arXiv:2305.03609v2 [stat.ML] 3 Nov 2023

差分隐私拓扑数据分析

Taegyu Kang* kang426@purdue.edu
D统计系 $f\,S$
普渡大学统计系，美国印第安
纳州西拉斐特，邮编 *47907*

Sehwan Kim* kim3009@purdue.edu
统计学系 普渡大学 西拉斐特，
印第安纳州 *47907*，美国

Jinwon Sohn* sohn24@purdue.edu
统计学系 普渡大学 西拉斐特，
印第安纳州 *47907*，美国

Jordan Awan† jawan@purdue.edu
普渡大学统计学系，美国印第
安纳州西拉斐特 *47907*

摘要

本文首次尝试对拓扑数据分析（TDA）进行差分隐私（DP）处理，以生成近乎最优的私有持久图。我们根据瓶颈距离分析持久图的敏感度，并证明，常用的 Čech复形的敏感度不会随着样本量 n 的增加而降低。这使得 Čech复形的持久图难以进行隐私化。作为替代方案，我们证明，由 L^1-距离测度（DTM）得到的持久图具有敏感度 $O(1/n)$。基于敏感性分析，我们提出采用指数机制，其效用函数根据 L^1-DTM 持久图的瓶颈距离定义。我们还推导出该隐私机制准确度的上界和下界；所得界表明，该机制的隐私误差近乎最优。我们通过模拟以及一个追踪人体运动的真实数据集，展示了隐私化持久图的性能。

关键词： Čech复形、到测度的距离、指数机制、持久图、持久同调

1 引言

技术的最新进展使得我们能够获取结构极其复杂的数据，以至于传统的数据分析方法无法妥善处理这些数据。为了分析如此复杂的数据，拓扑数据分析已成为数据科学中不可或缺的工具（Niyogi 等人，2011年；Khasawneh 和 Munch，2016年；Wasserman，2018年；Dindin

*：Kang、Kim 和 Sohn 为共同第一作者，对本文贡献相同。 †：通讯作者。

© 。许可协议：知识共享署名 4.0，详见
https://creativecommons.org/licenses/by/4.0/。

等，2020年；Rieck 等，2020年）。从本质上说，拓扑是能够讨论“邻近”概念的最基本数学结构；其普适性使其成为讨论极其复杂数据的恰当框架，因为这类数据通常并不具备向量空间、流形等更丰富的结构。拓扑数据分析是数据分析的一个新兴分支，旨在捕捉数据的拓扑结构，并且在过去几十年间得到了深入研究。全面综述请参见 Carlsson（2009年）。尤其是其标志性方法——持久同调，已得到广泛的理论研究，并应用于医学（Nicolau 等，2011年）、生物学（McGuirl 等，2020年）、神经科学（Xu 等，2021年；Caputi 等，2021年）、天文学（Xu 等，2019年）以及机器学习（Hensel 等，2021年；Betthauser 等，2022年）等诸多不同领域。

与此同时，随着规模更大、种类更多的数据变得触手可及，如何保护数据中个人的隐私信息也日益受到关注。出于这一考虑，人们对具有形式化保证的隐私保护程序的需求不断增长。这种范式推动了数据科学领域对完善的隐私保护框架的日益关注。差分隐私（Dwork 等人，2006年）是最先进的框架，它以形式化方式量化隐私及其保护。差分隐私要求隐私保护算法对任意两个仅相差一个数据点的数据集产生相似的结果。ϵ-差分隐私的确切定义将在下一节介绍；我们推荐阅读Dwork 和 Roth(2014年)，以全面了解差分隐私。近年来，差分隐私已成为数据科学领域的核心研究课题之一，并被用于解决深度学习（Shokri 和 Shmatikov，2015年；Abadi 等人，2016年）、函数型数据分析（Hall 等人，2013；Mirshani 等人，2019年）、社交网络（Karwa 和 Slavković，2016年；Karwa 等人，2017）等诸多问题。

尽管如上所述，差分隐私框架已广泛应用于数据科学中的众多方法，但其在拓扑数据分析中的应用尚未得到探讨。据我们所知，同时涉及差分隐私和拓扑数据分析的唯一一项工作是Hehir 等人 (2022年)，该研究仅将持久图用作传达随机响应算法效用的一种方法，并未尝试生成拓扑数据分析对象的隐私版本。我们认为，将差分隐私框架引入拓扑数据分析将成为一个新兴研究方向，因为拓扑数据分析方法已成功应用于许多包含个人敏感信息的数据领域。例如，Shnier 等人 (2019年) 使用持久图，区分自闭症谱系障碍患者与对照组的基因表达。此外，如上所述，拓扑数据分析方法还应用于医学和神经科学领域的其他若干问题，例如脑连接（Caputi 等人，2021年）、乳腺癌（Nicolau 等人，2011年）以及神经系统疾病（Lee 等人，2011年）。最后，拓扑数据分析近年来还与其他流行的机器学习方法相结合，例如卷积神经网络（Love 等人，2023年）、自编码器（Hofer 等人，2019年）等。因此，将差分隐私引入拓扑数据分析，可能会对数据科学产生深远影响。

我们的贡献：本文探讨如何将差分隐私（DP）的概念引入拓扑数据分析（TDA）框架。我们的关键观察是，要利用现有的隐私机制，需要采用一种抗离群值鲁棒的TDA方法。这一观察与差分隐私中由来已久的直觉原则相符：任何单个个体的具体数据都不应对分析结果产生显著影响，才能实现隐私保护；例如，参见

Dwork 和 Lei (2009年), Avella-Medina (2021年). 为阐明这一原则如何适用于TDA，我们考察持久图的瓶颈距离的敏感度。持久图是持久同调最广泛使用的表示形式，本文考察由两种不同构造得到的持久图：由Čech复形得到的持久图，我们发现其不具备抗离群值鲁棒性；以及由到测度的距离（DTM）得到的持久图，其具备抗离群值鲁棒性。我们的考察说明了为何Čech复形的持久图难以直接实现隐私保护，以及DTM的持久图如何克服这一困难。此外，我们讨论抗离群值鲁棒性的程度如何影响瓶颈距离敏感度的变化率，并提出使用 L^1-DTM，以实现最小敏感度。基于敏感性分析，我们提出首个针对持久图、提供 ϵ-差分隐私的机制，并使用指数机制。我们还为该机制的精度误差建立了上界和下界。这些界表明，该机制的隐私误差接近最优。我们的贡献可更具体地概括如下：

- 我们证明，以瓶颈距离定义的 Čech复形持久图的敏感度不会随着样本量增加而趋于零。

- 我们提出使用距离到测度（DTM）的持久图作为替代方案，并证明 L^p-DTM 持久图的敏感度（以瓶颈距离定义）有保证，为 $O(n^{-1/p})$。因此，我们采用可保证敏感度为 $O(n^{-1})$ 的 L^1-DTM 持久图。

- 我们应用指数机制，其效用函数根据 L^1-DTM 持久图的瓶颈距离定义，以生成差分隐私化的持久图。据我们所知，我们的算法首次尝试开发一种能够生成差分隐私化持久图的机制。我们还推导出该机制精度误差的上界和下界。

- 我们证明，应用于 L^1-DTM 持久图的任何隐私机制，其精度都不可能达到优于与我们机制对应的隐私误差衰减阶上界的衰减阶。这一结果表明，我们的机制可能具有最优隐私误差。

组织结构：本文其余部分组织如下。第2节简要回顾拓扑数据分析和差分隐私的背景及符号。第3节首先考察由Čech复形构造的持久图的敏感度，以及Chazal等人提出的基于DTM获得的抗离群值持久图构造方法的敏感度Chazal等人(2011年）。第4节基于第3节给出的敏感性分析，采用指数机制生成隐私化持久图，并推导其精度的上界和下界。第5节介绍实现该算法的模拟研究。第6节将该算法应用于一个真实世界数据集，其中包含三个人在建筑物内行走时的位置数据，这些数据由智能手机随时间记录。所有证明以及真实数据分析的补充结果均列于附录中。

2 预备知识

本节将介绍持久图，它是一种关于数据形状的统计量，并探讨持久图空间中的一种度量——瓶颈距离及其稳定性。此外，我们还将回顾 ϵ-差分隐私（ϵ-差分隐私）以及指数机制，后者是满足 ϵ-差分隐私的算法之一。

记号 在本文中，对于可能依赖于参数 $n \in \mathbb{Z}_+$ 的实数 A 和 B，我们使用渐近记号 $A \lesssim B$ 或 $A = O(B)$ 表示界 $|A| \leqslant CB$，其中某个绝对常数为 $C > 0$。若常数 C 依赖于某些参数，我们会明确指出；例如，$A \lesssim_{k,d} B$ 表示界 $|A| \leqslant C_{k,d}B$，其中常数 C_k，d 仅依赖于 k 和 d。若 $A \lesssim B$ 和 $A \gtrsim B$，我们将其记为 $A \approx B$。我们还使用记号 $A = o(B)$ 表示渐近极限 $\lim_{n\to\infty} A/B = 0$。此外，对于可能依赖于参数 $n \in \mathbb{Z}_+$ 的随机变量 X 和 Y，我们用 $X = O_p(Y)$ 表示 X/Y 依概率有界，用 $X = o_p(Y)$ 表示 X/Y 依概率收敛于零；这是概率论中的标准记号。设 $(X_n)_{n=1}^{\infty}$ 是一个随机变量序列。我们写作 $X_n = \tilde{O}_p\big(f(n)\big)$，表示对于某个 $k \in \mathbb{Z}_+$，$X_n = O_p\big(f(n)\log^k n\big)$。基本上，所有这类记号都描述了在 n 足够大时的渐近关系。

对于给定的度量空间 $(\mathcal{X}, \mathrm{d})$，$\mathcal{D}_n := \mathcal{D}_n(\mathcal{X}) := \mathcal{X}^n$ 表示对于每个 $n \in \mathbb{Z}_+$，$\mathcal{X}$ 中所有 n 元组构成的集合。

2.1 持久同调与图表

这里，我们简要介绍两种构建数据的持久同调及其对应持久图的方法，这两种方法将在本文的主要讨论中出现。第一种是 Čech复形的持久同调，第二种是连续函数的子水平集的持久同调。我们认为，对持久同调进行直观且具有说明性的描述，就足以理解本文的研究结果。有关持久同调的更详细背景知识，以及一些单纯同调的基础知识，请参见附录 A。若要更深入、全面地理解持久同调，我们建议读者参阅持久同调领域的相关文献，例如 Edelsbrunner 和 Harer (2008年, 2009年); Zomorodian 和 Carlsson (2005年)。关于代数拓扑的基本概念，我们建议读者参阅代数拓扑方面的标准教材，例如 Munkres(1984年); Bredon(1997年)。

设 $D = \{x_1, \ldots, x_n\}$ 是度量空间 $(\mathcal{X}, d)$ 的有限子集。设 $r > 0$ 为正实数。对于每个点 x_i，我们以 x_i 为球心放置一个半径为 r 的球 $B(x_i; r)$。D 上 Čech 复形的持久同调刻画了当 r 变化时，并集 $\cup_{j=1}^{n} B(x_j; r)$ 的同调结构的演变。例如，其第 0 维同调特征表示连通分支，第 1 维同调特征表示其中的环。图 1 展示了如何构造 Čech 复形的持久同调。随着半径 r 变化，一些同调特征会出现并消失，这种同调特征的“诞生”与“消亡”以称为持久图的多重集表示。更准确地说，数据集 D 上 Čech 复形的一个 q 维持久图是一个由有限个点组成的多重集，设这些点的数量为 m，其形式为 (b_i, d_i)，并满足 $0 \leq b_i \leq d_i \leq \infty$ 对每个

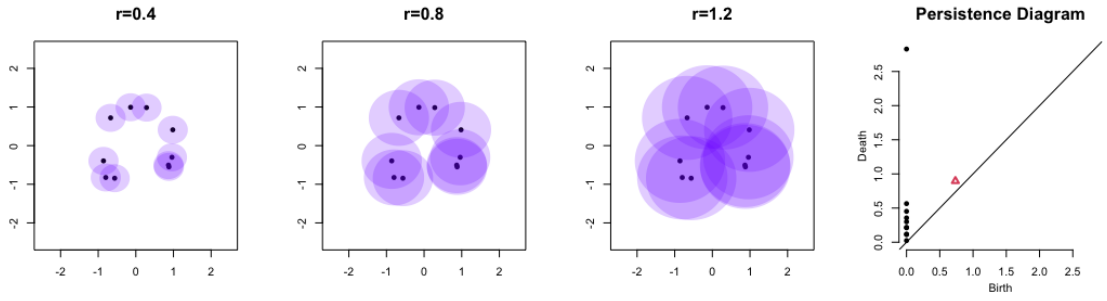


图 1：Čech复形及其持久图的构造：左侧三幅图展示了如何在圆上九个点构成的支撑集上构造 Čech复形。当$r = 0.4$时，存在若干个连通分支，但没有环。当 $r = 0.8$时，出现了一个捕捉圆形状的一维环。当 $r = 1.2$时，环消失，只剩下一个可缩连通分支。最右侧的图是 Čech复形的持久图。每个黑点表示各连通分支的诞生和消亡时刻，红色三角形表示环的诞生和消亡时刻。

$i = 1, \ldots\ldots, m$；每个点 (b_i, d_i) 的存在意味着存在一个 q 维同调特征，它在半径 b_i 时出现，并在半径 d_i 时消失。

对于另一种方法，设 $f_D : \mathcal{X} \to \mathbb{R}$ 是定义在度量空间 $(\mathcal{X}, d)$上的连续函数，且可能取决于给定集合 D。对于每个 $r \in \mathbb{R}$，都可以考虑次水平集 $L_r := \{x \in \mathcal{X} : f_D(x) \leqslant r\}$。与之前构造中考察球并集的演变类似，现在我们考察次水平集 L_r随 r 变化时的演变。图 2展示了某个连续函数的这种演变。在图中，三个连通分支和一个环先后出现，并在某个时刻消失，最终只剩下一个连通分支。各个维度上的出生-死亡对可以表示为持久图，与 Čech复形的情形相同。

一般来说，给定拓扑空间的一个过滤 $\{U_r\}_{r \in R}$，其中 R 是一个线性有序集；也就是说，对于 R 中满足 $r_1 \leq r_2$ 的任意 r_1 和 r_2，都有 $U_{r_1} \subseteq U_{r_2}$。于是，可以定义该序列的持久同调及相应的持久图。在第一个例子中，每个 U_r 都是半径为 r 的球的并集（或者，由这些球得到的单纯复形）；在第二个例子中，每个 U_r 都是子水平集 L_r。

2.2 瓶颈距离下持久图的稳定性

持久图 $\mathcal{P} = \{(b_i, d_i)\}_{i=1}^m$ 本质上是由出生-死亡对 b_i 和 d_i 构成的多重集，它们满足 $b_i \leq d_i$。将持久图“向量化”为某个向量空间中的元素有多种方法。其中最流行的方法之一是，将每个出生-死亡对 (b, d) 表示为位于 (b, d) 处的狄拉克测度 $\delta_{(b,d)}$，并将整个图$\mathcal{P}$ 表示为点测度 $\sum_{i=1}^m \delta_{(b_i,d_i)}$，该测度定义在集合 $\mathcal{T} := \{(x, y) : 0 \leq x \leq y \leq \infty\}$ 上（有关这种向量化方法的详细描述，请参见 Owada (2022) 中的第2节）。通过将持久图视为测度，可以利用测度之间的距离来定义两个持久图之间的距离。一

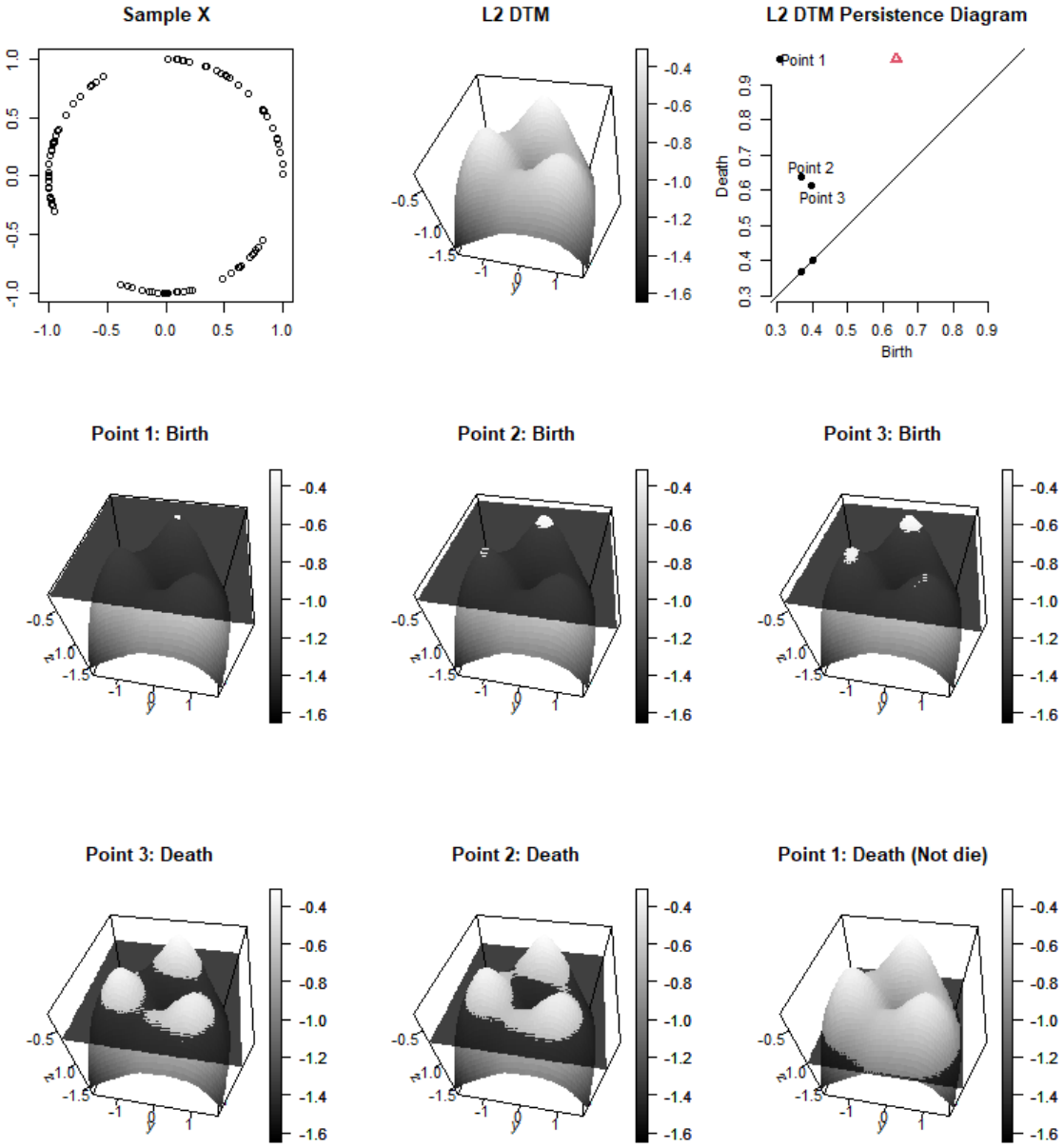


图 2：与圆数据集的 L^2-DTM 对应的过滤：该数据集分布在一个圆上，并对数据的 L^2-DTM 函数进行了可视化（为方便起见，展示的是乘以 -1 后的函数）。图中给出了由该函数构造的持久图。第二列和第三列展示了该函数的子水平集过滤过程。随着过滤过程的推进，三个连通分支分别在 0.308、0.367 和 0.397 处出现。其中一个分支在 0.613 处消亡，另一个在 0.636 处消亡。最后一个分支一直存续到演化结束。

