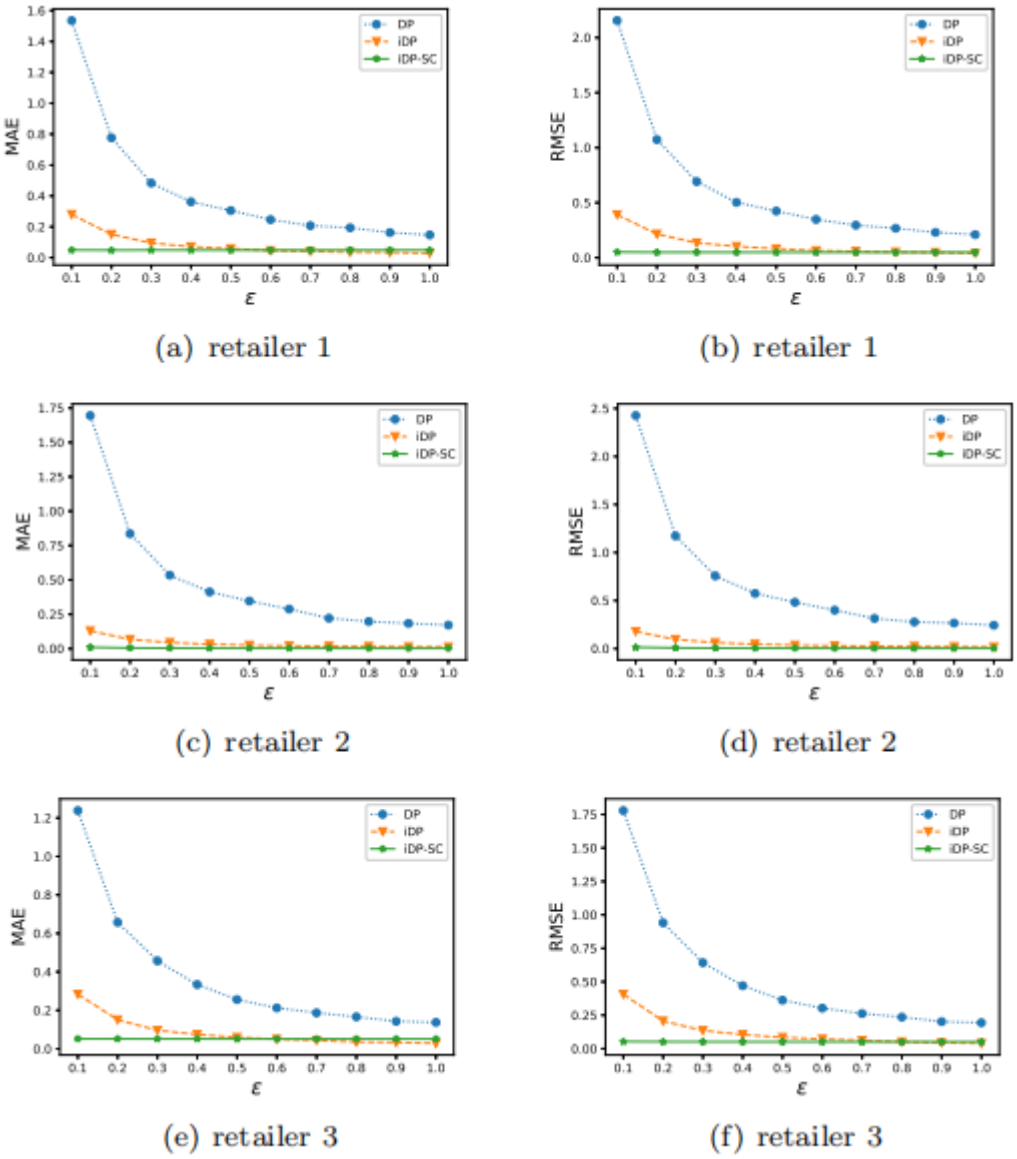

图 4.3　虚拟数据集 2 上各隐私模型在不同 ε 下的数据可用性对比图

4.2.2 聚类数目 K 的影响

在 *iDP-SC* 中，K 是控制聚类数量的参数。根据 3.4 节，聚类数量直接影响查询结果的数据效用。本届设计实验研究了 K 的影响，在隐私预算 ε 控制在 0.1 的前提下，观察参数 K 从 2 变化到 6 会对数据效用产生什么样的影响。

