Journal of Machine Learning Research 25 (2024) 1-42 Submitted 5/23; Revised 11/23; Published 5/24

Differentially Private Topological Data Analysis

Taegyu Kang* KANG426@PURDUE.EDU
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Sehwan Kim* SEHWAN_KIM@HPHCI.HARVARD.EDU
Department of Population Medicine
Harvard Medical School/Harvard Pilgrim Health Care Institute
Boston, MA, 02215, USA

Jinwon Sohn* SOHN24@PURDUE.EDU
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Jordan Awan† JAWAN@PURDUE.EDU
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Editor: Manuel Gomez-Rodriguez

Abstract

This paper is the first to attempt differentially private (DP) topological data analysis (TDA), producing near-optimal private persistence diagrams. We analyze the sensitivity of persistence diagrams in terms of the bottleneck distance, and we show that the commonly used Čech complex has sensitivity that does not decrease as the sample size n increases. This makes it challenging for the persistence diagrams of Čech complexes to be privatized. As an alternative, we show that the persistence diagram obtained by the L^1-distance to measure (DTM) has sensitivity $O(1/n)$. Based on the sensitivity analysis, we propose using the exponential mechanism whose utility function is defined in terms of the bottleneck distance of the L^1-DTM persistence diagrams. We also derive upper and lower bounds of the accuracy of our privacy mechanism; the obtained bounds indicate that the privacy error of our mechanism is near-optimal. We demonstrate the performance of our privatized persistence diagrams through simulations as well as on a real data set tracking human movement.

Keywords: Čech complex, Distance to a measure, Exponential mechanism, Persistence diagram, Persistent homology

*: Kang, Kim, and Sohn are co-first authors and they contribute equally to this paper. †: Corresponding author.

1. Introduction

Recent advances in technology make it possible to obtain data with such complicated structure that traditional data analysis methodologies cannot deal with them appropriately. To analyze such complex data, topological data analysis has been an indispensable tool in data science (Niyogi et al., 2011; Khasawneh and Munch, 2016; Wasserman, 2018; Dindin et al., 2020; Rieck et al., 2020). Essentially, topology is the most fundamental mathematical structure where the notion of "*nearness*" can be discussed, and its generality makes it an appropriate framework for discussing extremely complicated data which are not expected to have more equipped structures such as vector spaces, manifolds, and so on. Topological data analysis is a novel branch of data analysis which was invented to capture the topological structure of data, and it has been deeply studied for the last couple of decades. See Carlsson (2009) for a comprehensive overview. Especially, persistent homology, its flagship method, has been extensively studied theoretically and applied to many different disciplines such as medicine (Nicolau et al., 2011), biology (McGuirl et al., 2020), neuroscience (Xu et al., 2021; Caputi et al., 2021), astronomy (Xu et al., 2019), and machine learning (Hensel et al., 2021; Betthauser et al., 2022), to name a few.

At the same time, as bigger and more diverse data have become accessible, the issue of protecting private information of individuals in the data has also gained attention. Due to this concern, there is an increasing demand for privacy-protecting procedures with formal guarantees. Such a paradigm has accelerated the growing attention to a well-formulated framework of privacy protection in data science. Differential privacy (DP), introduced by Dwork et al. (2006), is the state-of-the-art framework that formally quantifies the notion of privacy and its protection. Differential privacy requires that a privacy-protecting algorithm produces similar results for any two data sets, which differ at only one data point. The exact definition of ϵ-DP will be introduced in the following section and we recommend Dwork and Roth (2014) for a comprehensive introduction to DP. Recently, DP has been a growing research topic in data science, tackling problems in deep learning (Shokri and Shmatikov, 2015; Abadi et al., 2016), functional data analysis (Hall et al., 2013; Mirshani et al., 2019), social networks (Karwa and Slavković, 2016; Karwa et al., 2017), as well as many others.