最受欢迎的选择之一是使用测度的 L^∞ Wasserstein 距离，这称为瓶颈距离。具体而言，设 $\mathcal{P}$，$\mathcal{P}'$ 为两个持久图。那么，$\mathcal{P}$ 与 $\mathcal{P}'$ 之间的瓶颈距离定义为

$$\mathrm{d_B}(\mathcal{P},\mathcal{P}') := \min_{g:\breve{\mathcal{P}}\leftrightarrow\breve{\mathcal{P}}'} \max_{z\in\breve{\mathcal{P}}} \| z - g(z) \|_\infty,$$

其中，$\breve{\mathcal{P}}$ 和 $\breve{\mathcal{P}}'$ 分别表示持久图 $\mathcal{P}$ 和 $\mathcal{P}'$，并分别包含对角线上所有点的副本；$g : \breve{\mathcal{P}} \leftrightarrow \breve{\mathcal{P}}'$ 遍历 $\breve{\mathcal{P}}$ 与 $\breve{\mathcal{P}}'$之间的所有双射。换言之，$\mathrm{d_B}(\mathcal{P}, \mathcal{P}')$ 是将一个图中的出生-死亡点逐一配对到另一个图中的出生-死亡点时，以 ℓ_∞ 距离衡量的最小最大代价。当两个图包含的出生-死亡点数量不同时，一个图中剩余的点将与对角线上的点配对。

瓶颈距离的一个关键性质是以下稳定性性质（更多详情，参见Cohen-Steiner等人（2007年）；Chazal等人（2016a年））。设 $\mathcal{P}_q(D)$ 和$\mathcal{P}_q(D')$ 是 q个持久图，它们由度量空间 D中两组 D'和（$\mathcal{X}$，d）的 Čech复形构造而成，则

$$\mathrm{d_B}(\mathcal{P}_q(D), \mathcal{P}_q(D')) \lesssim d_H(D, D'), \tag{2.1}$$

其中

$$d_H(D, D') := \max\left\{ \sup_{x\in D} \inf_{y\in D'} d(x, y), \sup_{y\in D'} \inf_{x\in D} d(x, y)\right\}$$

表示 D 与 D'之间的豪斯多夫距离。类似地，如果 $\mathcal{P}_q(D)$ 和 $\mathcal{P}_q(D')$ 分别由连续驯服函数 f_D 和 $f_{D'}$的次水平集过滤得到，那么

$$\mathrm{d_B}(\mathcal{P}_q(D), \mathcal{P}_q(D'))) \le \sup_{x\in\mathcal{X}} |f_D(x) - f_{D'}(x)|. \tag{2.2}$$

驯服函数的精确定义见定义 24。关于更全面的讨论，请参阅 Cohen-Steiner等人（2007年）。直观地说，一个 $\mathbb{R}$值函数 f称为驯服函数，是指其次水平集的同调变化次数至多有限。

2.3 差分隐私

差分隐私（DP）是一种数学框架，旨在量化所提出的随机算法（称为机制）的隐私泄露，由 Dwork 等人于（2006年）提出。衡量此类隐私风险的第一步，是明确哪些数据库被视为“仅有一项不同”，我们将其称为相邻数据库。如果 D 和 D' 满足 $d(D, D') \le 1$，则称二者相邻，其中 $d(\cdot,\cdot)$ 是数据库之间的某种度量。本文采用汉明距离 $H(\cdot,\cdot)$，它计算 D 与D'之间不同条目的数量。隐私机制 $\mathcal{M} : \mathcal{D}_n \to \mathcal{Y}$ 对任意 $D \in \mathcal{D}_n$均返回一个随机变量 $\mathcal{M}(D)$，算法 $\mathcal{M}$ 的隐私风险可根据定义评估如下：

定义 1（ϵ-差分隐私（ϵ-差分隐私）：Dwork 等人，2006年）给定 $\epsilon \ge 0$，一个隐私机制 $\mathcal{M}$ 在输出空间 $\mathcal{Y}$ 满足 ϵ-差分隐私，当且仅当

$$\mathbb{P}(\mathcal{M}(D) \in S) \le e^\epsilon \mathbb{P}(\mathcal{M}(D') \in S), \tag{2.2}$$

对于每个可测集合 $S \subset \mathcal{Y}$ 以及所有满足 $H(D, D') \le 1$的 D 和 D'。

<escape>该定义刻画了当 ϵ 中的某个实体与 D 中的实体不同时，通过隐私预算参数 D'可能发生的隐私泄露程度。 ϵ越小，就越难以区分 $\mathcal{M}(D)$ 和 $\mathcal{M}(D')$的概率分布，因此也就越难以确定分析 D 还是 D' 时使用了哪个数据集（ $\mathcal{M}$ (Wasserman 和 Zhou, 2010）。</escape>

为了便于理解，我们设想数据集 D 包含通过调查获得的 100 个人的信息。我们称 D 中的一人为人物 A，并假设采用了一个 ϵ- 差分隐私隐私机制 $\mathcal{M}$ ，以发布来自 $\mathcal{M}$ （ D ）的摘要，其中包含关于 D 的一些有用信息。由此可知，对于任何与 D 有 99 人相同、但不包括人物 A 的其他数据集 D' ，$\mathcal{M}$（D）和 $\mathcal{M}$（D'）的概率分布很难区分。因此，很难判断人物 A 是否包含在实际数据集 D 中。请注意，人物 A 是从 D 中任意选取的，因此 D 中每个人的隐私都得到了保护。

虽然任何满足 ϵ-差分隐私的机制都能保护隐私，但并非所有机制都能确保对底层效用而言具有良好的性能。不难想象，脱敏统计量可能会因保护隐私而引入过多噪声，从而严重损害效用。相比之下，指数机制是一种通用技术，既能兼顾效用，又能将隐私泄露控制在预算 ϵ 之内。

命题 2（指数机制：McSherry 和 Talwar，2007年）设 $n \in \mathbb{Z}_+$并设 $\{u_D : \mathcal{Y} \to \mathbb{R} : D \in \mathcal{D}_n\}$ 为一组效用函数。假设其敏感度 Δ（u）是有限的：

$$\Delta(u) = \sup_{H(D,D')\leq 1} \sup_{y\in\mathcal{Y}} |u_D(y) - u_{D'}(y)| < \infty, \tag{2.3}$$

其中，上确界取遍所有相邻的 D 和 D'，并假设对于所有 $D \in \mathcal{D}$，都有 $\int \exp\left(u_D(y)\right) d\nu(y) < \infty$，其中 ν 是 $\mathcal{Y}$ 中的一个测度。如果 Δ 满足 $\Delta(u) \leq \Delta < \infty$，则机制集合 $\{\mathcal{M}(D) : D \in \mathcal{D}\}$ 中的每个机制，其相对于 ν 都具有概率密度。

$$p_D(y) \propto \exp\left(\frac{\epsilon}{2\Delta} u_D(y)\right), \tag{2.4}$$

满足 ϵ-*DP*。

指数机制可以轻松应用于各种具有效用函数的问题。最简单的例子之一是计数统计量。我们将数据集 D 的计数统计量 count(D) 定义为数据集 D 中具有某种特定属性的数据点数量，并定义效用函数 $u_D(y) := -|y - \text{count}(D)|$，其敏感度为 1。当 y 接近 count(D) 时，该效用函数会赋予更高的值。许多机器学习和统计推断问题都可以通过指数机制以隐私保护的方式加以处理，前提是能够定义适当的效用函数，例如经验风险或似然函数（ Huang 和 Kannan，2012；Awan 等人，2019年；Cummings 等人，2019年；Lu 等人，2022年）。

命题 3（指数机制的效用：Dwork 和 Roth，2014年）设$\text{OPT}_D = \max_{y\in\mathcal{Y}} u_D(y)$ 为效用函数所能达到的最优值

对于所有输出，给定数据库 D。设 Y 为一个随机变量，其密度由式（2.4）给出。则，

$$\mathbb{P}\left[u_D(Y) \leq \mathrm{OPT}_D - \frac{2\Delta}{\epsilon}\left(\log|\mathcal{Y}| + t\right)\right] \leq e^{-t}, \tag{2.5}$$

对每个 $t \geq 0$ 都成立。因此，

$$u_D(Y) = \mathrm{OPT}_D + O_p\left(\frac{\Delta \log|\mathcal{Y}|}{\epsilon}\right). \tag{2.6}$$

必须谨慎选择指数机制中的效用函数，以确保命题 3 给出的误差率能够转化为私有输出的最优速率。例如，Awan 等人 (2019年) 证明，当效用函数在其最大值处具有二次泰勒展开时，指数机制中用于保护隐私的随机性通常会产生 $O_p(1/\sqrt{n})$ 噪声，其通常与非私有统计估计问题中的噪声处于同一数量级。另一方面，Reimherr 和 Awan (2019年) 证明，对于某些在局部可由绝对值函数近似的效用函数，隐私保护所需的随机性可能低至 $O_p($ $1/n)$。

为了生成差分隐私持久图，我们提出使用指数机制，其效用函数定义为私有持久图与非私有持久图之间瓶颈距离的相反数。

3 持久图在瓶颈距离下的敏感度

大多数差分隐私算法都需要量化：在给定数据中改变一个点，会使某个统计量的值发生多大变化。统计量可能发生的最大变化量通常被称为该统计量的敏感度。在本研究中，我们将由数据集 D 构造的持久图视为一种统计量，用以估计数据所依托空间的同调结构，并使用瓶颈距离在持久图空间上定义一个度量。因此，为了将差分隐私机制应用于持久图，我们首先应当估计持久图关于瓶颈距离的敏感度；也就是说，我们将分析瓶颈距离 $\mathrm{d_B}(\mathcal{P}_D,$ $\mathcal{P}_{D'})$ 可能达到多大，其中，数对 (D, D') 表示一对相邻数据集。注意， $\mathcal{P}_D$ 和 $\mathcal{P}_{D'}$ 分别表示在给定的持久同调构造方式下，由数据集 D 和 D' 构造的持久图。

在差分隐私中，为了确保估计量具有一致性，数据规模增大时，敏感度必须趋于 0。我们的关键观察是，如果所选的持久同调构造方式对离群值不具有鲁棒性，那么即使数据规模（记作 n）不断增大，相应持久图的敏感度也可能不会趋于 0。

我们证明，Čech复形的持久图的敏感度即使在数据规模增长至无穷大时，也无法收敛于0。为解决这一问题，我们提出使用称为到测度的距离（DTM）的概念，Chazal等人 Chazal等人(2018年) 对此进行了深入讨论，从而诞生了对异常值具有鲁棒性的持久图。此外，在DTM的各种构造版本中，我们提出使用 L^1-DTM，其敏感度最小。

在继续讨论主要的敏感性分析之前，我们先澄清一下术语。在本节引言中，我们一直用“敏感度”一词来指

两种不同的量：瓶颈距离的敏感度，以及指数机制效用函数的敏感度。为避免混淆，接下来我们将持久图的瓶颈距离敏感度称为基础敏感度，这一术语由Awan 和 Wang （2022年）引入。Čech复形中第 q阶持久图的瓶颈距离基础敏感度记为 $\Delta_q^{\check{\mathrm{C}}\mathrm{ech}}$，而来自 L^1-DTM 的则记为 Δ_q^{DTM}。它们的精确定义将在以下各小节中分别给出。除此之外，我们将“敏感度”一词专用于指数机制给定效用函数的敏感度。

3.1 Čech复形的持久图敏感度

下面通过一个例子说明，Čech复形的构造无法具有递减的基础敏感度。图 3很好地展示了以下例子中的情形。请注意，图 3 使用 Vietoris-Rips复形而非 Čech复形绘制图形。Vietoris-Rips复形是 Čech复形的一种变体，其计算效率更高。事实上，Vietoris-Rips复形的过滤与 Čech复形的过滤所包含的信息基本相同。Vietoris-Rips复形的定义及其与Čech复形的关系见附录A.3。

例 1 设 D 是 $\mathbb{R}^2$ 中的 n 个点，这些点紧密聚集成恰好两个簇。用 x 表示位于这两个簇中点的点，并将 D 中的一个点移动到 x，从而得到数据集 D'。现在进一步设想，随着 n 增大，D 和 D' 中的点配置保持不变，并推导由 D 和 D' 的 *Čech*复形得到的第0维持久图。此时，D 中的连通分支会迅速收缩为两个簇，而孤立点 $x \in D'$ 则会产生一个持续时间更长的额外连通分支。两个持久图之间的这种差异，使它们之间的瓶颈距离无法趋于0。更准确地说，它们之间的瓶颈距离始终不小于点 x 到 D 中簇的距离。

下面的定理表明，这种现象十分普遍。我们将由数据集 D 上的Čech复形构造的第 q个持久图记为$\mathcal{P}_q^{\check{\mathrm{C}}\mathrm{ech}}(D)$。

引理4 设 $D = \{x_1, \ldots, x_n\}$ 是欧几里得空间 $\mathbb{R}^d$的一个子集。设 $\{d_1, \ldots, d_m\}$ 为 $\mathcal{P}_0^{\mathrm{Cech}}(D)$ 中不同有限死亡时间的集合，其中 $0 < d_1 < \cdots < d_m < \infty$。设 $\delta = d_m - d_{m-1}$（若 $m = 1$，则设 $\delta = d_1$）。那么，可以取一个集合 D'，使得 $|D \setminus D'| + |D' \setminus D| \leqslant 1$满足

$$\mathrm{d_B}\left(\mathcal{P}_0^{\check{\mathrm{C}}\mathrm{ech}}(D), \mathcal{P}_0^{\check{\mathrm{C}}\mathrm{ech}}(D')\right) \geq \min\{\delta, d_m/2\}.$$

粗略地说，可以通过构造一个数据集来证明该定理 D'，使其在 D 的 Čech复形过滤中最“显著的连通分支”中点处额外包含一个点，也就是说，这些连通分支在时刻 d_m 消亡。详细证明见附录B.1。

除非另有说明，以下均假设所有数据集都支撑在 $\mathbb{R}^d$ 的一个有界子集 E 中。我们定义关于 Čech复形的基础敏感度 Δ_q^{Cech}:

$$\Delta_q^{\check{\text{C}}\text{ech}} := \sup_{H(D,D')\leq 1} \mathrm{d_B}\left(\mathcal{P}_q^{\check{\text{C}}\text{ech}}(D), \mathcal{P}_q^{\check{\text{C}}\text{ech}}(D')\right).$$

注意，稳定性定理（2.1）蕴含基础敏感度的以下上界:

$$\Delta_q^{\check{\text{C}}\text{ech}} \leq \mathrm{diam}E$$

对每个非负整数 q。引理 4 给出了 $q = 0$的基础敏感度的相应下界。此外，这些上下界表明，效用函数 v_D 定义为

$$v_D(\mathcal{P}) := -\mathrm{d_B}\left(\mathcal{P}_0^{\check{\text{C}}\text{ech}}(D), \mathcal{P}\right), \tag{3.1}$$

其敏感度为常数量级:

定理5 设给定的数据生成过程由欧几里得空间中的一个有界子集E 支撑。则有

$$\Delta_0^{\check{\text{C}}\text{ech}} \geq \frac{\mathrm{diam}E}{4}.$$

此外，效用函数 v_D 定义于（*3.1*）中，满足

$$\frac{1}{4}\mathrm{diam}E \leq \sup_{H(D,D')\leq 1} \sup_{\mathcal{P}} |v_D(\mathcal{P}) - v_{D'}(\mathcal{P})| \leq \mathrm{diam}E$$

定理 5 表明了为何开发适用于Čech复形的隐私机制具有挑战性：从其持久图的瓶颈距离来看，Čech复形具有极高的敏感性，以至于效用函数 $v_D(\cdot)$ 的敏感度始终保持不变，而不受数据集规模 n 的影响。这意味着，即使 n 不断增大，使用该效用函数的指数机制仍会持续添加固定规模的噪声。这使得即使在 n极其庞大的情况下，瓶颈距离也无法变得很小。

3.2 DTM 持久图的敏感度

DTM由Chazal等人（2011年）提出，为克服对离群点的敏感度提供了一种新方法。DTM通过测量每个点与概率测度支撑集的稠密部分之间的距离来进行分析。这样，离群点对应的距离就相对较大。因此，在考察DTM次水平集的过滤时，离群点产生的拓扑特征会在过滤的后期出现，也可能在整个过滤过程中都不会出现。关于DTM的更深入讨论可参见Chazal等人（2018年）；Anai等人（2020年）；Oudot （2015年）。凭借DTM的性质，基于DTM的持久图很可能具有小得多的敏感度，因此它或许能为我们提供一种合适的拓扑数据分析统计量，以此构建我们的隐私化机制。事实上，我们证明，基于DTM的持久