从图 4.4-图 4.6 可以看出，随着 K 值的增加，*iDP-SC* 的性能显著优于其他两个隐私模型，这与图 4.1-图 4.3 的结果一致。需要解释的是，由于没有引入聚类，DP 和 *iDP* 都与参数 K 无关，因此误差曲线随着 K 的变化适中保持水平。

图 4.4-图 4.6 还展示了 K 对 *iDP-SC* 性能的具体影响。通过观察在所有数据集上的实验结果，我们发现当 K 小于 4 时，*iDP-SC* 下的 *MAE* 和 *RMSE* 显著低于 *iDP*。同时，随着 K 的增加，*iDP-SC* 下的 *MAE* 和 *RMSE* 变化都不显著，但当 K 大于 4 时，*MAE* 和 *RMSE* 都

随着 K 的增加而显著增加，直到逐渐接近 *iDP* 的 *MAE* 和 *RMSE*。如图 4.4 所示，对于分销商 2，当 K 小于 4 时，*MAE* 保持在 0.02 左右。之后，随着 K 的增长，*MAE* 不断增大，当 $K=6$ 时，几乎等于 *iDP* 的 *MAE*。与图 4.4 中的分销商 3 相似，在 $K<4$ 时 *iDP-SC* 的 *MAE* 保持稳定，并且比 *iDP-SC* 的 *MAE* 小得多。当 $K\geq 4$ 时，*MAE* 逐渐增大直至接近 *iDP* 的 *MAE*。*RMSE* 的情况跟 *MAE* 的情况是类似的，同时，图 4.5 和图 4.6 展现的效用随 K 变化的趋势也与此类似。