While the DP framework has been widely adapted to numerous methodologies in data science as mentioned above, its application to TDA has yet to be discussed. To the best of our knowledge, the only work involved with both DP and TDA is Hehir et al. (2022), which solely used persistence diagrams as a method of communicating the utility of a randomized response algorithm, and did not attempt to produce a private version of a TDA object. We believe that introducing the DP framework to TDA will be an emerging direction of research because many areas where TDA methods have been successfully utilized use data containing people's sensitive information. For example, Shnier et al. (2019) applied persistence diagrams to differentiate gene expressions in individuals with autism spectrum disorders from those in a control group. Furthermore, TDA methods are used in several other problems in medical domain and neuroscience, as mentioned above, such as brain connectivity (Caputi et al., 2021), breast cancer (Nicolau et al., 2011), and neurological disorder (Lee et al., 2011). Finally, TDA has recently been combined with other popular machine learning methods such as convolutional neural networks (Love et al., 2023), auto-encoders (Hofer et al., 2019), etc. Hence, introducing DP to TDA may have far-reaching influence in data science.

Our Contributions: This paper is concerned with how to introduce the concept of differential privacy (DP) into the framework of topological data analysis (TDA). Our key observation is that, to exploit currently available privacy mechanisms, one needs an outlier-robust TDA method. Such an observation agrees with a long-standing intuitive principle in differential privacy saying that the specific data of any one individual should not have a significant effect on the outcome of the analysis to achieve privacy protection; for instance, see Dwork and Lei (2009), Avella-Medina (2021). To illuminate the adaptation of this principle to TDA, we examine the sensitivity of the bottleneck distance of persistence diagrams, which is the most widely used presentation of persistent homology, obtained by two different types of construction: persistence diagrams obtained from Čech complexes, which we show is not outlier-robust; and persistence diagrams obtained from the distance to a measure (DTM), which is outlier-robust. Our examination shows why persistence diagrams of Čech complexes are not readily privatized, and how persistence diagrams of the DTM can overcome such a difficulty. Moreover, we discuss how the magnitude of outlier-robustness affects the rate of sensitivity of the bottleneck distance, and propose to use L^1-DTM in order to achieve a minimal sensitivity. Based on the sensitivity analysis, we propose the first differentially private mechanism for persistence diagrams that provides ϵ-differential privacy, using the exponential mechanism. We also establish upper and lower bounds for the accuracy error of our mechanism. The established bounds indicate that the privacy error of our mechanism is near-optimal. Our contributions can be summarized more specifically as follows:

- We prove that the sensitivity of the persistence diagram of Čech complexes, defined in terms of the bottleneck distance, does not diminish to zero as the sample size increases.
- We propose using the persistence diagram of the distance to measure (DTM) as an alternative, and we prove that the L^p-DTM persistence diagram is guaranteed to have sensitivity, which is defined in terms of the bottleneck distance, $O(n^{-1/p})$. This leads us to use the L^1-DTM persistence diagram that guarantees the sensitivity $O(n^{-1})$.
- We apply the exponential mechanism whose utility function is defined in terms of the bottleneck distance of L^1-DTM persistence diagrams in order to produce differentially privatized persistence diagrams. To the best of our knowledge, our algorithm is the first attempt of developing a mechanism generating differentially privatized persistence diagrams. We also find upper and lower bounds of the accuracy error of our mechanism.
- We prove that any privacy mechanism applied to the L^1-DTM persistence diagrams cannot have accuracy whose decay order is superior to the upper bound of the decay order of the privacy error corresponding to our mechanism. This result indicates that our mechanism may have optimal privacy error.

Organization: The remainder of the paper is organized as follows. In Section 2, we briefly review the background and notation of TDA and DP. In Section 3, we first examine the sensitivity of persistence diagrams constructed from the Čech complexes, as well as for an outlier-robust construction of persistence diagrams obtained from the DTM, introduced by Chazal et al. (2011). In Section 4, based on the sensitivity analysis given in Section 3, we employ the exponential mechanism to generate privatized persistence diagrams. We also derive upper and lower bounds of its accuracy. Simulation studies which implement our

algorithm are given in Section 5. In Section 6, we apply our algorithm to a real-world data set including information about the locations of three people walking in a building recorded on smartphones over time. All proofs as well as additional results of real data analyses are presented in the appendices.