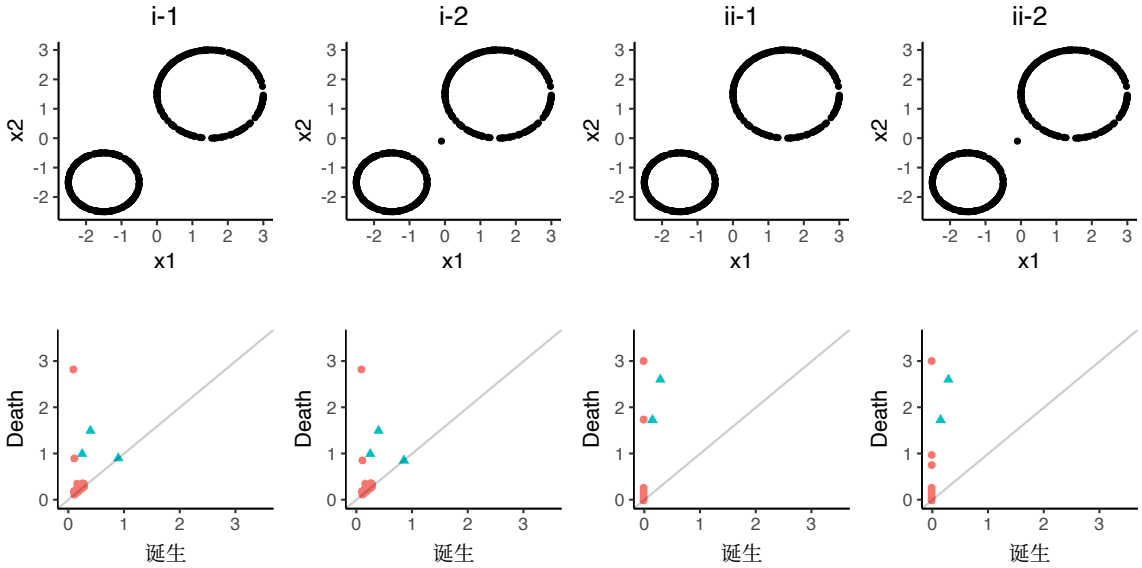


图 3： D 和 D'上的持久图：红色圆形和绿色三角形分别表示连通分支和环。列 (i-1) 和 (i-2) 分别对应 D 和 D' 上使用 L^1-DTM得到的结果，这两个图在连通分支方面的瓶颈距离为 0.042。列 (ii-1) 和 (ii-2) 来自 Vietoris-Rips复形，两个图之间的距离为 0.762。

图表的敏感度会随着 n 趋于无穷而收敛至0，但衰减速率取决于我们采用哪一类DTM设计。

基本上，DTM 被定义为某个函数的 L^p 范数。DTM 的原始版本被定义为一种 L^2 型量。我们证明， L^2 型 DTM 所生成的持久图的基础敏感度有上界 $O(n^{-1/2})$；并且我们发现，每个 L^p-DTM 都会得到类似的基础敏感度上界： $O(n^{-1/p})$。基于这一观察，我们聚焦于基础敏感度衰减最快的 L^1-DTM。此外，我们还验证了由 L^1-DTM 得到的持久图的基础敏感度具有常数倍 n^{-1} 的下界。换言之，我们对 L^1-DTM 的敏感性分析在常数因子范围内是紧的。

我们给出一般 L^p-DTM 的定义及其经验实现。要获得持久图的上界，关键性质是 DTM 的所谓 Wasserstein 稳定性；过去的文献对此已有广泛讨论，例如参见Chazal等人 (2016年b)。根据 Wasserstein 稳定性，我们推导出 $n^{-1/p}$ 的上界，其速率为 L^p-DTM。通过构造一个恰好达到该下界的具体示例，我们建立了 n^{-1} 的匹配下界，其速率为 L^1-DTM。所有证明均列于附录。

定义 6（到测度的距离）设 μ 为一个概率测度， X 为一个随机变量，其概率分布为 μ。对于给定的 μ、 $0 < m < 1$和 $p \geq 1$,

L^p 在分辨率为 m 时，到测度的距离 μ 定义为

$$\delta^{(p)}(x) := \delta^{(p)}_{\mu,m}(x) := \left[\frac{1}{m}\int_{u=0}^{m}\left(G_x^{-1}(u)\right)^p du\right]^{1/p},$$

其中 $G_x(t) = P\big[\parallel X - x \parallel \leq t\big]$。这里，$\parallel \cdot \parallel$ 表示欧几里得空间中的 ℓ^2-范数。

超参数 m 决定所采用的平滑效应强度，这与核密度估计中带宽的作用相似。一个自然的经验近似如下。

定义 7（DTM 的经验版本）设 X_1、...、X_n 是从概率分布 μ 中获得的独立同分布样本，μ_n 是定义在这些样本上的经验概率测度，即：

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$$

在分辨率 m 下，μ 的经验 L^p-*DTM* 记为 $\hat{}\delta^{(p)}$，其定义为在分辨率 m 下，μ_n 的 L^p-*DTM*；即：

$$\hat{\delta}^{(p)}(x) := \delta^{(p)}_{\mu_n,m}(x) = \left[\frac{1}{k}\sum_{X_i\in N_k(x)}\parallel X_i - x\parallel^p\right]^{1/p},$$

其中 $k = \lceil mn\rceil$ 和 $N_k(x)$ 是包含 x 在 X_1、...、X_n 中最近的 k 个邻居的数据点集合。这里，数据点之间的距离以欧几里得空间中的 ℓ^2-范数度量。

L^p-DTM 的关键定量性质称为其 Wasserstein 稳定性，具体如下：设 μ 和 ν 是定义在同一度量空间上的概率测度，则

$$\sup_x\left|\delta^{(p)}_{\mu}(x) - \delta^{(p)}_{\nu}(x)\right| \leq \frac{1}{m^{1/p}}W_p(\mu,\nu), \tag{3.1}$$

其中 $W_p(\mu,\ \nu)$ 表示 p-Wasserstein 距离，即 μ 与 ν之间的距离。更多细节请参见Chazal等人(2016年b)。设 D 和 D' 是相邻数据集。

设 $\mathcal{P}^{\mathrm{DTM}_p}{}_q\,(D)$ 表示由数据集 D 的经验分布的 L^p-DTM 次水平集过滤所构造的第 q 个持久图。关于 DTM 的 $\Delta_q^{\mathrm{DTM}_p}$ 基础敏感度为

$$\Delta_q^{\mathrm{DTM}_p} := \sup_{H(D,D')=1}\mathrm{d_B}\left(\mathcal{P}_q^{\mathrm{DTM}_p}(D), \mathcal{P}_q^{\mathrm{DTM}_p}(D')\right).$$

根据稳定性定理 (2.2) 和 Wasserstein 稳定性 (3.1)，通过量化相邻数据集上经验分布之间的 $\Delta_q^{\mathrm{DTM}_p}$-Wasserstein 距离，可以得到基础敏感度 p的以下上界。

定理 8（由 L^p-DTM 构造的持久图的敏感度）设 D 和 D' 是有界集合 E 中的有限子集，其中 $\mathbb{R}^d$ 满足 $|D| = |D'| = n$ 和$H(D, D') = 1$。那么，对于每个 $q \in \mathbb{Z}^{\geq 0}$，

$$\Delta_q^{\mathrm{DTM}_p} \leq \frac{\mathrm{diam}E}{m^{1/p} n^{1/p}}.$$

事实上，根据 Wasserstein 稳定性（3.1），通过估计以下对象上的经验分布之间的 p-Wasserstein 距离即可得到定理的结果： D和 D'。详细证明见附录 B.1。

根据定理 8，每个 L^p-DTM 的基础敏感度都有一个不超过 $O(n^{-1/p})$的上界。特别地，当p 取值为 1时，该保证上界最小：

$$\mathrm{d_B}\left(\mathcal{P}_q^{\mathrm{DTM}_1}(D), \mathcal{P}_q^{\mathrm{DTM}_1}(D')\right) \leq \frac{\mathrm{diam}E}{mn},$$

事实上， L^1-DTM 的上界在常数因子意义下是紧的。

命题 9（ L^1-DTM 的敏感度下界）假设 $m < 1/2$。那么，对于每个正整数 n，存在一对集合 D 和 D' 满足$|D| = |D'| = n$, $H(D, D') = 1$，并且

$$\mathrm{d_B}\left(\mathcal{P}_0^{\mathrm{DTM}_1}(D), \mathcal{P}_0^{\mathrm{DTM}_1}(D')\right) = \frac{\mathrm{diam}E}{2k},$$

其中 $k = \lceil mn \rceil$。

通过构造一对相邻数据集 D 和 D'，使其达到所提出的距离，即可得到证明。事实上，图 3 所示的数据集便达到该距离。详细证明请见附录 B.1。

现在，我们介绍用于设计隐私机制的效用函数。令 Pers 表示配备瓶颈距离的持久图空间。对于任意给定的数据集 D 和任意非负整数 q，我们将函数 $u_D^{(q)} : \mathrm{Pers} \to \mathbb{R}$定义如下：

$$u_D^{(q)}(\mathcal{P}) := -\,\mathrm{d_B}\left(\mathcal{P}, \mathcal{P}_q^{\mathrm{DTM}_1}(D)\right),$$

令 ℓ 为选定的非负整数。我们的效用函数 $u_D : \left(\mathrm{Pers}\right)^{\ell+1} \to \mathbb{R}$ 定义如下：

$$u_D(\mathcal{P}_0, \ldots, \mathcal{P}_\ell) := \sum_{q=0}^{\ell} u_D^{(q)}(\mathcal{P}_q). \tag{3.2}$$

根据基础敏感度的上界和下界，我们可以得到如下敏感度的上界和下界：

推论 10 对于选定的 $\ell \geq 0$，令效用函数 u_D 定义如（3.2）所示。则满足以下结论。

$$\frac{\mathrm{diam}E}{2\lceil mn \rceil} \leq \sup_{H(D,D')=1} \sup_{\mathcal{P} \in \mathsf{Pers}} |u_D(\mathcal{P}) - u_{D'}(\mathcal{P})| \leq (\ell+1)\frac{\mathrm{diam}E}{mn}$$

注 11 效用函数 u_D 的可加性使我们能够确立推论 *10*中的上界和下界。注意，该推论的下界是根据命题 *9* 的结果推导得出的，而该结果仅对 0 阶持久图有效；但是， u_D 的可加形式使其成为整个效用函数敏感度的下界。

注 12 请注意，虽然 L^1-*DTM* 的下界与其上界的速率一致，但目前我们尚未获得其他 L^p-*DTM* 的匹配下界。因此，一般 L^p-*DTM* 的基础敏感度或许可以改进。例如，在图*3*所示的情形中，我们通过实证验证了 L^2-*DTM* 对 D 和 D' 的持久图之间的瓶颈距离也为 $O(n^{-1})$。

4 使用指数机制与 L^1-DTM

本节将介绍如何实现指数机制，以生成隐私化持久图。具体而言，我们将给出生成隐私化持久图所依据的精确概率分布，并逐步概述一种用于近似该分布的马尔可夫链蒙特卡罗过程。

设 $D = \{X_1, \ldots, X_n\}$ 是由服从同一概率分布 μ 的独立同分布样本构成的数据集。为简洁起见，对于每个 $q \geq 0$ ，我们用 $\mathcal{P}_q(D)$ 表示由经验测度 μ_n 的 L^1-DTM 得到的第 q 个持久图；上一节中将其记作 $\mathcal{P}_q^{\mathrm{DTM}_1}(D)$。对于给定的 ℓ，我们令 $\mathcal{P}(D) := (\mathcal{P}_0(D) \ldots, \mathcal{P}_\ell(D))$。设 $\mathcal{P}_q(\mu)$ 为测度 μ 的 L^1-DTM $\delta_\mu^{(1)}$ 得到的第 q 个持久图；对于给定的 ℓ，定义 $\mathcal{P}(\mu) := (\mathcal{P}_0(\mu) \ldots, \mathcal{P}_\ell(\mu))$。此外，设 $\mathcal{P}_{\mathrm{DP}} = (\mathcal{P}_{0,\mathrm{DP}}, \ldots, \mathcal{P}_{\ell,\mathrm{DP}})$ 为通过我们的算法生成的一组隐私化持久图。

我们从两个不同的角度分析私有化持久图的误差。首先，我们估计 $\mathrm{d_B}(\mathcal{P}(\mu)\ \mathcal{P}_{\mathrm{DP}})$。这一量表示私有化持久图与原始数据生成过程的持久图之间的误差。从统计学视角来看， $\mathcal{P}(\mu)$ 可视为表征真实数据生成过程的参数。因此，量 $\mathrm{d_B}(\mathcal{P}(\mu)\ \mathcal{P}_{\mathrm{DP}})$ 衡量的是通过隐私化实际统计量 $\mathcal{P}(D)$ 得到的隐私化统计量 $\mathcal{P}_{\mathrm{DP}}$ 对参数 $\mathcal{P}(\mu)$ 进行估计时的误差。第二种方法是估计量 $\mathrm{d_B}(\mathcal{P}(D)\ \mathcal{P}_{\mathrm{DP}})$，它衡量隐私化过程对原始非隐私化统计量造成的扭曲程度。

4.1 生成隐私化持久图

指数机制的设计通过指定一个输出空间 $\mathcal{Y}$和一个效用函数 u_D 来表述： $\mathcal{Y} \to \mathbb{R}$ 对于每个数据集 D。由于我们要进行隐私保护的目标是持久图，因此自然的选择是将所有可能的持久图构成的空间作为输出空间，我们将其记为 Pers。效用函数的第一个候选是函数 u_D ，其定义见（3.2）。然而，其输出空间 Pers包含具有任意多个诞生-消亡对的持久图；也就是说，从整个 Pers 中采样持久图不可避免地是一个无限维问题，在计算上存在技术难题。为了绕过这种

为了解决这一问题，我们预先指定一个超参数 $M \in \mathbb{Z}_+$，即一个正整数，并且只考虑在每个维度 q 上至多具有 M 个出生-死亡对的持久图空间 Pers_M。对于每个受限空间 Pers_M，给定任意数据集 D，我们将函数 $u_D^{(q)} : \mathrm{Pers}_M \to \mathbb{R}$ 重新定义如下：

$$u_D^{(q)}(\mathcal{P}) := -\,\mathrm{d_B}(\mathcal{P}, \mathcal{P}_q(D)).$$

效用函数 u_D 也以与（3.2）相同的方式，利用重新定义的 $u_D^{(q)}$ s 重新定义。也就是说，效用函数 u_D： $\left(\mathrm{Pers}_M\right)^{\ell+1} \to \mathbb{R}$ 定义为

$$u_D(\mathcal{P}_0, \ldots, \mathcal{P}_\ell) := \sum_{q=0}^{\ell} u_D^{(q)}(\mathcal{P}_q). \tag{4.1}$$

请注意，推论 10 中得到的上界和下界，对于（ u_D 4.1）中定义的效用仍然成立。

在（4.1）中选择效用函数 u_D 后，可以确定生成隐私化持久图所依据的概率分布。在介绍指数机制之前，我们先引入 Pers_M 的离散化版本，以便于第 4.2 节的分析。请注意，Pers_M 中的每个持久图都可以¯视为左上三角形 $\mathcal{T}$ 中的至多 M 个点组成的族：$= \{(x, y) : 0 \le x \le y \le \mathrm{diam}E\}$。我们不直接使用 $\mathcal{T}$，而是将其离散化为有限个离散点；¯例如，$\mathcal{T}$ 中的一组等间距有限点就可以作为这样的离散化。通过用 N^2 个离散点对集合 $\bar{\mathcal{T}}$ 进行离散化，可以得到 Pers_M 的离散化；也就是说，Pers_M 的离散化版本是由至多含有 M 个点的集合构成的族，这些点位于¯的离散化版本中。请注意，Pers_M的这种离散化的基数至多为 N^{2M}。对于给定的正整数 N，我们将 $\widetilde{\mathrm{Pers}}_{M,N}$定义为通过用 N^2 个等间距离散点对 $\bar{\mathcal{T}}$ 进行离散化而得到的 Pers_M 的离散化。因此，我们的$\widetilde{\text{}}$ $\widetilde{\text{}}$指数机制实际上是在空间 $\mathrm{Pers}_{M,N}$上执行的。空间 $\mathrm{Pers}_{M,N}$ 是实际的输出空间，由以下机制生成的私有持久图都属于该空间。

命题 13 设 $\ell \ge 0$ 为固定值，效用 u_D 定义于（4.1）中。令 $p(\cdot)$ 表示由以下表达式刻画的概率密度函数：

$$p(\mathcal{P}_{\mathrm{DP}}) \propto \exp\left(\frac{\epsilon}{2\Delta} u_D(\mathcal{P}_{DP})\right) = \exp\left(-\frac{\epsilon}{2\Delta}\sum_{q=0}^{\ell} \mathrm{d_B}\left(\mathcal{P}_q(D), \mathcal{P}_{q,\mathrm{DP}}\right)\right) \tag{4.2}$$

相对于集合 $(\widetilde{\mathrm{Pers}}_{M,\ N})^{\ell+1}$ 上的均匀分布作为基准测度。在（4.2）中， Δ 定义如下。

$$\Delta := (\ell + 1)\frac{\mathrm{diam}E}{mn}$$

并且 $\mathcal{P}_{\mathrm{DP}} = (\mathcal{P}_{0,\ \mathrm{DP}},\ \ldots,\ \mathcal{P}_{\ell,\ \mathrm{DP}})$。那么，由密度（4.2）刻画的指数机制满足 ϵ-差分隐私。