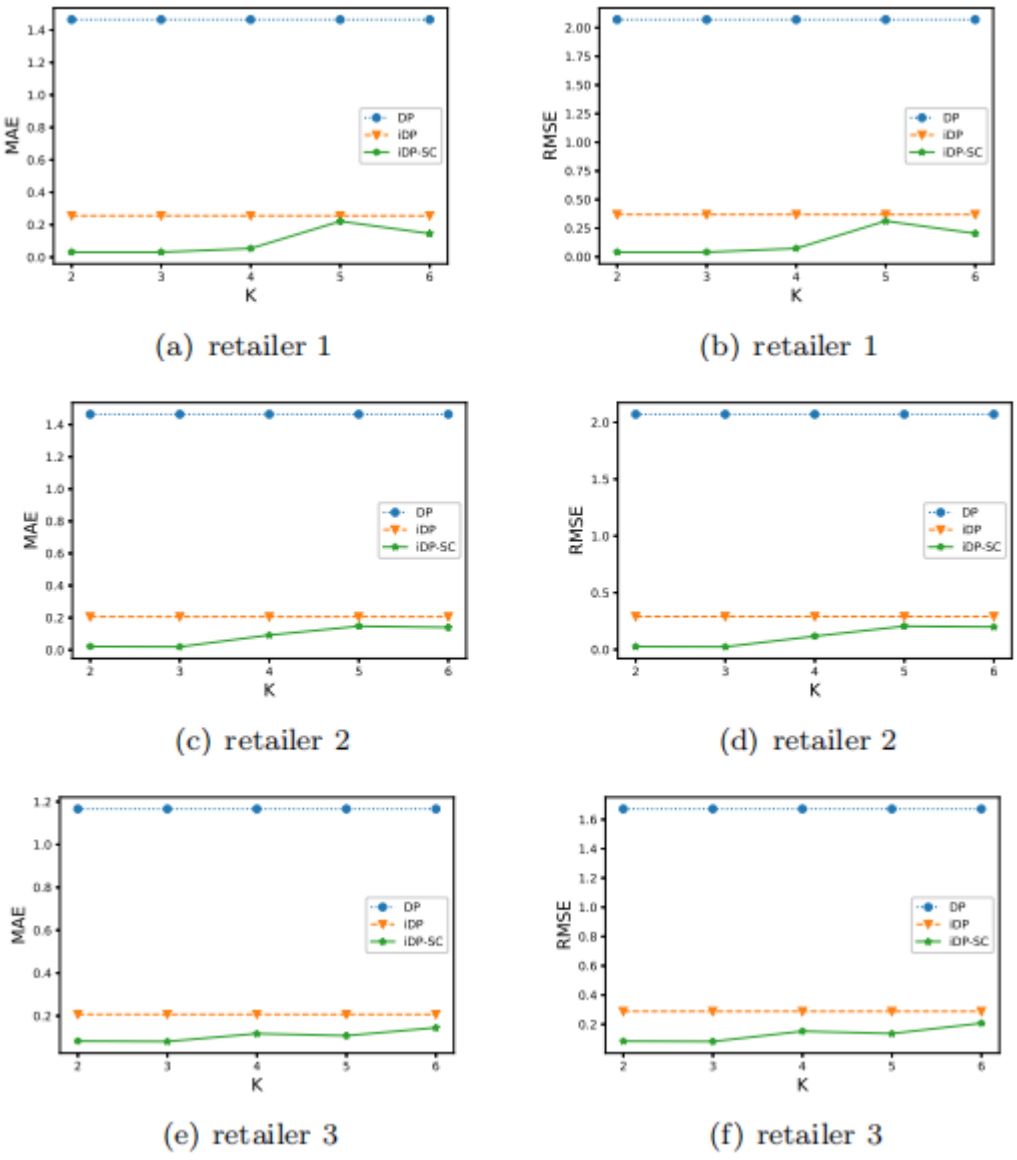

图 4.4　Automobile 数据集上各隐私模型在不同 K 下的数据可用性对比图

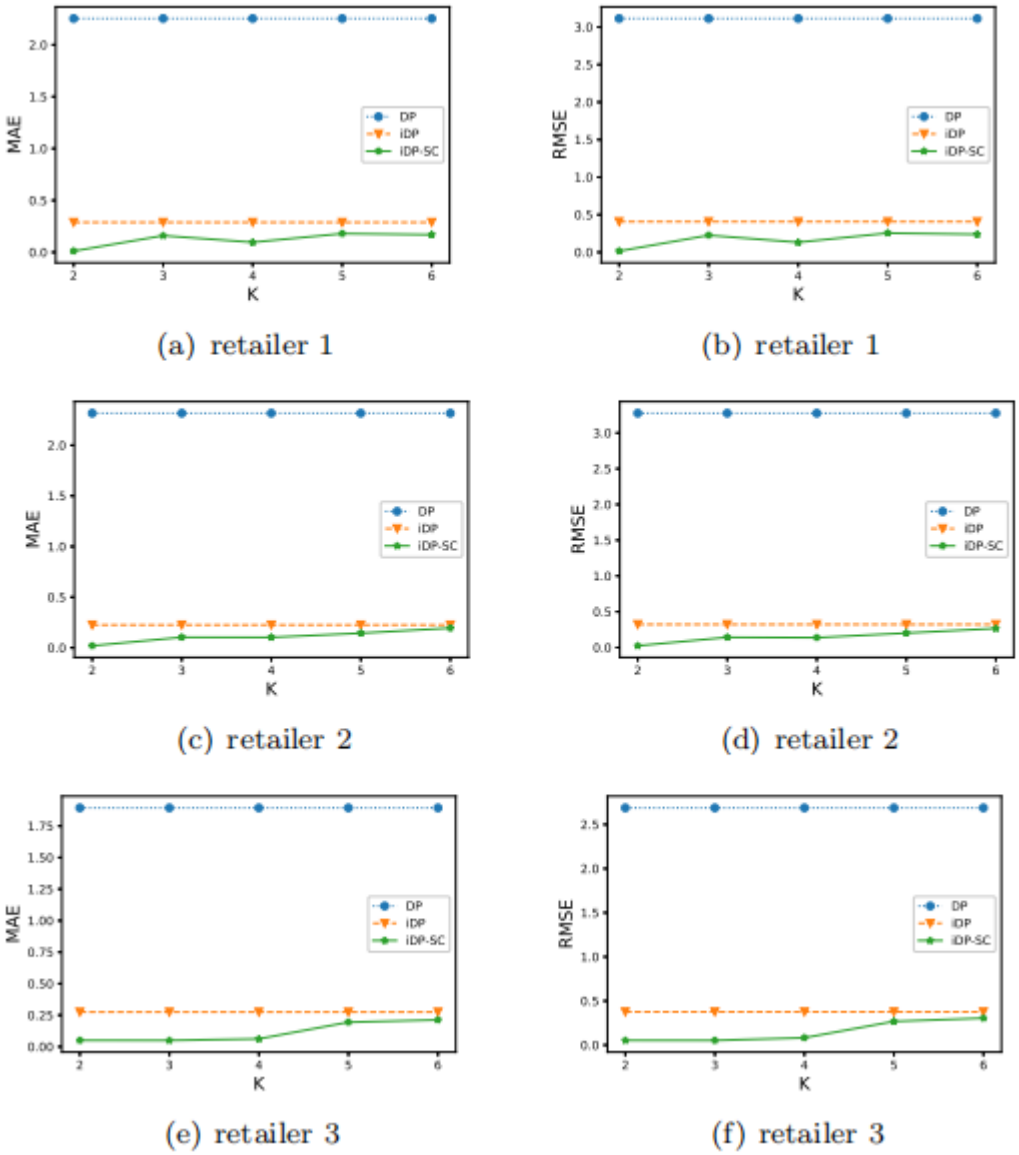

图 4. 5 虚拟数据集 1 上各隐私模型在不同 K 下的数据可用性对比图

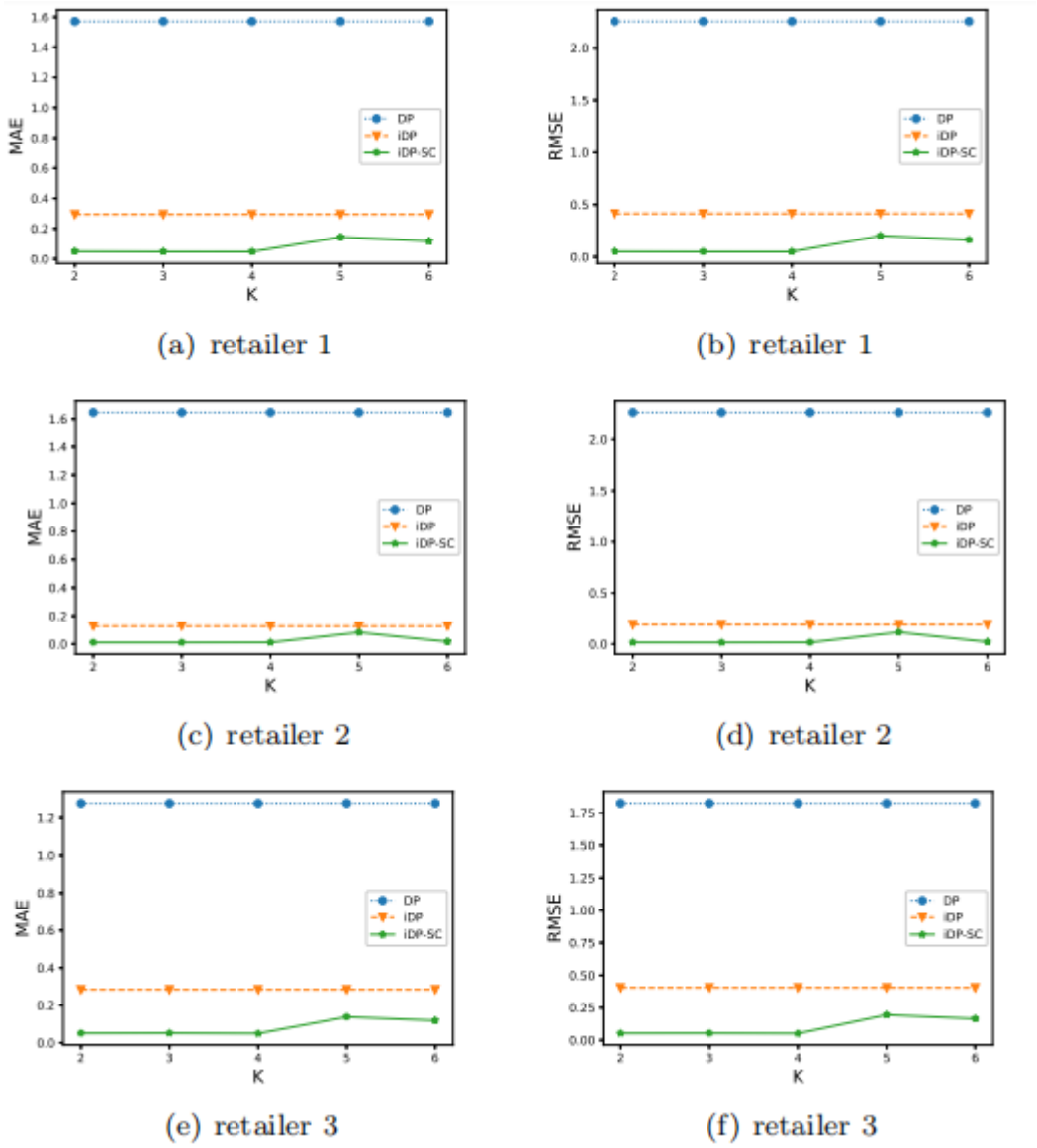

图 4.6　虚拟数据集 2 上各隐私模型在不同 K 下的数据可用性对比图

这样的结果很容易解释。K 越大表示 *CRSC* 聚类替换质心得到的数据集 $\overline{D}$ 越接近原始数据集 D，所以随着 K 的增加，由聚类引入的信息失真误差逐渐减小。另一方面，*iDP-SC* 的敏感度随着 K 的增加而增加，直到逐渐接近 *iDP-SC* 的敏感度。也就是说，K 越大，聚类带来的敏感度的降低量越不显著，所以噪声误差的降低量随着 K 的增加而增加。噪声误差和信息失真误差构成了总误差，因此总误差是波动的，但一般都随着 K 的增加而增加。这部分的实验结果也可以证实前文的理论分析：K 越大，*iDP-SC* 的数据效用越低。

上述实验结果表明了聚类数 K 与查询结果的数据效用之间的关系。这些关系为 *iDP-SC* 中控制聚类数和提高查询结果的数据效用提供了启发。