2. Preliminaries

In this section, we introduce the persistence diagram, which is a statistic about the shape of the data, and look at bottleneck distance, a metric in the persistence diagram space as well as its stability. Also, we review the ϵ-differential privacy (ϵ-DP) and the exponential mechanism, which is one of the algorithms that satisfies ϵ-DP.

Notation Throughout the paper, for real numbers A and B which possibly depend on a parameter $n \in \mathbb{Z}_+$, we use the asymptotic notation $A \lesssim B$ or $A = O(B)$ to denote the bound $|A| \leqslant CB$ for some absolute constant $C > 0$. If the constant C depends on some parameters, we will explicitly indicate them; for instance, $A \lesssim_{k,d} B$ means the bound $|A| \leqslant C_{k,d}B$ with a constant $C_{k,d}$ depending only on k and d. If $A \lesssim B$ and $A \gtrsim B$, we write denote it by $A \approx B$. We also use the notation $A = o(B)$ to denote the asymptotic result $\lim_{n\to\infty} A/B = 0$. In addition, for random variables X and Y which possibly depend on a parameter $n \in \mathbb{Z}_+$, we write $X = O_p(Y)$ to mean that X/Y is bounded in probability and $X = o_p(Y)$ to mean that X/Y converges to zero in probability, which are standard notations in probability theory. Let $(X_n)_{n=1}^{\infty}$ be a sequence of random variables. We write $X_n = \tilde{O}_p\big(f(n)\big)$ to mean that $X_n = O_p\big(f(n)\log^k n\big)$ for some $k \in \mathbb{Z}_+$.

For a given metric space $(\mathcal{X}, \mathrm{d})$, $\mathcal{D}_n := \mathcal{D}_n(\mathcal{X}) := \mathcal{X}^n$ denotes the set of all n-tuples of elements in $\mathcal{X}$ for every $n \in \mathbb{Z}_+$.

2.1 Persistent Homology and Diagrams

Here, we briefly introduce two methods of constructing persistent homology and corresponding persistence diagrams of data, which will show up in our main discussion. The former one is the persistent homology of Čech complex and the latter one is the persistent homology of the sub-level sets of a continuous function. We believe that an intuitive and illustrative description of persistent homology will suffice to understand the results of this paper. More detailed background knowledge about persistent homology along with some fundamental knowledge about simplicial homology is presented in Appendix A. For a deeper and comprehensive understanding for persistent homology, we refer the reader to the literature of persistent homology; for instance, Edelsbrunner and Harer (2008, 2009); Zomorodian and Carlsson (2005). For fundamental concepts about algebraic topology, we refer the reader to standard texts in algebraic topology such as Munkres (1984); Bredon (1997).

Let $D = \{x_1, \ldots, x_n\}$ be a finite subset of a metric space $(\mathcal{X}, d)$. Let $r > 0$ be a positive real number. At every point x_i, we place a ball $B(x_i; r)$ with radius r centered at x_i. The persistence homology of the Čech complexes on D captures the evolution of the homological structure of the union $\cup_{j=1}^{n} B(x_j; r)$ as r varies. For instance, the 0th homological feature represents the connected components of it and the 1st homological feature represents the loops in it. Figure 1 portrays how to construct the persistent homology of Čech complexes. As the radius r varies, some homological features show up and disappear, and such "birth"