为了从分布（4.2）中生成样本，我们采用 Metropolis-Hastings 算法。该算法的详细流程总结于附录 C.1。

注 14 请注意，建立命题 *13*并不需要离散化，但要推导下一节中的效用结果（例如命题 *15* 和定理*17*），则需要进行离散化。通过使用更复杂的技术，有可能从我们的分析中移除这种离散化。

4.2 隐私化持久图的分析

令 $\ell \geq 0$ 确定。回顾一下，我们已将隐私算法生成的持久图的输出空间限制为每个维度包含 M 个点，并且这些点落在连续持久图空间的离散化版本上。为了在误差量化中纳入这些限制，我们定义 $\mathcal{P}_{\mathrm{OPT}}$ 为来自 $\mathcal{P}(D) = (\mathcal{P}_0(D) \ldots, \mathcal{P}_\ell(D))$ 的最接近持久图，该图可以由隐私算法生成。更准确地说，对于每个 q

$$\mathcal{P}_{q,\mathrm{OPT}} := \underset{\mathcal{P} \in \mathsf{Pers}_M}{\operatorname{argmin}} \ \mathrm{d_B}(\mathcal{P}, \mathcal{P}_q(D)),$$

其中，$\mathcal{P}$ 遍历所有至多包含 M 个元素的持久图，且

$$\mathcal{P}_{\mathrm{OPT}} := (\mathcal{P}_{0,\mathrm{OPT}}, \ldots, \mathcal{P}_{\ell,\mathrm{OPT}}).$$

$\widetilde{\text{类似地，}}$ $\mathcal{P}_{\mathrm{OPT}}$ 在离散空间 $\mathsf{Pers}_{M,N}$ 上的对应概念定义如下。对于每一个 $q \geq 0$,

$$\widetilde{\mathcal{P}}_{q,\mathrm{OPT}} := \underset{\mathcal{P} \in \widetilde{\mathsf{Pers}}_{M,N}}{\operatorname{argmin}} \ \mathrm{d_B}\left(\mathcal{P}, \mathcal{P}_q(D)\right)$$

and

$$\widetilde{\mathcal{P}}_{\mathrm{OPT}} := (\widetilde{\mathcal{P}}_{0,\mathrm{OPT}}, \ldots, \widetilde{\mathcal{P}}_{\ell,\mathrm{OPT}}).$$

此外，对于任意一对持久图的 $(\ell+1)$-元组 $\mathcal{P} = (\mathcal{P}_0, \ldots, \mathcal{P}_\ell)$ 和 $\mathcal{P}' = (\mathcal{P}'_0, \ldots, \mathcal{P}'_\ell)$，我们定义

$$\mathrm{d_B}(\mathcal{P}, \mathcal{P}') := \sum_{q=0}^{\ell} \mathrm{d_B}(\mathcal{P}_q, \mathcal{P}'_q).$$

一般而言，在有关指数机制的文献中，针对指数机制所覆盖空间中使误差最小的取值，已有广泛的分析。例如，参见Dwork 和 Roth （2014年）。关于$\mathcal{P}_{q,\ \mathrm{OPT}}$的一项关键结果总结于命题2.5中。因此，可以得到以下估计。$\widetilde{\text{回顾一下，}}$离散化空间 $\mathsf{Pers}_{M,\ N}$ 是通过将左上三角形 $\bar{T}$ 离散化为 N^2 个等间距离散点而得到的。

命题 15 设 $\mathcal{P}_{\mathrm{OPT}}$ 如上所定义，并设 $\mathcal{P}_{\mathrm{DP}}$ 为由第 *4.1*节所述的指数机制得到的私有持久图。假设将左上三角形 $\bar{T}$ 离散化为 N^2 个等间距点，则有以下结论。

$$\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}_{\mathrm{DP}}) = O_p\left(\frac{(\ell+1)^2 M \log N}{n\epsilon} + \frac{1}{N}\right)$$

特别地，如果取 $N = n$，则有

$$\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}_{\mathrm{DP}}) = O_p\left(\frac{(\ell+1)^2 M \log n}{n\epsilon}\right) = \tilde{O}_p\left(\frac{(\ell+1)^2 M}{n\epsilon}\right).$$

注 16 事实上，指数机制本身只能直接保证隐私化图集中在离散空间中的最优图上。更具体地说，我们有

$$\mathrm{d_B}(\widetilde{\mathcal{P}}_{\mathrm{OPT}}, \mathcal{P}_{\mathrm{DP}}) = O_p\left(\frac{(\ell+1)^2 M \log N}{n\epsilon}\right). \tag{4.3}$$

另一方面，只要我们采用足够精细的离散化，就显然有，~距离 $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}_{\mathrm{OPT}})$ 与误差（4.3）相比可以忽略。例如，取$N = N(n) = n$ 可确保此类近似误差的阶为 $O_p(n^{-1})$，而（4.3）中的项 $\log N$ 只会增加 $\log n$ 量级的误差。这保证了命题 15 中的结果。

为了利用上述结果，我们可以如下估计这两类误差。

$$\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}(\mu)) \le \mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}_{\mathrm{OPT}}) + \mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(\mu)) \tag{4.4}$$

and

$$\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}(D)) \le \mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}_{\mathrm{OPT}}) + \mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(D)). \tag{4.5}$$

因此，剩下的工作是分别估计 $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(\mu))$ 和 $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(D))$。

在陈述本节的主要定理之前，我们想总结一下所使用的术语，以称呼我们所关注的各个误差项。首先，我们将 $\mathrm{d_B}(\mathcal{P}(D)\ \mathcal{P}(\mu))$ 称为估计误差，因为 $\mathcal{P}(D)$ 可以被视为估计 $\mathcal{P}(\mu)$ 的统计量。按照差分隐私文献中的惯例，将 $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}_{\mathrm{OPT}})$ 称为隐私误差。另一方面，我们将量 $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(D))$ 称为近似误差，因为 $\mathcal{P}_{\mathrm{OPT}}$ 是空间 Pers_M 中对 $\mathcal{P}(D)$ 的最佳近似。与前两个项不同，据我们所知，文献此前尚未分析过形式为 $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(\mu))$ 的量。正如我们所指出的，$\mathcal{P}(\mu)$ 可以被视为描述生成数据的概率测度 μ D 的总体参数。基于这一观点，我们将量 $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(\mu))$ 称为约束估计误差，并将相应的量 $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}(\mu))$ 称为隐私-估计误差，以表明该量可以解释为使用隐私化统计量 $\mathcal{P}_{\mathrm{DP}}$ 估计总体参数 $\mathcal{P}(\mu)$ 时产生的误差量。

如果我们可以选择足够大的 M，那么项 $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(\mu))$ 和 $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(D))$ 都可以根据经验分布在 D 上收敛到测度 μ 的情况，使用 Wasserstein 距离 W_1 进行估计（见命题25）。结果表明，这两项都以 $O_p((\ell+1)n^{-1/d})$ 为界。因此，令 $M = M(n)$ 为缓慢递增的序列，我们就能达到这样的界，同时不会降低命题 15 中得到的隐私误差。

定理 17（隐私估计误差的上界）令 $M = M(n) = \log n$，且 $N(n) = n$。那么，对所有足够大的 n，以下估计成立。

$$\begin{aligned}\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}(\mu)) &\le \mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}_{\mathrm{OPT}}) + \mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(\mu)) \\ &= \tilde{O}_p\left(\frac{(\ell+1)^2}{n\epsilon} + \frac{(\ell+1)}{n^{1/d}}\right).\end{aligned}$$

自然会让人想知道所得到的上界有多精确。对于总体估计误差（以及估计误差），遗憾的是，只要论证依赖于 $n^{-1/d}$ 上经验测度到测度 D 的 Wasserstein 收敛 μ，就不可避免地会得到这一收敛速率。这意味着，如果可能获得更快的速率，就需要采用不同的方法，以便更精确地考察 DTM 的次水平集的同调特征的诞生与消亡。在拓扑数据分析的文献中，有一些研究借助几何学工具考察了 Čech复形的这类特征。例如，参见 Bobrowski 和 Adler (2014年)。这些方法或许能为如何更精确地分析 DTM 的持久图提供启示。

关于隐私误差，我们认为，除常数和对数因子外，该误差是精确的。回顾一下，$\mathcal{P}_{\mathrm{OPT}}$ 被定义为我们的隐私算法输出范围内、与 $\mathcal{P}(D)$距离最小的持久图。根据这一定义，我们可以推测，在相当大的概率下，距离 $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}_{\mathrm{OPT}})$ 可能小于距离 $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}(D))$。这意味着，如果我们能够找到一个与 $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}_{\mathrm{OPT}})$ 的上界相匹配的 $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}(D))$ 下界，就说明我们的估计可能是精确的。在下面的定理中，我们证明，在某些温和条件下，不存在隐私误差小于 $1/(n\epsilon)$ 的 ϵ-差分隐私机制，其中隐私误差是相对于来自 L^1-DTM 的持久图而言的。下文中，我们回顾一下，$\mathcal{P}_0(D)$ 表示将 L^1-DTM 应用于给定数据集上的经验测度所得到的零维持久图，其定义如前所述。

定理 18 假设 $m < 1/2$。令 n 为正整数，且 $\mathcal{M}$ 为任意一个 ϵ-差分隐私机制，该机制会生成数据集 D的隐私化持久图 $\mathcal{M}(D)$ 。假设 ϵ 满足 $1/n \leq \epsilon \leq 1$。那么，对于每个满足 $|D| = n$的数据库 D， $\mathcal{M}$ 都不可能达到 $\mathrm{d_B}(\mathcal{P}_0(D)\ \mathcal{M}(D)) = o_p\left(\frac{1}{n\epsilon}\right)$ 。

5 模拟研究

在以下模拟研究和真实世界数据分析中，我们只考虑第 0 阶和第 1 阶持久图；也就是说，我们使用的效用通过取 $\ell = 1$ 给出，即我们设定 $u_D : \left(\mathrm{Pers}_M\right)^2 \to \mathbb{R}$ by

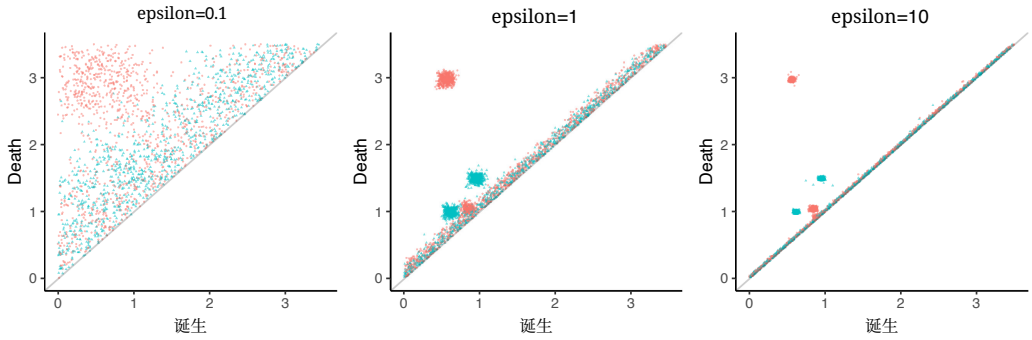
$$u_D(\mathcal{P}_0, \mathcal{P}_1) := u_D^{(0)}(\mathcal{P}_0) + u_D^{(1)}(\mathcal{P}_1).$$

作出此限制仅仅是为了简洁地介绍我们的算法；该算法可以轻松扩展，以纳入更高维度的特征。

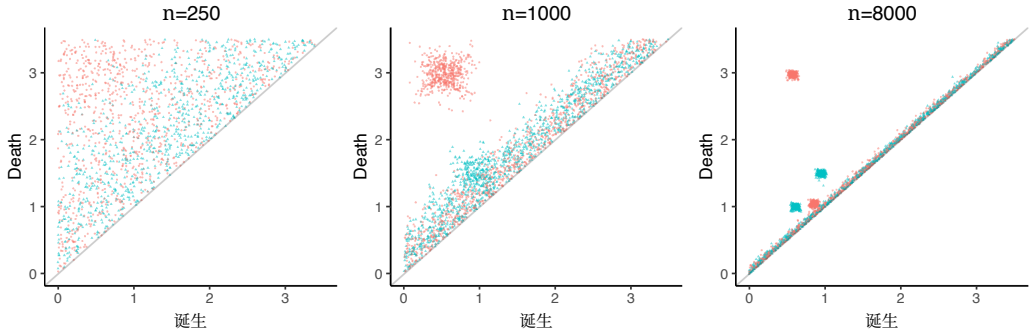
我们生成差分隐私持久图，并研究关键超参数的影响：隐私预算 ϵ 和样本量 n，其中 DTM 的分辨率 m 设为 0.2。对于指数机制，默认参数为$\epsilon = 1$, $m = 0.2$, $n = 4000$，以及 $M = 5$。这些超参数是通过初步模拟选定的。为了从指数机制中抽样，我们使用马尔可夫链蒙特卡洛算法，具体见附录 C.1。我们选取 $T = 10000$个蒙特卡洛图中的最后一个迭代结果，作为用于分析并报告的隐私化图*。

该模拟基于图 3中的示例，其中有两个圆，圆心位于原点(1.5,1.5) 和 $(-1.5, -1.5)$，半径分别为 1.5 和 1。每个圆由 200

*代码 R 可在 https://github.com/jwsohn612/DPTDA获取。



(a) 不同 ϵ 的私有化持久图，其中 $n = 4000$。



(b) 不同 n 的私有化持久图，其中 $\epsilon = 1$。

图4：如第 5节所述，针对500个重复数据集的隐私化图： 0维连通分支（橙色）和 1维环（绿色）。

沿圆的边界均匀生成的点的观测值。在（i-2）和（ii-2）中，圆的中间还多了一个点。生成 Vietoris-Rips图时，最大过滤尺度设为 3。所有分析均基于 500 个抽样数据集，并且我们对每次重复都应用隐私机制。

图 4 展示了指数机制在改变 ϵ 和 n 时的输出。为体现图表的变化，我们考虑了 500 个重复数据集。我们的指数算法独立应用于每个数据集，且算法仅报告最终图。对全部 500 次重复试验重复这一过程后，我们获得了 500 个报告的私有图。图 4 中的每个面板均根据这 500 个私有图绘制，以展示私有图分布的形状。正如预期，随着 ϵ 或 n 增大，隐私化持久图的变异性会减小。

瓶颈距离方面的总体趋势如图5所示。请注意，x轴采用对数尺度表示。图中的灰色区域显示了非私有图与其私有图之间瓶颈距离的95%逐点分位数区间。图5表明，对数尺度下的 ϵ 和 n 与对数瓶颈距离之间近似呈线性关系，因为阴影区域较好地包含了红色虚线，例如，$\log \mathrm{d_B}(\mathcal{P}(D) \cdot) \approx -\log\, n + c$，其中 c为某个常数。这些结果从启发式角度˜支持了 $d_B((P)_{DP}, \mathcal{P}(D)) = O_p(1/(n\epsilon))$（将 ℓ 视为固定值）。

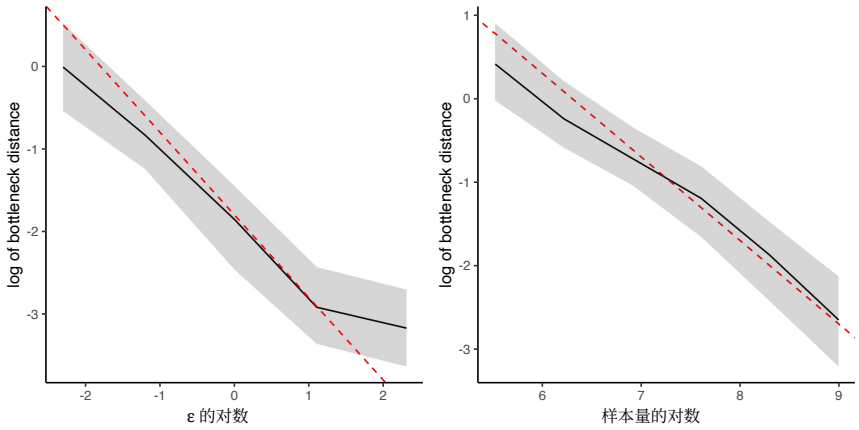


图5：在500次重复实验中，ϵ（左）和 n（右）分别增大时，$\mathrm{d_B}$ between $\mathcal{P}(D)$与对应隐私图的95%分位数区间。阴影区域中的红色虚线斜率均为-1。

6 真实数据分析

在本节中，我们将该机制应用于一个真实世界数据集[†]，该数据集记录了三个人在建筑物内行走的位置，数据由智能手机记录。我们将这些人称为 Walker A、B 和 C。每个人的位置3维坐标(x, y, z)随时间测量了20000 次，因此该数据集包含20000 个位置向量(x, y, z)，分别对应 Walker A、B 和 C。

我们计算与每个行走者对应的持久图，并应用我们的机制对其进行隐私保护。需要指出的是，我们关注的不是单个行走者的隐私，而是某个人的时间戳发生变化时的隐私。如果某个行走者的持久图因某一时间戳的位置变化而发生显著变化，则位置信息可能面临隐私泄露的风险。我们的隐私机制在保护与每个时间戳相关信息的同时，保留了每个行走者行程的拓扑结构。

为了获得隐私化图，我们在机制中执行 50000 次马尔可夫链蒙特卡洛迭代；将最后一次迭代得到的持久图作为报告的隐私化图。我们将采样空间大小设为 $M = 5$，L^1-DTM 的分辨率设为 $m = 0.05$，隐私预算设为 $\epsilon = 1$。

图7 展示了 L^1-DTM 对应 Walker C 的持久图与其隐私化图的比较结果。可以看出，这些图表非常相似。事实上，瓶颈距离 $\mathrm{d_B}(\mathcal{P}_0(D)\mathcal{P}_{0,\mathrm{DP}})$ 和 $\mathrm{d_B}(\mathcal{P}_1(D)\,\mathcal{P}_{1,\mathrm{DP}})$ 均为 0.01，这表明我们的机制具有很高的准确性。请注意，对角线附近的点不被视为显著点，也不会对瓶颈距离产生实质性影响。在图 7右侧的图中，我们看到瓶颈距离收敛到区域 $< .025$，且马尔可夫链似乎在经过 ≈ 5000 次迭代后收敛。