4.2.3 总结和建议

上述实验显示参数 K 和隐私预算 ε 对本文提出的 *iDP-SC* 性能的影响。图 4.1-4.3 展现了隐私预算 ε 对评估指标 *MAE* 和 *RMSE* 的影响，我们观察到，对于所有数据集，与 *DP* 和

iDP-SC 相比，*iDP-SC* 具有更高的数据效用，能够适应更强的隐私要求（即更小的隐私预算 ε ）而不会导致误差过大。事实上，正如我们在前文所分析的，与 *DP* 和 *iDP-SC* 相比，*iDP-SC* 的敏感度降低了，由此直接导致了噪声误差的降低。虽然先前的 CRSC 也会产生信息失真误差，但与差分隐私下噪声误差相比，它实际上要小得多，这也正显示了我们提出的 *iDP-SC* 在实践上的巨大优势。基于上述观察，并结合之前的理论结果，可以得知 *iDP-SC* 利用 CRSC 达到了较低的敏感度，这使得它能够提供与 DP 相同的隐私保障，同时达到了数据效用的目的。在研究聚类数 K 影响的实验中，当 K 小于 4 时，*iDP-SC* 的性能明显优于 *iDP-SC*。我们还发现，对于所有数据集，K 越大，*iDP-SC* 的数据效用越低。随着 K 的增加，聚类引起的信息失真误差逐渐减小，但 *iDP-SC* 的敏感度增加，因此噪声误差增加。噪声误差的增加要大于信息失真误差的减少，因此总误差通常随 K 的增加而增加。