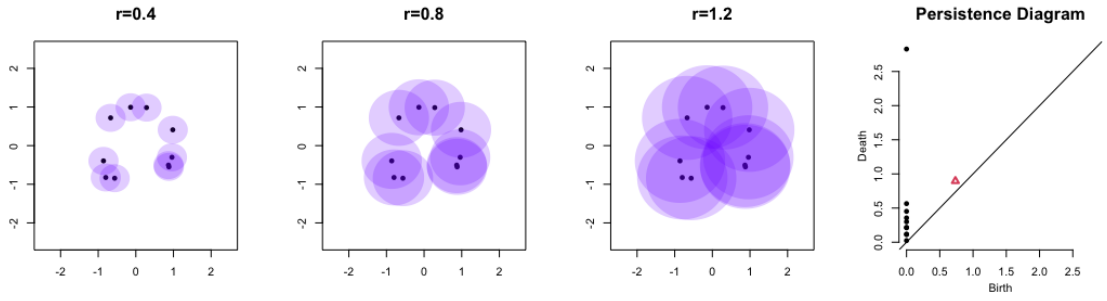


Figure 1: Constructing of Čech complexes and its persistence diagram: The left three figures illustrate how Čech complexes on nine points supported on a circle are constructed. When $r = 0.4$, there are several connected components; but there is no loop. When $r = 0.8$, there exists a 1-dimensional loop that captures the shape of the circle. When $r = 1.2$, the loop disappears and there is only a single contractible connected component. The rightmost figure is the persistence diagram of the Čech complexes. Each black dot represents the birth-and-death times of each connected component and the red triangle represents the birth-and-death times of the loop.

and "death" of homological features are presented as multisets called persistence diagrams. More precisely, a qth persistence diagram of the Čech complexes on the data set D a multiset that consists of finitely many, say m, points (b_i, d_i) satisfying $0 \leq b_i \leq d_i \leq \infty$ for every $i = 1, \ldots, m$; the presence of each point (b_i, d_i) means that there exists a q-dimensional homological feature that shows up at radius b_i and disappears at radius d_i.

As for the other method, let $f_D : \mathcal{X} \to \mathbb{R}$ be a continuous function defined on metric space $(\mathcal{X}, d)$, possibly depending on the given set D. For each $r \in \mathbb{R}$, one can consider the sub-level set $L_r := \{x \in \mathcal{X} : f_D(x) \leqslant r\}$. As we consider the evolution of the union of balls in the previous way of construction, we now consider the evolution of the sub-level sets L_r as r varies. Figure 2 illustrates such an evolution of a certain continuous function. In the figure, three connected components and one loop show up once and disappear at some time except for a single connected component. The birth-death pairs at each dimension can be presented as a persistence diagram, just as for the Čech complex.

In general, let a filtration of topological spaces $\{U_r\}_{r \in R}$ be given, where R is a linearly ordered set; that is, for any r_1 and r_2 in R satisfying $r_1 \leq r_2$, $U_{r_1} \subseteq U_{r_2}$. Then, one can define the persistent homology and the corresponding persistence diagram of the sequence. In our first example, each U_r is the union of balls with radius r (or, the simplicial complex obtained from the balls); in our second example, each U_r is the sub-level set L_r.

2.2 Stability of Persistence Diagrams in the Bottleneck Distance

A persistence diagram $\mathcal{P} = \{(b_i, d_i)\}_{i=1}^m$ is essentially a multiset of birth-death pairs b_i and d_i, which satisfy $b_i \leq d_i$. There are numerous ways to "vectorize" a persistence diagram into an element in some vector space. One of the most popular ways is to represent each birth-death pair (b, d) by the Dirac measure $\delta_{(b,d)}$ at (b, d), and represent the whole diagram $\mathcal{P}$ by the point measure $\sum_{i=1}^m \delta_{(b_i,d_i)}$ which is a measure on the set $\mathcal{T} := \{(x, y) : 0 \leq$