[†]数据可从以下地址获取
http://bertrand.michel.perso.math.cnrs.fr/Enseignements/TDA/Tuto-Part1.html

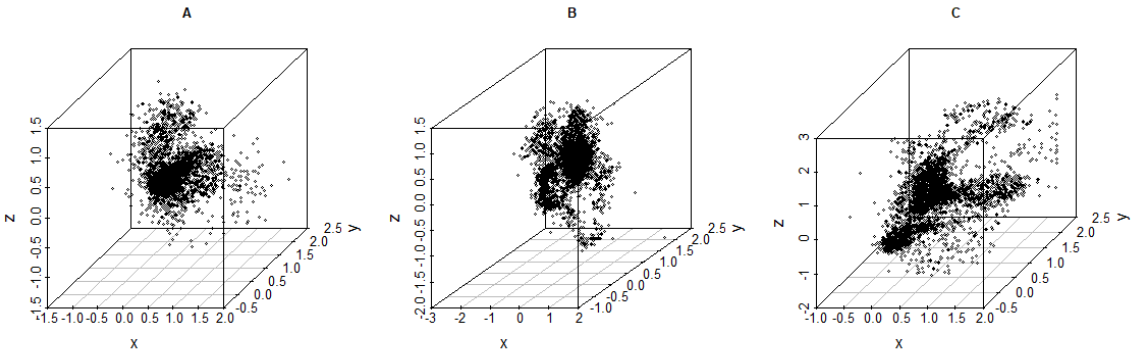


图 6：Walker A、B 和 C 的位置信息散点图。

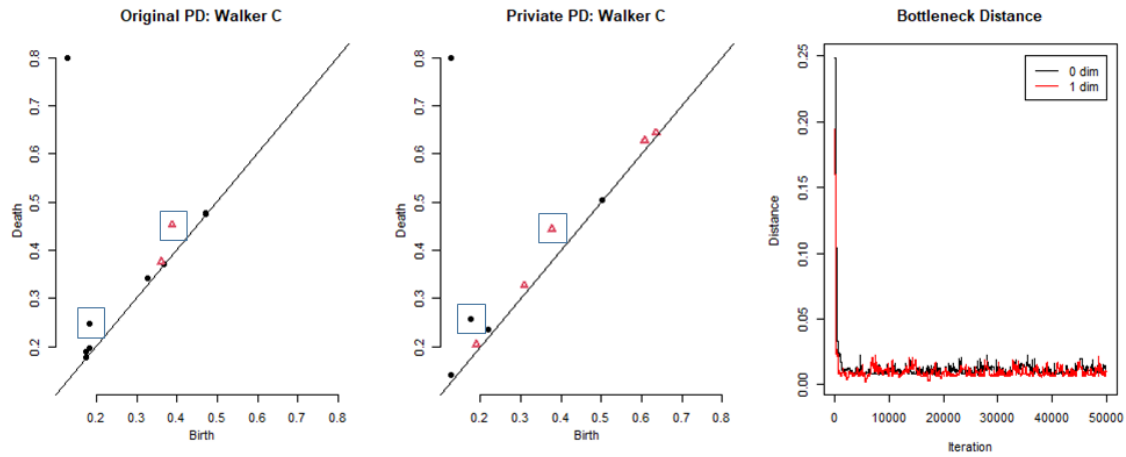


图 7：（左）Walker C 的真实持久图；（中）MCMC 过程最后一次迭代时 Walker C 的隐私化持久图；（右）真实图与隐私化图在 MCMC 迭代过程中的瓶颈距离 $\mathrm{d_B}(\mathcal{P}_0(D)\mathcal{P}_{0,\mathrm{DP}})$（黑色）和 $\mathrm{d_B}(\mathcal{P}_1(D)\mathcal{P}_{1,\mathrm{DP}})$（红色）。

7 讨论

本文提出了首个用于生成差分隐私持久图的机制，并强调了抗离群值鲁棒性在敏感性分析中的作用。尽管我们的研究成果对隐私保护下的拓扑数据分析作出了重要贡献，也加深了我们对鲁棒拓扑数据分析测度的一般认识，但本研究仍存在一些重要的不足之处，也有一些未来研究方向：

正如其他论文中所指出的（例如，Minami 等人，2016年；Ganesh 和 Talwar，2020年；Seeman 等人，2021年；Awan 和 Rao，2023年），指数机制的马尔可夫链蒙特卡洛实现可能会带来额外的隐私成本；在更严谨的研究中，应将其纳入分析。上述论文提出的技术可以应用于我们的指数机制实例，以确保在端到端隐私保护中考虑马尔可夫链蒙特卡洛近似所带来的影响。

本文建议将持久图中的分量数量 M 设为足够大的常数，或设为诸如 $M = \log n$ 这样的递增函数；此外，也可以将 M 视为一个未知量，并以隐私保护的方式进行学习。正如一位匿名审稿人所建议的，在这种情况下，或许可以开发一种可逆跳跃马尔可夫链蒙特卡洛算法，从指数机制中采样。这种方法面临的一些挑战包括：1）在无限维空间 $\prod_{M=1}^{\infty}(\widetilde{\text{Pers}}_{M,\,N})^{\ell+1}$ 上构建基准测度，以确保指数机制产生有效的概率分布；2）确定一种允许在不同维度之间转换的可逆跳跃规则；3）对这一新机制进行定制化的效用分析。我们将这一研究方向留待未来的研究者探索。

在拓扑数据分析方面，我们想指出，还可以讨论其他一些抗离群点拓扑数据分析方法，以用于隐私保护。例如，Chazal等人也曾讨论过的核距离（Chazal等人（2018年）可能是一个很好的候选方案。

除了拓扑数据分析之外，我们提出隐私机制的整体框架也可以应用于其他任何取值于度量空间的统计量。针对取值于度量空间的统计量，可以类似于我们对持久性图和瓶颈距离所采用的方式定义效用函数；这一点已在 Dwork 等人（2006年）中得到认识。然而，对于每个不同问题的效用分析，都需要对所涉及统计量和度量空间的具体结构与性质进行独特的理解。

我们方法的一个理论局限在于其随所考虑的维度数量（记为 $\ell+1$）而扩展。众所周知，ϵ-差分隐私机制中的误差通常随维度线性扩展，而我们的指数机制实例也不例外。由于这是 ϵ-差分隐私的局限，因此只能通过采用不同的隐私框架来解决。未来的研究者可以考虑在近似差分隐私（Dwork 和 Roth，2014年）、零集中差分隐私（Bun 和 Steinke，2016年）或高斯差分隐私（Dong 等人，2022年）框架下开发差分隐私拓扑数据分析方法；这些框架通常允许隐私噪声的幅度仅随维度的平方根扩展。

我们工作的另一个局限性在于，效用分析依赖于对持久图空间进行人为离散化。这一局限性是由采用命题3所导致的，并且有望通过使用更先进的技术加以解决。

最后，Dong 等人(2020年) 为指数机制提出了一种替代敏感度的量，这可能会改善有限样本性能。

致谢与经费披露

作者感谢匿名审稿人提出的宝贵意见，这些意见显著改善了本文的呈现效果。Taegyu Kang 的研究部分得到了普渡大学美国空军科学研究办公室（AFOSR）项目资助（FA9550-22-0238）的支持。Jordan Awan 的研究部分得到了美国国家科学基金会（NSF）项目资助（SES-2150615）的支持。

参考文献

Martin Abadi、Andy Chu、Ian Goodfellow、H Brendan McMahan、Ilya Mironov、Kunal Talwar 和 Li Zhang。采用差分隐私的深度学习。载于*2016*年*ACM SIGSAC*计算机与通信安全会议论文集，第308–318页，2016年。

Hirokazu Anai、Frédéric Chazal、Marc Glisse、Yuichi Ike、Hiroya Inakoshi、Raphaël Tinar- rage 和 Yuhei Umeda。基于DTM的过滤。载于拓扑数据分析，第15卷，第33–66页。施普林格，2020年。

Marco Avella-Medina。隐私保护参数推断：稳健统计的一个案例。美国统计学会会刊，116(534)：969–983，2021年。

Jordan Awan和Vinayak Rao。隐私感知拒绝采样。《机器学习研究杂志》，24(74):1–32，2023年。

Jordan Awan和Yue Wang。差分隐私Kolmogorov-Smirnov型检验。arXiv:2208.06236，2022年。

Jordan Awan、Ana Kenney、Matthew Reimherr和Aleksandra Slavković。指数机制的优势与陷阱：在希尔伯特空间和函数型主成分分析中的应用。发表于国际机器学习会议，第374–384页。PMLR，2019年。

Leo Betthauser、Urszula Chajewska、Maurice Diesendruck和Rohith Pesala。利用潜在空间表示发现分布偏移。arXiv:2202.02339，2022年。

Omer Bobrowski和Robert J. Adler。距离函数、临界点与随机Čech复形的拓扑。同调、同伦与应用，16(2):311–344，2014年。

Glen E. Bredon。《拓扑与几何》。施普林格出版社，纽约，1997年。

Mark Bun 和 Thomas Steinke。集中差分隐私：简化、扩展与下界。载于《密码学理论会议》，第 635–658 页。施普林格，2016年。

Luigi Caputi、Anna Pidnebesna 和 Jaroslav Hlinka。拓扑数据分析用于脑连接分析的机遇与陷阱。《NeuroImage》，238:118245，2021年。

Gunnar Carlsson。拓扑与数据。《美国数学学会通报》，46 (2):255–308，2009年。

Frédéric Chazal、David Cohen-Steiner 和 Quentin Mérigot。概率测度的几何推断。《计算数学基础》，11:733–751，2011年。

Frédéric Chazal、Vin de Silva、Marc Glisse 和 Steve Oudot。《持久模的结构与稳定性》。施普林格，2016a年。Frédéric Chazal、Pascal Massart 和 Bertrand Michel。稳健几何推断的收敛速率。《电子统计学杂志》，10：2243–2286，2016b。Frédéric Chazal、Brittany Fasy、Fabrizio Lecci、Bertrand Michel、Alessandro Rinaldo 和 Larry Wasserman。稳健拓扑推断：到测度的距离与核距离。《机器学习研究杂志》，18：1–40，2018年。David Cohen-Steiner、Herbert Edelsbrunner 和 John Harer。持久图的稳定性。《离散与计算几何》，37：103–120，2007年。Rachel Cummings、Varun Gupta、Dhamma Kimpara 和 Jamie Morgenstern。隐私与公平性的兼容性问题。载于《第27届用户建模、适应与个性化会议附属出版物》，第309–315页，2019年。Meryll Dindin、Yuhei Umeda 和 Frederic Chazal。通过模块化神经网络进行心律失常检测的拓扑数据分析。载于《人工智能进展：第33届加拿大人工智能会议，加拿大人工智能会议2020，渥太华，加拿大安大略省，2020年5月13日至15日，会议论文集33》，第177–188页。施普林格，2020年。Jinshuo Dong、David Durfee 和 Ryan Rogers。指数机制的最优差分隐私组合。载于《国际机器学习会议》，第2597–2606页。PMLR，2020年。Jinshuo Dong、Aaron Roth 和 Weijie J Su。高斯差分隐私。《皇家统计学会杂志 B辑》，84(1)：3–37，2022年。Cynthia Dwork 和 Jing Lei。差分隐私与稳健统计。载于《STOC'09——2009年ACM计算理论国际研讨会论文集》，第371–380页，2009年。Cynthia Dwork 和 Aaron Roth。《差分隐私的算法基础》。Now Publishers Inc.，2014年。Cynthia Dwork、Frank McSherry、Kobbi Nissim 和 Adam Smith。在隐私数据分析中根据敏感度校准噪声。载于《密码学理论会议》，第265–284页。施普林格，2006年。Herbert Edelsbrunner 和 John Harer。持久同调——一项调查。载于《离散与计算几何调查》，《当代数学》。美国数学学会，2008年。Herbert Edelsbrunner 和 John Harer。《计算拓扑：导论》。美国数学学会，2009年。Nicolas Fournier 和 Arnaud Guillin。经验测度在 Wasserstein 距离下的收敛速率。《概率论及相关领域》，162(4)：707–738，2015年。

Arun Ganesh 和 Kunal Talwar。通过对离散化朗之万马尔可夫链蒙特卡罗进行 Rényi 散度分析，加速差分隐私采样器。神经信息处理系统进展，33:7222–7233，2020年。Rob Hall、Alessandro Rinaldo 和 Larry Wasserman。函数和函数型数据的差分隐私。机器学习研究杂志，14(1):703–727，2013年。Jonathan Hehir、Siddharth Vishwanath、Aleksandra Slavković 和 Xiaoyue Niu。局部差分隐私下随机图相关问题，2022年。论文发表于 JMM 2022。Felix Hensel、Michael Moor 和 Bastian Rieck。拓扑机器学习方法综述。人工智能前沿，4:52，2021年。Christoph Hofer、Roland Kwitt、Marc Niethammer 和 Mandar Dixit。通过持久同调实现连通性优化的表征学习。载于国际机器学习会议，第2751–2760页。PMLR，2019年。Zhiyi Huang 和 Sampath Kannan。社会福利的指数机制：隐私保护、真实且近乎最优。载于*2012*年*IEEE*第*53*届计算机科学基础年度研讨会，第140–149页。IEEE，2012年。Vishesh Karwa 和 Aleksandra Slavković。利用噪声度数进行推断：差分隐私 β-模型与合成图。统计年刊，44(1):87–112，2016年。Vishesh Karwa、Pavel N Krivitsky 和 Aleksandra B Slavković。共享社交网络数据：指数族随机图模型的差分隐私估计。英国皇家统计学会杂志*C*辑（应用统计），第481–500页，2017年。Firas A Khasawneh 和 Elizabeth Munch。利用持久同调检测车削中的颤振。机械系统与信号处理，70:527–541，2016年。Hyekyoung Lee、Moo K Chung、Hyejin Kang、Bung-Nyun Kim 和 Dong Soo Lee。脑网络的判别性持久同调。载于*2011*年*IEEE*生物医学成像国际研讨会：从纳米到宏观，第841–844页，2011年。Ephy R. Love、Benjamin Filippenko、Vasileios Maroulas 和 Gunnar Carlsson。用于深度学习的拓扑卷积层。《机器学习研究杂志》，24:1–35，2023年。Zhigang Lu、Hassan Jameel Asghar、Mohamed Ali Kaafar、Darren Webb 和 Peter Dickinson。基于凸化损失函数的深度学习差分隐私框架。*IEEE*信息取证与安全汇刊，17:2151–2165，2022年。Melissa R. McGuirl、Alexandria Volkening 和 Björn Sandstede。斑马鱼图案的拓扑数据分析。美国国家科学院院刊，117(10):5113–5124，2020年。Frank McSherry 和 Kunal Talwar。通过差分隐私实现机制设计。载于第*48*届*IEEE*计算机科学基础年度研讨会（*FOCS'07*），第94–103页，2007年。

Kentaro Minami、HItomi Arai、Issei Sato 和 Hiroshi Nakagawa。无需敏感度的差分隐私。神经信息处理系统进展，29，2016年。Ardalan Mirshani、Matthew Reimherr 和 Aleksandra Slavković。基于高斯扰动的函数型数据形式化隐私。载于国际机器学习会议，第4595–4604页。PMLR，2019年。James R. Munkres。代数拓扑基础。Addison-Wesley Publishing Company，1984年。Monica Nicolau、Arnold J. Levine 和 Gunnar Carlsson。基于拓扑的数据分析发现了一组具有独特突变特征且生存率极佳的乳腺癌亚组。美国国家科学院院刊，108(17):7265–7270，2011年。Partha Niyogi、Stephen Smale 和 Shumuel Weinberger。对噪声数据无监督学习的拓扑视角。*SIAM*计算杂志，40(3):646–663，2011年。Steve Y. Oudot。持久理论：从箭图表示到数据分析。美国数学学会，2015年。Takashi Owada。稀疏情形下持久图的收敛性。应用概率年刊，32(6):4706–4736，2022年。Matthew Reimherr 和 Jordan Awan。Kng： k-范数梯度机制。神经信息处理系统进展，32，2019年。Bastian Rieck、Tristan Yates、Christian Bock、Karsten Borgwardt、Guy Wolf、Nicholas Turk-Browne 和 Smita Krishnaswamy。利用立方持久性揭示时变 fMRI 数据的拓扑结构。神经信息处理系统进展，33:6900–6912，2020年。Jeremy Seeman、Matthew Reimherr 和 Aleksandra Slavković。采用人工原子的指数机制马尔可夫链实现的精确隐私保证。神经信息处理系统进展，34:13125–13136，2021年。Daniel Shnier、Mircea Voineagu 和 Irina Voineagu。自闭症脑转录组数据的持久同调分析。英国皇家学会界面杂志，16，2019年。Reza Shokri 和 Vitaly Shmatikov。隐私保护深度学习。载于第*22*届*ACM SIGSAC*计算机与通信安全会议论文集，第1310–1321页，2015年。Larry Wasserman。拓扑数据分析。统计及其应用年度评论，5(1):501–532，2018年。Larry Wasserman 和 Shuheng Zhou。差分隐私的统计框架。美国统计学会会刊，105(489):375–389，2010年。doi: 10.1198/jasa.2009.tm08651。Xiaoqi Xu、Nicolas Drougard 和 Raphaëlle N. Roy。拓扑数据分析作为 EEG 处理的新工具。神经科学前沿，15，2021年。

Xu Xin、Jessi Cisewski-Kehe、Sheridan B. Green 和 Daisuke Nagai。利用拓扑数据分析寻找宇宙空洞和丝状体环。*Astronomy and Computing*，27:34–52，2019年。