实验结果为使用 *iDP-SC* 时如何控制聚类数 K 和如何进一步提高查询结果的数据效用提供了启发。根据以上的实验结果，我们建议在使用本文提出的 *iDP-SC* 时使用的参数范围为 $0.1 \le \varepsilon \le 0.5$，$2 \le K \le 4$。

5 结论

Skyline 查询是一种功能强大的数据分析工具，广泛应用于多准则决策。在分销渠道绩效评价的背景下，如何解决 Skyline 查询中潜在的隐私泄露问题是现实中值得我们的问题。解决这一问题的现有研究大多采用加密技术，所提供的的隐私保护力度往往是有限的。近年来，差分隐私技术因其稳健性和可靠性而广泛受到学术界的关注，但是它也存在一定争议，因为差分隐私往往会牺牲数据效用。为了解决这一问题，我们提出了基于谱聚类的个体化差分隐私(iDP-SC)。该算法利用数据管理员在响应查询时可以掌握当前实际数据集这一事实，将差分隐私要求的查询结果的不可区分性从任意一对相邻数据集转换为实际数据集及其任意一个相邻数据集。因此，通过 iDP 噪声只需要调整到局部敏感度，而不需要差分隐私所要求的全局敏感度，这往往远高于局部敏感度。然后，我们利用谱聚类进一步降低 iDP 下的局部敏感度。通过基于谱聚类的质心替换，用基于谱聚类的局部敏感度代替局部敏感度，从而进一步降低了噪声，提高数据效用。本文的主要贡献可以总结如下：

（1）本文提出了一种基于差分隐私的隐私保护 Skyline 查询方法。与其他算法相比，我们的算法可以避免传统研究可能泄露关键信息的弊端，并且可以对隐私保护程度进行定量分析。

（2）所提出的 iDP-SC 为解决差分隐私数据效用差的问题提供了一种新的、更有效的方法。我们利用了 iDP 的保持数据效用的核心思想，并进一步结合谱聚类提出了 iDP-SC。因此，iDP-SC 在很大程度上保持了数据效用，而又不会降低差分隐私提供的隐私保护强度，而且即使隐私预算很小，加噪之后的查询结果也可以接近真实的查询结果。

（3）我们从隐私性和效用两个方面对所提出的 iDP-SC 进行了理论分析。结果表明，iDP-SC 能够同时满足 DP 和 iDP。在效用分析方面，我们分析了 iDP-SC 中的两种误差源，并进一步论证了该算法可以提高数据效用的原因。

（4）本文通过实验验证了所提出的 iDP-SC 比 iDP 和 DP 具有更好的性能。同时，我们也解释了所提算法的关键参数与查询结果的数据效用之间的关系。在此基础上，我们给出了算法的理想参数。

但是本文提出的 iDP-SC 方法并非没有局限性。未来在这个研究领域还有很多的问题值得研究：首先，为了更可靠的隐私保护，我们可以考虑馆长不被信任的情况。在这种情况下，需要考虑不同的设置，如局部差分隐私(local differential privacy, LDP)，在用户端添加噪声后，再将结果提交给管理员。其次，考虑到用户对隐私的可接受程度可能有不同的期望，个性化差异隐私(PDP)将是 iDP-SC 的又一延伸。此外，iDP-SC 算法假设了不同分销商之间的独立性，分销商之间存在耦合关系时，如何利用分销商之间的关系进一步降低敏感度也是未来有趣的研究方向。

参考文献

[1] 菲利普·科特勒，凯文·莱恩. 营销管理. 第 12 版. 上海：上海人民出版社，2007:351-368.

[2] 张闯.营销渠道管理.北京：清华大学出版社，2014.

[3] Bert Rosenbloom, Rolph Anderson. Channel management and sales management: Some key interfaces.Journal of the Academy of Marketing Science .2006 (3):223-229.

[4] Rauno Rusko. Conflicts of supply chains in multi-channel marketing: a casefrom northern Finland[J].Technology Analysis & Strategic Management .2016 (4):12-15,18.

[5] 伯特·罗森布罗姆.营销渠道：管理的视野.第 8 版.宋华等，译.北京：中国人民大学出版社，2014.

[6] Boldyreva A, Chenette N, Lee Y, 等 Order-Preserving Symmetric Encryption[C]. international cryptology conference, 2009: 224-241.

[7] Borzsony S, Kossmann D, Stocker K, 等 The Skyline operator. international conference on data engineering, 2001: 421-430.

[8] Bothe S, Cuzzocrea A, Karras P, 等 Skyline Query Processing over Encrypted Data: An Attribute-Order-Preserving-Free Approach. conference on information and knowledge management, 2014: 37-43.

[9] Chan H, Perrig A, Song D, 等 Random key predistribution schemes for sensor networks. ieee symposium on security and privacy, 2003: 197-213.

[10] Chen W, Liu M, Zhang R, 等 Secure outsourced Skyline query processing via untrusted cloud service providers. ieee international conference computer and communications, 2016: 1-9.

[11] Choi W, Liu L, Yu B, 等 Multi-criteria decision making with Skyline computation. information reuse and integration, 2012: 316-323.

[12] Curtmola R, Garay J A, Kamara S, 等 Searchable symmetric encryption: improved definitions and efficient constructions. computer and communications security, 2006: 79-88.

[13] Dwork C. Differential privacy. international colloquium on automata languages and programming, 2006: 1-12.

[14] Dwork C. A firm foundation for private data analysis. Communications of The ACM, 2011, 54(1): 86-95.

[15] Dwork C, Rothblum G N. Concentrated Differential Privacy. arXiv: Data Structures and Algorithms, 2016.

[16] Dwork C, Kenthapadi K, Mcsherry F, 等 Our data, ourselves : Privacy via distributed noise generation[J]. Lecture Notes in Computer Science, 2006: 486-503.

[17] Dwork C, Mcsherry F, Nissim K, 等 Calibrating noise to sensitivity in private data analysis. theory of cryptography conference, 2006: 265-284.