Afra Zomorodian 和 Gunnar Carlsson。计算持久同调。离散计算几何，33:249–274，2005年。

附录A。关于持久同调的更多内容

本部分旨在更详细地介绍如何构造持久同调及其对应的持久图。我们从单纯复形和单纯同调的定义开始，然后介绍如何构造持久同调。

A.1 单纯同调

我们先从单纯复形的定义开始。本小节的大部分内容基于 Munkres (1984年)。

定义 19（单纯复形）（抽象）单纯复形是由有限非空集合构成的一个集族K，并且如果 σ 是 K的一个元素，那么 σ的每个非空子集也都是其元素。

单纯复形 K 的每个元素 σ 都称为 K的一个单形。单形 σ 的维度定义为 $|\sigma| - 1$，即 σ 中的元素个数减一。当σ 是一个 q维单形时，我们简称 σ 为 q单形。单纯复形 K 的维度 $\dim K$ 定义为 K中单形的最大维度，即

$$\dim K := \max_{\sigma \in K} \dim \sigma.$$

如果集合 $\{\dim \sigma : \sigma \in K\}$ 无界，则令 $\dim K = \infty$。 σ 的每个非空子集都称为 σ的一个面。

设 K 为一个单纯复形。对于 K 中的每个单形 $\sigma = \{v_0, \ldots, v_q\}$，可以考虑 σ 中顶点的有序元组。具体而言，对于 $\{0, \ldots, q\}$ 上的每个置换 α，都存在一个有序元组 $(v_{\alpha(0)}, \ldots, v_{\alpha(1)})$。这样的有序元组称为 σ 的一个有序单纯形。 K 中每个单形的所有有序单纯形构成的集合称为 K 的有序单纯复形，记为 K_{ord}。

设 K_{ord} 为单纯复形 K的一个有序单纯复形。设 $v^\alpha = (v_{\alpha(0)}, \ldots, v_{\alpha(q)})$ 和 $v^\beta = (v_{\beta(0)}, \ldots, v_{\beta(q)})$ 是公共 q-单形 $\sigma = \{v_0, \ldots, v_q\}$ 的两个有序 q-单形。当且仅当 α 和 β 具有相同的符号时，声明 $v^\alpha \sim v^\beta$，即 α 与 β 仅相差偶数次对换。注意，这个关系在每个单形 σ 的有序单形集合上定义了一个等价关系。用 $[v_0, \ldots, v_q]$ 表示有序单形 $(v_0, \ldots, v_q)$ 的等价类。这样的等价类称为一个定向 q-单形。也就是说，每个带有 $q \geq 1$ 的 q- 单形都会诱导出两个定向 q- 单形。用 K_{ori} 表示 K 中每个单形的所有定向单形组成的集合。在不引起混淆的情况下，我们使用符号 σ 同时表示一个单形和一个定向单形。

对于每个自然数 $q \geq 0$，令 $K^q{}_{\text{ori}}$ 为 K 的所有定向 q-单形组成的集合。定义 $C_q(K)$ 为所有满足以下条件的函数 $c\colon K^q{}_{\text{ori}} \to \mathbb{Z}$ 的集合。

- $c(\sigma) = -c(\sigma')$，如果 σ 和 σ' 是同一个单形的相反定向。
- $c(\sigma) = 0$ 除有限多个定向的 q-单形 σ外均成立，即每个 c 都是有限支撑的。

通过将群运算定义为逐元素加法，可以在 $C_q(K)$ 上赋予群结构。注意，在该群结构下，$C_q(K)$ 是一个阿贝尔群。此外，显然 $C_q(K)$ 是一个自由阿贝尔群，其基可以通过为每个单形 σ 恰好选择一个定向单形来构造。$C_q(K)$ 中的每个元素c 都可以表示为 K 中定向 q-单形的有限 $\mathbb{Z}$-线性组合，即，每个 c 都可以写成

$$c = \sum_{i=1}^{k} n_i \sigma_i,$$

其中 k 是有限的，$n_i \in \mathbb{Z}$ 和 $\sigma_i \in K^q{}_{\text{ori}}$ 对所有 $1 \leq i \leq k$ 均成立。每个函数 c 称为 K 的一个 q-链，而 $C_q(K)$ 称为 K 的定向 q-链群。我们规定，当 $q < 0$ 或 $q > \dim K$ 时，$C_q(K) = 0$。

现在，我们来定义定向链复形的边界算子。

定义 20（边界算子）设 K 为一个单纯复形。对于每个整数q，定义

$$\partial_q : K_{\text{ori},q} \to C_{q-1}(K)$$

通过赋值

$$\partial_q : [v_0, \ldots, v_q] \mapsto \sum_{i=0}^{q} (-1)^i [v_9, \ldots, \hat{v}_i, \ldots, v_q],$$

其中，$[v_0, \ldots, \hat{v}_i, \ldots, v_q]$ 是从 $[v_0, \ldots, v_q]$中删除 v_i 后得到的 $q-1$) 定向单形。由于 $C_q(K)$ 是自由阿贝尔群，映射 ∂_q 可唯一地延拓为群同态 $\partial_q\colon C_q(K) \to C_{q-1}(K)$。这个同态称为边界算子。

边界算子的关键性质如下：

$$\partial_{q-1} \circ \partial_q = 0 \quad \text{for every } q.$$

换句话说，阿贝尔群与群同态的序列 $(C_q(K), \partial_q)_{q \in \mathbb{Z}}$ 构成一个链复形。该性质也可以表述如下。

$$\operatorname{Im} \partial_{q-1} \subseteq \operatorname{Ker} \partial_q \quad \text{for every } q,$$

其中，Ker 和 Im 分别表示同态的核与像。由于定向链复形的群序列构成一个链复形，因此可以定义它的同调群。此外，核 $\operatorname{Ker} \partial_q$ 通常写作 $Z_q(K)$，其每个元素称为 q-圈；而像 $\operatorname{Im} \partial_{q-1}$ 通常写作 $B_q(K)$，其每个元素称为 q- 边界。

定义 21（单纯同调）设 K 为一个单纯复形。对于每个整数 q，第 q 个单纯形同调群定义为以下商群：

$$Z_q(K)/B_q(K) = \frac{\mathrm{Ker}(\partial_q : C_q(K) \to C_{q-1}(K))}{\mathrm{Im}(\partial_{q-1} : C_q(K) \to C_q(K))},$$

并记作 $H_q(K)$

备注 *22* 不必将 $C_q(K)$构造成阿贝尔群，也可以考虑定义在相同基上的自由R-模，其中 R 是一个交换环。边界算子可以用相同方式定义，现在它可以唯一地延拓为一个 R-模同态 $\partial_q : C_q(K) \to C_{q-1}(K)$。由此得到的 R-模与 R-模同态构成的序列 $(C_q(K)\ \partial_q)$ 是一个 R-模链复形，因此可以用相同方式定义其单纯同调；在这种情况下，每个同调群 $H_q(K)$也成为一个 R-模。在这种情况下，我们将 K 的第 q阶单纯形同调模记为 $H_q(K; R)$，并称其为 K 以 R为系数的第 q阶单纯同调。

A.2 持久同调

设 $\{K_r\}_{r\in\mathbb{R}}$ 是一族满足 $K_{r_1} \subseteq K_{r_2}$ （当 $r_1 \le r_2$ 时）的单纯复形。这样的集合称为单纯复形的滤过（参数为 $\mathbb{R}$）。对于滤过中的每个单纯复形 K_r，都可以构造链复形 $(C_q(K_r)\ \partial_q)_{q\in\mathbb{Z}}$ 及相应的同调群 $(H_q(K_r))_{q\in\mathbb{Z}}$。此外，每个包含映射 $\iota_{r_1,r_2} : K_{r_1} \to K_{r_2}$ ($r_1 \le r_2$) 都诱导出一个群同态 $C_q(K_{r_1}) \to C_q(K_{r_2})$，事实上，对于每个整数 q，它就是包含映射 $C_q(K_{r_1}) \hookrightarrow C_q(K_{r_2})$；并且，所有这些同态（包含映射）都与边界算子交换，即每个包含映射都会在定向链的链复形之间诱导出一个链映射。因此，对于每个 q，每个包含 ι_{r_1,r_2} 都会在同调群之间诱导出一个同态 $\iota^q_{r_1,r_2} : H_q(K_{r_1}) \to H_q(K_{r_2})$。这就得到一族 $\{H_q(K_r)\}_{r\in\mathbb{R}}$ 单纯同调群，并且对于每个 q 以及每一对 $r_1 \le r_2$，都伴随有一个群同态 $\iota^q_{r_1,r_2} : H_q(K_{r_1}) \to H_q(K_{r_2})$。

对于每一对 $r_1 \le r_2$ 和每个 q，$\iota^q_{r_1,\ r_2} : H_q(K_{r_1}) \to H_q(K_{r_2})$ 的像，记作 $\mathrm{Im}\iota^q_{r_1,\ r_2}$，称为从 r_1 到 r_2持续的第 q 个持续同调群。群 $\mathrm{Im}\iota^q_{r_1,\ r_2}$ 的秩称为从 r_1 到 r_2 持续的第 q 个持续贝蒂数，记作$\beta^q_{r_1,\ r_2}$。直观来说，贝蒂数 $\beta^q_{r_1,\ r_2}$ 表示在过滤中于参数 r_1 之前诞生、且直到参数r_2 时尚未消亡的独立q-环的数量。此外，对于过滤中的每个 q-环，可以分别考察该环首次出现（诞生）和消失（消亡）时的参数值。

设 σ 是一个 q-循环，它在某个时刻出现在过滤中，即 σ 是$\mathrm{Ker}\,\partial_q(K_r)$ 对某个 r的一个元素。于是，可以考虑它的诞生和消亡时间（参数值）。将过滤中所有 q-循环的所有出生-死亡对汇集起来，可以形成一个由(b, d) 形式的点组成的 $b \le d \le \infty$多重集。该多重集称为过滤的第 q阶持久图。持久图的形式化构造涉及主理想整环上有限生成分次模的结构定理，这是抽象代数中的一个定理。关于更形式化且全面的讨论，请参阅Carlsson(2009年)以及Edelsbrunner 和 Harer(2008年)。

A.3 构造单纯复形过滤的几种方法

现在，我们介绍几种构造单纯复形过滤的方法，这些方法在本文的主要讨论中发挥作用。本小节的内容可见于Edelsbrunner 和 Harer(2009年）。

令 $D = \{x_1, \ldots, x_n\}$ 为度量空间（$\mathcal{X}$，d）的一个有限子集。对于每个非负实数 $r \geq 0$，考虑以每个$i \in \{1, \ldots, n\}$为中心的球 $B(x_i; r) := \{y \in \mathcal{X}: d(y, x_i) < r\}$。定义以 D 为基础、半径为 r 的Čech复形 $\check{C}(D; r)$如下：当且仅当 $\cap_{j=0}^{q} B(x_{i_j}; r) \neq \emptyset$时，$D$的子集 $\sigma = \{x_{i_0}, \ldots, x_{i_q}\}$ 属于 $\check{C}(D; r)$。注意，对于每一对 $r_1 \leq r_2$，$\check{C}(D; r_1) \subseteq \check{C}(D; r_2)$。因此，Čech复形的集合 $\{\check{C}(D; r)\}_{r \geq 0}$ 构成单纯复形的滤过。

Čech复形有几种变体。其中一种变体是Vietoris-Rips复形。在 D 上、半径为 r 的Vietoris-Rips复形VR(D; r)定义如下。子集 $\sigma = \{x_{i_0}, \ldots, x_{i_q}\}$ D D r B $x_{i_{j_k}}$ r $\cap B$ $x_{i_{j_l}}$ r $\neq$属于 $\emptyset$ k $l \in \{0, \ldots, q\}$ B x_{i_0} r $\ldots B$ x_{i_q} r VR(;)当且仅当对于每个,, (;) (;)，也就是说，球(;) (;)两两相交。显然，只要 $r_1 \leq r_2$，就有VR(D; r_1) $\subseteq$ VR(D; r_2)。因此，Vietoris-Rips复形的集合 $\{\text{VR}(D; r)\}_{r \geq 0}$ 构成了单纯复形的滤过。下面Čech复形与Vietoris-Rips复形之间的关系表明，在欧几里得空间中的有限子集上，Čech复形的过滤与Vietoris-Rips复形的过滤所包含的信息本质上相同。

命题23 设 $D = \{x_1, \ldots, x_n\}$ 是欧几里得空间的一个有限子集，并在其上赋予由 ℓ^2-范数诱导的度量。那么，对于任意 $r \geq 0$，以下结论成立。

$$\check{C}(D; r) \subseteq \text{VR}(D; r) \subseteq \check{C}(D; \sqrt{2}r).$$

最后一种构造方式来自定义在度量空间上的实值函数。设（$\mathcal{X}$，d）是一个可三角剖分的度量空间，且 $f: \mathcal{X} \to \mathbb{R}$ 是一个实值连续函数。对于每个 $r \in \mathbb{R}$，考虑子水平集 $L_r := f^{-1}((-\infty, r])$，它是$\mathcal{X}$的一个子集。注意，当 $r_1 \leq r_2$时，$L_{r_1} \subseteq L_{r_2}$。此外，由于 $\mathcal{X}$ 是可三角剖分的，因此可以在保持包含关系的同时，对所有子水平集进行三角剖分。因此，对子水平集族 $\{L_r\}_{r \in \mathbb{R}}$ 进行这样的三角剖分所得到的集合，构成了一个单纯复形的滤过。

在结束本节之前，我们介绍连续函数$f: \mathcal{X} \to \mathbb{R}$ 的一个特定条件，以确保 f 的变化不会过于剧烈。

定义 24（驯服函数） 设 $(\mathcal{X}, d)$ 是一个可三角剖分度量空间，且 $f: \mathcal{X} \to \mathbb{R}$ 是一个实值连续函数。令 X_r 为子水平集$f^{-1}((-\infty, r])$的三角剖分。设 $\iota_{r_1, r_2}^{q}: H_q(X_{r_1}) \to H_q(X_{r_2})$ 是由包含映射 $\iota_{r_1, r_2}: X_{r_1} \to X_{r_2}$ 对每一对 $r_1 \leq r_2$诱导的群同态。若不存在正数 $\epsilon > 0$ 使得对于每个维度 q，$\iota_{r-\epsilon, r+\epsilon}^{q}$ 都是同构，则称 $r \in \mathbb{R}$ 为同调临界值。若 f 产生的同调临界值只有有限多个，且其所有子水平集的所有同调群均为有限秩，则称函数 f 为驯服函数。

附录B. 主要结果的证明

在本部分中，我们将给出第 3节和第4节中各项定理的详细证明。在本节中，除非另有说明，符号 $\|\cdot\|$ 表示 ℓ^2-范数，其所在的欧几里得空间由各项证明中所讨论的数据点所在的空间决定。

B.1 第3节中结果的证明

证明 [引理 4] 的证明固定 $r > 0$，使得 $d_{m-1} < r < d_m$，并令 $G(D; r)$ 为顶点集为 $\mathcal{X}$、连接阈值为 r 的几何图；也就是说，D 是$G(D; r)$的顶点集，并且当 $d(x_i, x_j) \leq r$ 时，任意一对顶点 $\{x_i, x_j\}$ 构成其一条边。由于 $(0, d_m)$ 是图 $\mathcal{P}_0^{\text{Cech}}(D)$ 的一个元素，$G(D; r)$ 中至少存在两个连通分支 $\mathcal{Y}_1$ 和$\mathcal{Y}_2$，满足

$$\min_{x_1 \in \mathcal{Y}_1, x_2 \in \mathcal{Y}_2} d(x_1, x_2) = 2d_m. \tag{A.1.1}$$

令 $\mathcal{Y}_1$ 和 $\mathcal{Y}_2$ 为这样的连通分支，并令 $x_1 \in \mathcal{Y}_1$ 和 $x_2 \in \mathcal{Y}_2$ 为取得最小值的点，即 $d(x_1, x_2) = d_m$。令 D' 为在 x_1 和 x_2的中点处添加一个点（例如 z）后得到的集合。显然，$\mathcal{Y}_1$（或等价地，$\mathcal{Y}_2$）的消亡时间减半。请注意，在 D 的过滤中，添加点 z不会使其他连通分支的消亡时间变得更长。因此，我们可以写为

$$\mathcal{P}_0^{\check{\text{Cech}}}(D') = \{(0, d_1'), \ldots, (0, d_t'), (0, \infty)\} \cup \{(0, d_m/2)\}$$

与 $d_j' \leq d_{m-1}$对于所有 $j = 1$、.....、t。这里，元素 $(0, d_m/2)$ 的重数至少为 2. 要计算 $\mathcal{P}_0^{\text{Cech}}(D)$ 与 $\mathcal{P}_0^{\text{Cech}}(D')$ 之间的瓶颈距离，我们必须考察 $\mathcal{P}_0^{\text{Cech}}(D)$ 与 $\mathcal{P}_0^{\text{Cech}}(D')$ 之间所有可能的双射。所有这些双射可分为三类。第一，$(0, d_m) \in \mathcal{P}_0^{\text{Cech}}(D)$ 与某个 $j \in \{1, \ldots, t\}$ 对应的元素 $(0, d_j') \in \mathcal{P}$ Cech0 (D') 相关联。第二，$(0, d_m) \in \mathcal{P}$ Cech0 (D) 与 $(0, d_m/2) \in \mathcal{P}_0^{\text{Cech}}(D')$ 相关联。第三，$(0, d_m) \in \mathcal{P}_0^{\text{Cech}}(D)$ 与对角线上的一个点相关联。在第一种情况下，关于 $(0, d_m)$ 的可能最小距离不可能小于 δ。在第二种情况下，$(0, d_m)$ 与 $(0, d_m/2)$ 之间的距离显然为 $d_m/2$。在最后一种情况下，$(0, d_m)$ 与对角线之间的距离为 $d_m/\sqrt{2}$。由于瓶颈距离定义为对所有此类双射取最小值，因此所需结论得证。

■

证明 [定理 5] 的证明 n 设 a 为偶数。令 b 和 E 为集合 $|a - b| =$ 中的两个点，且满足 Ediam D，并令 $n/2$ 由 a 个 $n/2$ 的副本和 b个 D' 的副本组成。令 a由将其中一个 a 移动到 b和 c的中点而得到，即 。那么，显然有

$$\mathcal{P}_0^{\check{\text{Cech}}}(D) = \{(0, \text{diam}E/2), (0, \infty)\}$$

and

$$\mathcal{P}_0^{\check{\text{Cech}}}(D') = \{(0, \text{diam}E/4), (0, \text{diam}E/4), (0, \infty)\}.$$