[18] Han X, Li J, Yang D, 等 Efficient Skyline Computation on Big Data. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(11): 2521-2535.

[19] Hose K, Vlachou A. A survey of Skyline processing in highly distributed environments. very large data bases, 2012, 21(3): 359-384.

[20] Hua J, Zhu H, Wang F, 等 CINEMA: Efficient and Privacy-Preserving Online Medical Primary Diagnosis With Skyline Query. IEEE Internet of Things Journal, 2019, 6(2): 1450-1461.

[21] Jin W, Han J, Ester M, 等 Mining thick Skylines over large databases. european conference on principles of data mining and knowledge discovery, 2004: 255-266.

[22] Jorgensen Z, Yu T, Cormode G, 等 Conservative or liberal? Personalized differential privacy.

international conference on data engineering, 2015: 1023-1034.

[23] Kossmann D, Ramsak F, Rost S, 等 Shooting stars in the sky: an online algorithm for Skyline queries. very large data bases, 2002: 275-286.

[24] Li H, Xiong L, Ji Z, 等 Partitioning-Based Mechanisms Under Personalized Differential Privacy. pacific-asia conference on knowledge discovery and data mining, 2017: 615-627.

[25] Liu J, Yang J, Xiong L, 等 Secure Skyline Queries on Cloud Platform. international conference on data engineering, 2017: 633-644.

[26] Liu X, Lu R, Ma J, 等 Efficient and privacy-preserving Skyline computation framework across domains. Future Generation Computer Systems, 2016: 161-174.

[27] Liu X, Choo K R, Deng R H, 等 PUSC: Privacy-Preserving User-Centric Skyline Computation Over Multiple Encrypted Domains. trust security and privacy in computing and communications, 2018: 958-963.

[28] Liu X, Yang D, Ye M, 等 U-Skyline: A New Skyline Query for Uncertain Databases. IEEE Transactions on Knowledge and Data Engineering, 2013, 25(4): 945-960.

[29] Machanavajjhala A, Kifer D, Abowd J M, 等 Privacy: Theory meets Practice on the Map. international conference on data engineering, 2008: 277-286.

[30] Nissim K, Raskhodnikova S, Smith A, 等 Smooth sensitivity and sampling in private data analysis. symposium on the theory of computing, 2007: 75-84.

[31] Qaosar M, Alam K M, Zaman A, 等 A Framework for Privacy-Preserving Multi-Party Skyline Query Based on Homomorphic Encryption. IEEE Access, 2019: 167481-167496.

[32] Qaosar M, Zaman A, Siddique M A, 等 Privacy-Preserving Secure Computation of Skyline Query in Distributed Multi-Party Databases. Information-an International Interdisciplinary Journal, 2019, 10(3).

[33] Sanchez D, Domingoferrer J, Martinez S, 等 Utility-preserving differentially private data releases via individual ranking microaggregation[J]. Information Fusion, 2016, 30(30): 1-14.

[34] Soriacomas J, Domingoferrer J, Sanchez D, 等 Improving the Utility of Differentially Private Data Releases via k-Anonymity. trust security and privacy in computing and communications, 2013: 372-379.

[35] Soriacomas J, Domingoferrer J, Sanchez D, 等 Enhancing data utility in differential privacy via microaggregation-based $$k$$k-anonymity. very large data bases, 2014, 23(5): 771-794.

[36] Soriacomas J, Domingoferrer J, Sanchez D, 等 Individual Differential Privacy: A Utility-Preserving Formulation of Differential Privacy Guarantees. IEEE Transactions on Information Forensics and Security, 2017, 12(6): 1418-1429.

[37] Von Luxburg U. A tutorial on spectral clustering. Statistics and Computing, 2007, 17(4): 395-416.

[38] Yang Z, Zhong S, Wright R N, 等 Privacy-preserving queries on encrypted data[C]. european symposium on research in computer security, 2006: 479-495.

[39] Zaman A, Siddique M A, Annisa, 等 Secure Computation of Skyline Query in MapReduce. advanced data mining and applications, 2016: 345-360.

[40] Zenginler T. Secure Skyline querying. Master's thesis, Boğ aziçi University, 2007.

[41] Stern L W , El-Ansary A I , Coughlan A T . 市场营销渠道. 清华大学出版社, 2001.

[42] 伯特· 罗森布罗姆, 罗森布罗姆, 李乃和等. 营销渠道管理. 机械工业出版社, 2003.

[43] Johnson J L. Strategic integration in industrial distribution channels: managing the interfirm

relationship as a strategic asset. Journal of the Academy of marketing Science, 1999, 27(1): 4-18.

[44] Mehta R, Dubinsky A J, Anderson R E. Leadership style, motivation and performance in international marketing channels. European journal of marketing, 2003.

[45] Eichel E, Bender H E. Performance appraisal: A study of current techniques[M]. American Management Assoc. Research and Information Service, 2008.

[46] Weber J A, Dholakia U. Planning market share growth in mature business markets[J]. Industrial Marketing Management, 2013, 27(5): 401-428.

[47] Boldyreva A, Chenette N, O'Neill A. Order-preserving encryption revisited: Improved security analysis and alternative solutions. Annual Cryptology Conference. Springer, Berlin, Heidelberg, 2011: 578-595.

[48]Zhou SG, Li F, Tao YF, Xiao XK.Privacy preservation in database applications: A survey.Chinese Journal of Computes,2009,32(5):847－858 (in Chinese with English abstract).[doi: 10.3724/SP.J.1016.2009.00847]

[49] Liu YH, Zhang TY, Jin XL, Cheng XQ.Personal privacy protection in the era of big data.Journal of Computer Research and Development, 2015,52(1):229－247 (in Chinese with English abstract).[doi: 10.7544/issn1000-1239.2015.20131135]

[50] Mehmood A, Natgunanathan I, Xiang Y, Hua G, Guo S.Protection of big data privacy.IEEE Access on Theoretical Foundations for Big Data Applications, 2016,4:1821－1834.[doi: 10.1109/ACCESS.2016.2558446]

[51] Zhang X, Liu C, Nepal S, Yang C, Gou WC.A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud. Journal of Computer & SystemScience,2014,80(5):100－1020.[doi: 10.1016/j.jcss.2014.02.007]

[52] Mohammadian E, Noferesti M, Jalili R.FAST: Fast anonymization of big data streams.In: Proc.of the ACM Conf.on Big Data Science and Computing.Beijing: ACM Press, 2014.[doi: 10.1145/2640087.2644187]

[53] 杨旭东，高岭，王海，等 一种面向直方图发布的均衡差分隐私保护方法. 计算机学报, 2020(8).

[54] 贾春福，王雅飞，陈阳，孙梦洁，葛凤仪.机器学习算法在同态加密数据集上的应用.清华大学学报(自然科学版),2020,第 60 卷(6): 456-463

[55] MCSHEItRY F, TALWAR K. Mechanism design via differential privacy//48th IEEE Symposium on Foundations of Computer Science. Providence, USA, 2007.

[56] DWORK C, ROTH A. The algorithmic foundations of differential privacy. Foundations &Trends in Theoretical Computer Science, 2014.

[57] 佚名. 中华人民共和国网络安全法. 中华人民共和国全国人民代表大会常务委员会公报, 2016(6).

[58] Manescu D , Tribune J . General Data Protection Regulation. Juridical Tribune (Tribuna Juridica).

Utility-Preserving Differentially Private Skyline Query for for distribution channel performance evaluation

Qiujun Lan, Jiaqi Ma

(School of Business Administration, Hunan University, Changsha, 410000, China)

Abstract： Skyline query has been widely applied in multi-criterion decision making, whose potential privacy risk is practically concerned on distribution channel performance evaluation. Differential Privacy (DP) is a rigorous privacy-preserving method for its robustness and reliability. Considering DP will deteriorate the data utility, this paper proposes the Individual Differential Privacy via Spectral Clustering (iDP-SC) to address the privacy leakage in skyline query. It shifts the calculation of the local sensitivity from the original dataset to the one processed by spectral clustering, and through this, the sensitivity is reduced as well as the calibrated noise. As a result, it maintain higher utility without sacrificing the privacy preservation provided by differential privacy. Furthermore, compared with existing work for privacy-preserving skyline query, the proposed iDP-SC avoids the disclosure of key information, while providing the quantitative analysis of privacy protection level. The performance of the proposed iDP-SC was examined through comparison with the DP and Individual Differential Privacy (iDP) on both real and synthetic datasets. The experimental results demonstrates the capability and effectiveness of the proposed approach.

Keywords: Privacy Preserving, Differential Privacy, Skyline Query, Spectral Clustering