这证明了这两个图表之间的瓶颈距离以下界 $\mathrm{diam}E/4$为界，从而得到所需结果。当 n为奇数时，可以令 D包含 $(n-1)/2$个 a 的副本和 $(n+1)/2$ 个 b的副本，结果不变。

至于第二个结果，可以通过应用反三角不等式建立所提出的上界。为了建立下界，注意到对于任意一对集合 D和 D'，

$$\begin{aligned}\sup_{\mathcal{P}}|v_D(\mathcal{P})-v_{D'}(\mathcal{P})| &\geq \left|v_D\left(\mathcal{P}_0^{\check{\mathrm{C}}\mathrm{ech}}(D)\right)-v_{D'}\left(\mathcal{P}_0^{\check{\mathrm{C}}\mathrm{ech}}(D)\right)\right| \\ &= \mathrm{d_B}\left(\mathcal{P}_0^{\check{\mathrm{C}}\mathrm{ech}}(D'),\mathcal{P}_0^{\check{\mathrm{C}}\mathrm{ech}}(D)\right)\end{aligned}$$

根据第一个结果，对所有相邻对 D 和 D' 取最后一个表达式的上确界，其下界为 $\mathrm{diam}E/4$ 。证明完毕。 ■

证明 [定理的证明 8] 令 $\delta_D^{(p)}$ 和 $\delta_{D'}^{(p)}$ 为 L^p-DTM，分别对应于经验分布D 和 D'。根据 DTM 的稳定性定理（2.2）和 Wasserstein 稳定性（3.1，我们有

$$\mathrm{d_B}\left(\mathcal{P}_q^{\mathrm{DTM}_p}(D),\mathcal{P}_q^{\mathrm{DTM}_p}(D')\right)\leq\left\|\delta_D^{(p)}-\delta_{D'}^{(p)}\right\|\infty\leq\frac{1}{m^{1/p}}W_p\left(\hat{\mu}_D,\hat{\mu}_{D'}\right), \tag{A.1.2}$$

其中 $\hat{}\mu_D$ 和 $\hat{}\mu_{D'}$ 分别表示 D 和 D'上的经验分布。我们将建立不等式（A.1.2）右侧的一个上界。

假设 $H(D,D')=1$。令 x 为属于 D 但不属于 D' 的元素，令 z 为属于 D' 但不属于 D 的元素。令 π 为 $\hat{}\mu_D$ 和 $\hat{}\mu_{D'}$ 的耦合，定义如下：对于每个 $y\in D$，设

$$\pi(y,y)=\frac{1}{n},$$

and

$$\pi(x,z)=\frac{1}{n}.$$

不难验证， π 确实是$\hat{}\mu_D$ 和$\hat{}\mu_{D'}$的一个耦合。由此 π 我们有

$$\int_{(z_1,z_2)\in\mathbb{R}^d\times\mathbb{R}^d}\|z_1-z_2\|^p\,d\pi(z_1,z_2)=\|x-z\|^p\,\frac{1}{n}\leq\left(\mathrm{diam}E\right)^p\frac{1}{n}$$

根据 Wasserstein 距离 W_p的定义，我们有

$$\begin{aligned}W_p(\hat{\mu}_D,\hat{\mu}_{D'})^p &= \inf_{\nu}\int_{(z_1,z_2)\in\mathbb{R}^d\times\mathbb{R}^d}\|z_1-z_2\|^p\,d\nu(z_1,z_2) \\ &\leq \int_{(z_1,z_2)\in\mathbb{R}^d\times\mathbb{R}^d}\|z_1-z_2\|^p\,d\pi(z_1,z_2)\end{aligned}$$

where ν 遍历 P_n 和 P_n' 的所有耦合。因此，我们得到如下结果：

$$W_p(\hat{\mu}_D,\hat{\mu}_{D'})\leq\frac{\mathrm{diam}E}{n^{1/p}},$$

这意味着所需的结果。 ■

证明 [命题 9] 的证明 设 D 为一个数据集，其点分别以 50% 和 50% 的比例分布在两端 a、b 处，分别为 E。更具体地说，$\| a-b \|=$直径为 E，且集合D 包含 $2/n$ 个 a 的副本和 $2/n$ 个 b的副本。设 c 为 a 与 b的中点；即$\| a-c \|=\| b-c \|=$ 直径为 $E/2$。通过将 D 中的一个 a 移动到 c，构造 D'；也就是说，D' 包含$n/2-1$ 个 a的副本、$n/2$ 个 b的副本以及一个 c。设 δ_D 为在 D 上的经验分布、分辨率为 m 的 L^1-DTM，$\delta_{D'}$ 同理。于是，我们有

$$\delta_D(x)=\begin{cases}0 & \text{if } x=a,\\ 0 & \text{if } x=b,\\ \operatorname{diam} E/2 & \text{if } x=c.\end{cases}$$

另一方面，

$$\delta_{D'}(x)=\begin{cases}0 & \text{if } x=a,\\ 0 & \text{if } x=b,\\ \frac{k-1}{k}\frac{\operatorname{diam} E}{2} & \text{if } x=c.\end{cases}$$

回顾一下 $k=\lceil mn \rceil$。注意，线段 ab 上的任意点 x 都满足 $\delta_D(x) \leq \delta_D(c)$，且 $\delta_{D'}(x) \leq \delta_{D'}(c)$。因此，$D$ 的零维持久图 $\mathcal{P}_0^{\mathrm{DTM}_1}(D)$如下所示：

$$\mathcal{P}_0^{\mathrm{DTM}_1}(D)=\{(0,\operatorname{diam} E/2),(0,\infty)\}.$$

类似地，$\mathcal{P}_0^{\mathrm{DTM}_1}(D')$如下所示：

$$\mathcal{P}_0^{\mathrm{DTM}_1}(D')=\left\{(0,(k-1)\operatorname{diam} E/(2k)),(0,\infty)\right\}.$$

上述两个图表之间的瓶颈距离计算如下：

$$\mathrm{d_B}(\mathcal{P}_0^{\mathrm{DTM}_1}(D),\mathcal{P}_0^{\mathrm{DTM}_1}(D'))=\frac{\operatorname{diam} E}{2}-\left(\frac{k-1}{k}\right)\frac{\operatorname{diam} E}{2}=\frac{1}{k}\frac{\operatorname{diam} E}{2}=\frac{\operatorname{diam} E}{2\lceil mn \rceil}.$$

■

证明 [推论 10] 的证明。上界通过应用反三角不等式得到。

至于下界，请注意

$$\begin{aligned}\sup_{\mathcal{P}}|u_D(\mathcal{P})-u_{D'}(\mathcal{P})| &\geq |u_D(\mathcal{P}^{\mathrm{DTM}_1}(D))-u_{D'}(\mathcal{P}^{\mathrm{DTM}_1}(D))|\\ &= \mathrm{d_B}\left(\mathcal{P}^{\mathrm{DTM}_1}(D'),\mathcal{P}^{\mathrm{DTM}_1}(D)\right) \geq \mathrm{d_B}\left(\mathcal{P}_0^{\mathrm{DTM}_1}(D'),\mathcal{P}_0^{\mathrm{DTM}_1}(D)\right).\end{aligned}$$

根据命题 直径$E/(2\lceil mn \rceil)$，最后一个表达式的下界为 9。 ■

B.2 第4节结果的证明

证明 [命题的证明 15] 注意，不等式（2.5）给出

$$\mathbb{P}\left[\,|u_D(\mathcal{P}_{\mathrm{DP}}) - u_D(\widetilde{\mathcal{P}}_{\mathrm{OPT}})| \geq \frac{2\Delta}{\epsilon}(\log|\widetilde{\mathsf{Pers}}^{(\ell+1)}_{M,N}| + t)\right] \leq e^{-t},$$

对每个 $t \geq 0$ 均成立。由反三角不等式可得

$$\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \widetilde{\mathcal{P}}_{\mathrm{OPT}}) \geq |u_D(\mathcal{P}_{\mathrm{DP}}) - u_D(\widetilde{\mathcal{P}}_{\mathrm{OPT}})|.$$

结合这两点可得

$$\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \widetilde{\mathcal{P}}_{\mathrm{OPT}}) = O_p\left(\frac{\Delta}{\epsilon}(\ell+1)\log\left|\widetilde{\mathsf{Pers}}_{M,N}\right|\right).$$

回忆一下，Δ 被选为 $\Delta = (\ell+1)\mathrm{diam}E/(mn)$，且 $|\widetilde{\mathsf{Pers}}_{M,N}| = N^M$。因此，我们有

$$\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \widetilde{\mathcal{P}}_{\mathrm{OPT}}) = O_p\left(\frac{(\ell+1)^2 M \log N}{n\epsilon}\right).$$

现在，回忆一下，左上三角形 $\bar{T}$ 被 $N^2 = N^2(n)$ 个 $= n^2$ 等距点离散化；每个间距的长度被某个常数 C 乘以 $C\mathrm{diam}E/n$ 所界定，其中该常数仅取决于所选的欧几里得距离。因此，对于所有足够大的 n，用 $\mathcal{P}_{\mathrm{OPT}}$ 近似 $\mathcal{P}_{\mathrm{OPT}}$ 的误差满足

$$\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \widetilde{\mathcal{P}}_{\mathrm{OPT}}) = O_p(n^{-1}),$$

至此，证明完成。 ■

定理 17 可通过证明以下结果得出。

命题25 设 $M = M(n)$ 是一个非递减的正整数序列，满足 $M(n) \geq |\mathcal{P}_q(\mu)|$ 对所有充分大的 n 均成立。则对每个 $q \geq 0$，有

$$\mathrm{d_B}(\mathcal{P}_{q,\mathrm{OPT}}, \mathcal{P}_q(\mu)) = O_p(n^{-1/d}).$$

此外，我们还有

$$\mathrm{d_B}(\mathcal{P}_{q,\mathrm{OPT}}, \mathcal{P}_q(D)) = O_p(n^{-1/d}).$$

证明 [命题 25] 的证明。取充分大的 n，我们可以假设 $M \geq |\mathcal{P}_q(\mu)|$。换言之，$\mathcal{P}_q(\mu)$ 属于至多含有 M 个元素的持久图空间。因此，根据 $\mathcal{P}_{q,\mathrm{OPT}}$ 的定义，

$$\mathrm{d_B}(\mathcal{P}_q(D), \mathcal{P}_{q,\mathrm{OPT}}) \leq \mathrm{d_B}(\mathcal{P}_q(D), \mathcal{P}_q(\mu)).$$

对于 $\mathrm{d_B}(\mathcal{P}_q(\mu)\mathcal{P}_{q,\mathrm{OPT}})$，由三角不等式可得

$$\begin{aligned}\mathrm{d_B}(\mathcal{P}_q(\mu), \mathcal{P}_{q,\mathrm{OPT}}) &\leq \mathrm{d_B}(\mathcal{P}_q(\mu), \mathcal{P}_q(D)) + \mathrm{d_B}(\mathcal{P}_q(D), \mathcal{P}_{q,\mathrm{OPT}}) \\ &\leq 2\,\mathrm{d_B}(\mathcal{P}_q(\mu), \mathcal{P}_q(D)).\end{aligned}$$

根据 Fournier 和 Guillin （2015年）中的定理 2，以及瓶颈距离的稳定性定理（2.2）和 DTM 的 Wasserstein 稳定性（3.1），可以直接推导出

$$\mathrm{d_B}(\mathcal{P}_q(\mu), \mathcal{P}_q(D)) = O_p(n^{-1/d})$$

对每个 $q \geq 0$。 ■

证明 [定理的证明 17]根据命题25，我们有

$$\mathrm{d_B}(\mathcal{P}_{q,\mathrm{OPT}}, \mathcal{P}_q(\mu)) = O_p(n^{-1/d})$$

对每个 $q \geq 0$。此估计结合命题 15给出的估计，即可得到所需结果。

■

定理18的证明通过建立以下三个引理来完成。第一个引理相当技术性。

引理 26 设 $\mathcal{P} = \mathcal{P}_0^{\mathrm{DTM}_1}$。假设 *DTM* 的分辨率 m 选择为满足$m \leq 1/2$。对于任意一对满足 $1 \leq K \leq n$的正整数 K 和 n，存在一对满足 $|X_n| = |Y_n| = n$, $H(X_n, Y_n) = K$的数据集 X_n和 Y_n，以及 $\mathrm{d_B}(\mathcal{P}(X_n)\ \mathcal{P}(Y_n)) \geq \frac{CK}{n}$，其中某个常数 C 与 K 和 n无关，且 $H(X_n, Y_n)$ 表示 X_n 和 Y_n之间的汉明距离。

证明

回顾一下 $k = \lceil mn \rceil$。整个情况将分为三种情形：(i) $1 \leq K \leq \min\{k, n/2 - k\}$, (ii) $\min\{k, n/2 - k\} < K < \max\{k, n/2 - k\}$，以及 (iii)$K \geq \max\{k, n/2 - k\}$。

首先，假设 $1 \leq K \leq \min\{k, n/2 - k\}$。选择两个满足$\| a - b \| = \mathrm{diam}\ E$的点 a和 b；例如，在 $E = [0,1]^d$的情况下，可以选择 $a = (0, \ldots, 0)$ 和 $b = (1, \ldots, 1)$。选择由 $n/2$ 个 a 副本和 $n/2$ 个 b 副本组成的数据集 X_n （如果 n 为奇数，则取 $(n-1)/2$ 个 a 副本和 $(n+1)/2$ 个 b副本；结果相同）。另一方面，选择通过将 K 个 a 副本移动到 a 和 b 的中点（记为 c）而构造的数据集 Y_n，即作为多重集， X_n 和 Y_n 可以表示如下：

$$X_n = \{ (a, n/2), (b, n/2) \} \text{ and } Y_n = \{ (a, n/2 - K), (c, K), (b, n/2) \}.$$

由于 $M \leqslant n/2 - k$，点 a 仍然在数据集 Y_n 中拥有超过 k 个点。因此，我们有

$$\delta_{Y_n}(x) = \begin{cases} 0 & \text{if } x = a, \\ \frac{k-K}{k} \frac{\mathrm{diam}\, E}{2} & \text{if } x = c, \\ 0 & \text{if } x = b. \end{cases}$$

设 $x(t)$ 为线段 ac 上的点，它将 ac 按 $t : (1-t)$ 的比例分割，其中$t \in [0,1]$。那么，我们有

$$\delta_{Y_n}(x(t)) = \begin{cases} t \frac{\mathrm{diam}\, E}{2} & \text{if } 0 \leq t \leq 1/2, \\ \frac{(k-K)t\, \mathrm{diam}\, E/2 + K(1-t)\, \mathrm{diam}\, E/2}{k} & \text{if } 1/2 < t \leq 1. \end{cases}$$

现在，我们进一步将情况分为两种：(i-1) $(k-2K) \geq 0 \Leftrightarrow K \leq k/2$，以及 (i-2) $(k-2K) < 0 \Leftrightarrow K > k/2$。在 (i-1) 的情况下，$\delta_{Y_n}(x(t))$ 关于 t 单调递增。因此，$\mathcal{P}(Y_n)$ 可得

$$\mathcal{P}(Y_n) = \{(0,\infty), (0, \delta_{Y_n}(c))\}.$$

请注意，得到的 $\mathcal{P}(X_n)$ 为

$$\mathcal{P}(X_n) = \{(0,\infty), (0, \operatorname{diam} E/2)\}.$$

因此，

$$\begin{aligned} \mathrm{d_B}(\mathcal{P}(X_n), \mathcal{P}(Y_n)) = \frac{\operatorname{diam} E}{2} - \delta_{Y_n}(c) &= \frac{\operatorname{diam} E}{2} - \left(\frac{k-K}{k}\right)\frac{\operatorname{diam} E}{2} \\ &= \frac{K}{k}\frac{\operatorname{diam} E}{2} \\ &= \frac{\operatorname{diam} E}{2m}\frac{K}{n}. \end{aligned}$$

在情况 (i-2) 中，$\delta_{Y_n}((x(t))$ 从 $t = 1/2$ 降至 $t = 1$。因此，$\mathcal{P}(Y_n)$ 被给定为

$$\mathcal{P}(Y_n) = \{(0,\infty), (\delta_{Y_n}(c), \operatorname{diam} E/4), (0, \operatorname{diam} E/4)\}.$$

$\mathcal{P}(X_n)$ 与 $\mathcal{P}(Y_n)$ 之间的瓶颈距离可以通过比较两种不同的情形推导得出。第一种情形对应于 $\mathcal{P}(Y_n)$ 中的 $(0, \operatorname{diam}/4)$ 与 $\mathcal{P}(X_n)$ 中的 $(0, \operatorname{diam} E/2)$ 之间的匹配。由此情形得到的距离必须大于或等于 $\operatorname{diam} E/4$。另一种情形对应于 $\mathcal{P}(X_n)$ 中的 $(0, \operatorname{diam} E/2)$ 与 $\mathcal{P}(Y_n)$ 中的 $(\delta_{Y_n}(c), \operatorname{diam} E/4)$ 之间的匹配。因此，$\mathcal{P}(Y_n)$ 中的 $(0, \operatorname{diam} E/4)$ 必须与对角线上的某个点相对应。由此情形得到的距离必须大于或等于 $\operatorname{diam} E/(4\sqrt{2})$。因此，

$$\mathrm{d_B}(\mathcal{P}(X_n), \mathcal{P}(Y_n)) \geq \frac{\operatorname{diam} E}{4\sqrt{2}}.$$

现在，让我们将注意力转向情况 (ii)，该情况假设 $\min\{k, n/2-k\} < K < \max\{k, n/2-k\}$。首先，考虑情况 $k < n/2-k$，因此 $k < K < n/2-k$。在这种情况下，a 和 c 都至少有 k 个点，因此

$$\delta_{Y_n}(x) = 0 \text{ for all } x = a, b, \text{and } c.$$

上述结果向我们表明

$$\mathcal{P}(Y_n) = \{(0,\infty), (0, \operatorname{diam}/4), (0, \operatorname{diam} E/4)\}.$$

因此，

$$\mathrm{d_B}(\mathcal{P}(X_n), \mathcal{P}(Y_n)) = \frac{\operatorname{diam} E}{4\sqrt{2}}.$$

其次，考虑情形 $k > n/2-k$，因此 $n/2-k < K < k$。在这种情况下，a 和 c 的点数都少于 k。因此，

$$\delta_{Y_n}(x) = \begin{cases} \frac{k-n/2+K}{k}\frac{\operatorname{diam} E}{2} & \text{if } x = a, \\ \frac{k-K}{k}\frac{\operatorname{diam} E}{2} & \text{if } x = c, \\ 0 & \text{if } x = b \end{cases}.$$

采用我们在情形 (i) 中使用的类似论证，可以证明所需结果在这种情况下同样成立。

最后，考虑情形 (iii)，其中 $K \geq \max\{k, n/2-k\}$。在这种情形下，$\mathcal{P}(Y_n)$ 至少包含一个元素 (0，diam /4)。因此，它与 $\mathcal{P}(X_n)$ 之间的瓶颈距离始终大于或等于 diam $E/4$。至此，引理的证明完成。

■

接下来的两个引理从假设检验的框架来阐述差分隐私的概念。引理 27是差分隐私文献中广为人知的一个经典结果。利用 f-差分隐私框架可以轻松推导出该结果（Dong 等人，2022年）。我们给出一个直接证明，无需使用 f-差分隐私。

引理 27 设 X 和 X' 是相邻数据集，$\mathcal{M}$ 是任意一个 ϵ-差分隐私机制。对于假设检验 $H_0 : \mathcal{M}(X)$ 对比 $H_1 : \mathcal{M}(X')$，

$$\text{Type I error} + \text{Type II error} \geq \frac{2}{1+e^{\epsilon}}.$$

证明 记 $\mathcal{Y}$ 为 $\mathcal{M}(X)$ 所处的概率空间。记 μ_X 为 $\mathcal{M}(X)$ 在 $\mathcal{Y}$ 上的概率测度。根据 Awan 等人（2019年，命题 2.3），存在一个基准测度 ν，它支配所有数据库 X对应的 μ_X。记 f_X 为 μ_X 关于 ν 的密度；根据 Awan 等人（2019年，命题 2.3），对于相邻数据库 X 和 X'，该密度几乎处处满足 $f_X \leq e^{\epsilon} f_{X'}$ ν。

设 X 和 X' 是相邻数据库，且 $\phi : \mathcal{Y} \to [0,1]$ 是一个检验。那么第一类错误和第二类错误分别为 $\mathrm{I} = \mathbb{E}\phi(\mathcal{M}(X))$ 和 $\mathrm{II} = 1 - \mathbb{E}\phi(\mathcal{M}(X'))$。于是

$$\begin{aligned}\mathrm{I} = \mathbb{E}\phi(\mathcal{M}(X)) = \int \phi(t) f_X(t) d\nu &\geq e^{-\epsilon} \int \phi(t) f_{X'}(t) d\nu \\ &= e^{-\epsilon}\mathbb{E}\phi(\mathcal{M}(X')) \\ &= e^{-\epsilon}(1-\mathrm{II}).\end{aligned}$$

重复 使用 $\phi' = 1 - \phi$ 重复上述论证，并交换 X 和 X'的角色，我们得到

$$\mathrm{II} \geq e^{-\epsilon}(1-\mathrm{I}).$$

于是，

$$\mathrm{I} + \mathrm{II} \geq e^{-\epsilon}[2 - (\mathrm{I}+\mathrm{II})],$$

这意味着 $\mathrm{I} + \mathrm{II} \geq \frac{2}{1+e^{\epsilon}}$。 ——

■

引理 *28* 设 $(\epsilon_n)_{n=1}^{\infty}$ 是满足以下条件的正数序列：$1/n \leq \epsilon_n \leq 1$ 对每个n均成立。令 $K_n = \lfloor 1/\epsilon_n \rfloor$。对于每个 n，给定正数序列 $(\Delta_n)_{n=1}^{\infty}$，设$\{(X_n, Y_n)\}_{n=1}^{\infty}$ 是满足以下条件的有限数据集序列：对每个 n，$H(X_n, Y_n) = K_n$ 且 $d_{\mathrm{B}}(\mathcal{P}(X_n)\ \mathcal{P}(Y_n)) \geq K_n\Delta_n$。这里，$H$ 表示集合间的汉明距离，$\mathcal{P}$表示任意持久图。那么，对于任何能生成隐私化持久图的 $\epsilon(n)$-差分隐私机制 $\mathcal{M}$，以下两个条件不可能同时成立：$d_{\mathrm{B}}(P_{X_n}, \mathcal{M}(X_n)) = o_p(\Delta_n/\epsilon_n)$ 和 $d_{\mathrm{B}}(P_{Y_n}, \mathcal{M}(Y_n)) = o_p(\Delta_n/\epsilon_n)$。

证明 为简化记号，我们将省略 X、 Y、 ϵ、 Δ和 K 对n的依赖关系。我们将为 H_0 构造一个假设检验： $\mathcal{M}(X)$ 对比 H_1： $\mathcal{M}(Y)$。注意，由于$\mathcal{M}$ 对于大小为 1 的群组满足 ϵ-差分隐私，因此对于大小为 $K\epsilon$ 的群组，它满足 K (Dwork 和 Roth，2014年，定理 2.2）。

定义集合 S_X 和 S_Y 如下：

$$S_X = \{\mathcal{P} \mid \mathrm{d_B}(\mathcal{P}, \mathcal{P}(X)) < K\Delta/2\}$$
$$S_Y = \{\mathcal{P} \mid \mathrm{d_B}(\mathcal{P}, \mathcal{P}(Y)) < K\Delta/2\}$$

并将我们的检验定义为 $\phi(\mathcal{M}(\cdot)) = I(\mathcal{M}(\cdot) \in S_Y)$，它是事件 $\mathcal{M}(\cdot) \in S_Y$ 的指示函数。于是

$$\text{Type I error } = \mathbb{P}(\mathcal{M}(X) \in S_Y) \leq \mathbb{P}(\mathcal{M}(X) \notin S_X)$$
$$\text{Type II error} = \mathbb{P}(\mathcal{M}(Y) \notin S_Y).$$

根据引理27，我们有

$$\mathbb{P}(\mathcal{M}(X) \notin S_X) + \mathbb{P}(\mathcal{M}(X') \notin S_Y) \geq \frac{2}{1+e^{k\epsilon}} \geq \frac{2}{1+e},$$

因为 $k\epsilon \leq 1$，这意味着以下情况之一成立：

$$\mathbb{P}(\mathrm{d_B}(\mathcal{M}(X), \mathcal{P}(X)) \geq K\Delta/2) \geq \frac{1}{1+e},$$

or

$$\mathbb{P}(\mathrm{d_B}(\mathcal{M}(Y), \mathcal{P}(Y)) \geq K\Delta/2) \geq \frac{1}{1+e}.$$

这排除了二者都是 $o_p(K\Delta) \leq o_p(\Delta/\epsilon)$ 的可能性。 ■

证明 [定理的证明 18] 根据引理 26，可以保证 0阶持久性图的 L^1-DTM 过滤 $\mathcal{P}_0^{\mathrm{DTM}_1}$满足引理 28 中规定的条件，其中 $\Delta_n = \frac{C}{mn}$ 对于某个与 C 无关的常数 n。为简洁起见，令 $\mathcal{P} = \mathcal{P}_0^{\mathrm{DTM}_1}$。然后，对于满足该条件的数据集序列 (X_n, Y_n) ，引理 28 告诉我们，以下两种情况之一成立：

$$\mathbb{P}(\mathrm{d_B}(\mathcal{M}(X_n), \mathcal{P}(X_n)) \geq CK/n) \geq \frac{1}{1+e}$$

or

$$\mathbb{P}(\mathrm{d_B}(\mathcal{M}(Y_n), \mathcal{P}(Y_n)) \geq CK/n) \geq \frac{1}{1+e}$$

成立，因此排除了它们是 $o_p(K/n) \leq o_p(1/(n\epsilon))$ 的可能性。证明至此完成。

■

附录 ix C. 模拟和真实数据分析的补充内容 is

C.1 我们算法的更详细描述

这里将详细介绍我们在第 4 节中提出的隐私机制算法。对于给定的数据集 D，令 S 表示该数据集上 L^1-DTM 函数的最大值，并令 M 为指定的正整数。

为了获得初始图 $\mathcal{P}^{(0)}$ 差分隐私 $= (\mathcal{P}_0^{(0)}$，差分隐私 $\mathcal{P}_1^{(0)}$，差分隐私$)$，从闭区间 $[0, S]$ 上的均匀分布中独立同分布地生成样本 $x_1, \ldots, x_M, z_1, \ldots, z_M$，即 $x_1, \ldots, x_n, z_1, \ldots, z_M \overset{i.i.d.}{\sim}$ 均匀分布 $[0, S]$，用符号表示。对于每个 $i \in \{1, \ldots, M\}$，设

$$y_i = x_i + (1 - x_i)z_i;$$

图 $\mathcal{P}_0^{(0)}$，差分隐私 的构造如下：

$$\mathcal{P}_{0,\mathrm{DP}}^{(0)} = \{(x_1, y_1), \ldots, (x_M, y_M)\}.$$

另一个初始图 $\mathcal{P}_1^{(0)}$，差分隐私 以同样的方式生成，且独立于 $\mathcal{P}_0^{(0)}$，差分隐私。注意，每个初始图都由 M 个在左上三角形上均匀分布的点组成$\{(x,\ y)$：$\ x,\ y \in [0, S]$ 和 $y \geqslant x\}$。

对于生成的 $\mathcal{P}^{(t)}$ 差分隐私 $= (\mathcal{P}_0^{(t)}$，差分隐私 $\mathcal{P}_1^{(t)}$，差分隐私$)$，通过向第 t 步中每个图的每个元素添加高斯噪声，得到下一个候选图 $\mathcal{P}^{(t+1)}$ 差分隐私。具体而言，记

$$\mathcal{P}_{0,\mathrm{DP}}^{(t)} = \{(x_1^{(t)}, y_1^{(t)}), \ldots, (x_M^{(t)}, y_M^{(t)})\}.$$

从均值为 (0，0)、协方差矩阵为 $\sigma^2 I_2$ 的二维高斯分布中生成独立同分布样本 $Z_1^{(t)}$，……，$Z_M^{(t)}$，其中 σ 是预先指定的正数，I_2 是 2×2 单位矩阵。令

$$\mathcal{P}_0' = \{(x_1^{(t)}, y_1^{(t)}) + Z_1^{(t)}, \ldots, (x_M^{(t)}, y_M^{(t)}) + Z_M^{(t)}\}$$

$\mathcal{P}_1'$ 的构造方式与 $\mathcal{P}_0'$ 相同，且二者相互独立，并令 $\mathcal{P}' = (\mathcal{P}_0'$，$\mathcal{P}_1')$。然后，计算以瓶颈距离定义的 Metropolis-Hastings 采样器 p 中的接受/拒绝概率：

$$p = \min\left\{0, -\frac{\epsilon}{2\Delta}\left(u_D(\mathcal{P}_{\mathrm{DP}}^{(t)}) - u_D(\mathcal{P}')\right)\right\},$$

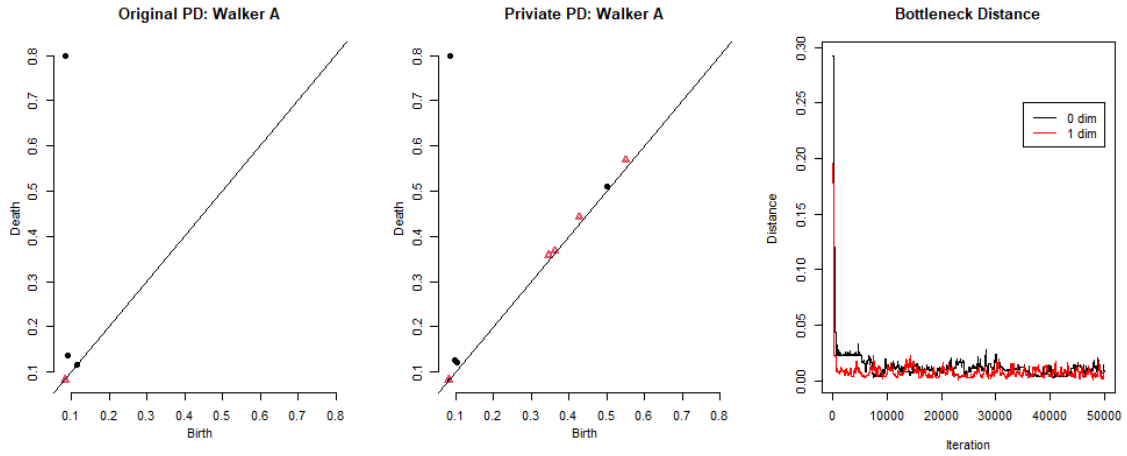
其中

$$\Delta = \frac{2\sqrt{2}}{mn}.$$

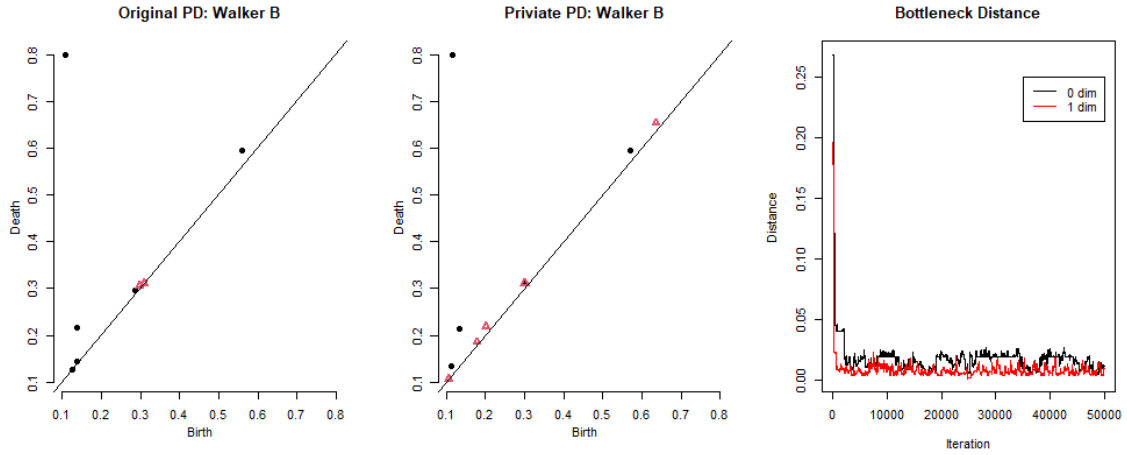
生成 $U \sim$ 均匀分布$(0,\ 1)$。如果 $\log U \leq p$，则取 $\mathcal{P}^{(t+1)}$ 差分隐私$= \mathcal{P}'$；否则，取 $\mathcal{P}^{(t+1)}$ 差分隐私$= \mathcal{P}^{(t)}$ 差分隐私。重复执行此过程，并将最后一次迭代得到的 $\mathcal{P}^{(t)}$ 差分隐私作为隐私化持久图。整个过程总结于算法 1。

C.2 真实数据分析中的其他结果

下图（图8）展示了 Walker A 和 B的隐私化持久图的准确性。机制的实现过程与第6节中描述的 Walker C 相同。可以看到，在这些情况下，我们也获得了相当准确的隐私化图表。



(a) Walker A



(b) Walker B

图 8：图 8-(a) 展示了 Walker A 数据的 L^1-DTM 持久图及其隐私化图。同时还展示了真实图与隐私化图之间的瓶颈距离随马尔可夫链蒙特卡洛迭代的变化。在最终迭代时，我们得到 $\mathrm{d_B}(P_0(D)P_{0,\ \mathrm{DP}}) = 0.01$ and $\mathrm{d_B}(P_1(D)P_{1,\ \mathrm{DP}}) = 0.009$。图 8-(b) 展示了关于 Walker B 的同类信息。此处，在最终迭代时，我们得到 $\mathrm{d_B}(P_0(D)P_{0,\ \mathrm{DP}}) = 0。011$ 和 $\mathrm{d_B}(P_1(D)P_{1,\ \mathrm{DP}}) = 0.009$。

算法 1 指数机制的马尔可夫链蒙特卡洛实现

输入： $\mathcal{P}_0(D)$ $\mathcal{P}_1(D)$，以及一个正整数 M

初始化： $\mathcal{P}^{(0)}_{0,\mathrm{DP}}$ $\mathcal{P}^{(0)}_{1,\mathrm{DP}} \sim$ 均匀分布(Pers$_M$)

适用于 $i = 1, 2, \ldots,$ **do**

$\mathcal{P}'_0 = \mathcal{P}^{(t-1)}_{0,\text{差分隐私}} + \mathrm{N}(0, \sigma^2 I_2)$, $\mathcal{P}'_1 = \mathcal{P}^{(t-1)}_{1,\text{差分隐私}} + \mathrm{N}(0, \sigma^2 I_2)$

$\mathcal{P}' = (\mathcal{P}'_0, \mathcal{P}'_1)$

$p =$ 最小化 $\left\{0, \frac{\epsilon}{2\Delta}\left(u_D(\mathcal{P}^{(t-1)}) - u_D(\mathcal{P}')\right)\right\}$

$U \sim$ 均匀分布(0,1)

如果 $\log U \leq p$ **则**

$\mathcal{P}^{(t)}_{0,\mathrm{DP}} = \mathcal{P}'_0, \mathcal{P}^{(t)}_{1,\mathrm{DP}} = \mathcal{P}'_1$

else

$\mathcal{P}^{(t)}_{0,\mathrm{DP}} = \mathcal{P}^{(t-1)}_{0,\mathrm{DP}}, \mathcal{P}^{(t)}_{1,\mathrm{DP}} = \mathcal{P}^{(t-1)}_{1,\mathrm{DP}}$

结束条件判断

结束循环