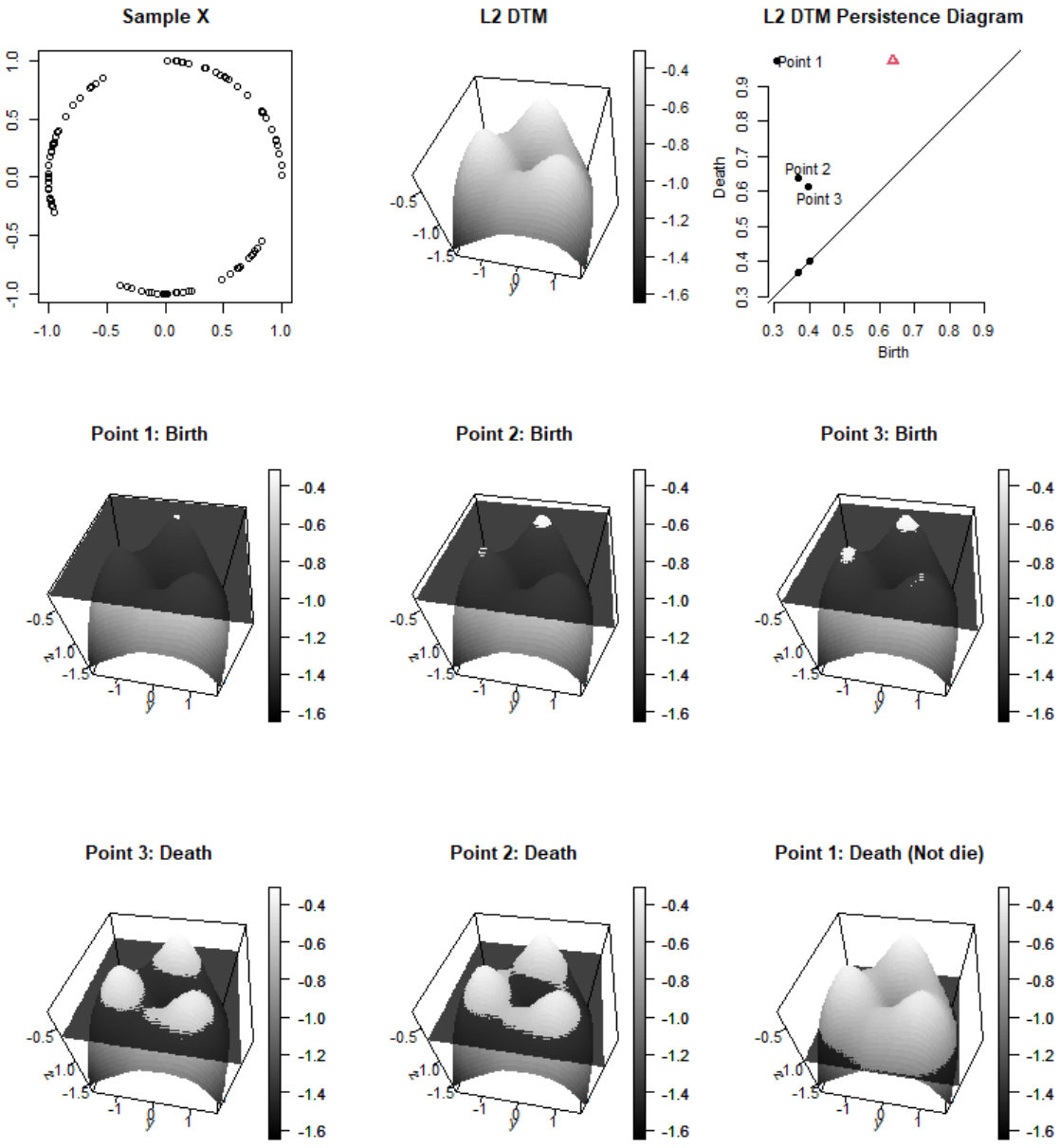


Figure 2: Filtration corresponding to the L^2-DTM of a circle data set: The data set is supported on a circle, and the L^2-DTM function of the data is visualized (for convenience, the function multiplied by -1 is presented). The persistence diagram constructed from the function is presented. The second and third columns present how the filtration of the sub-level sets of the function evolves. As the filtration evolves, three connected components show up at values 0.308, 0.367, and 0.397 respectively. One component dies at a value of 0.613, another one dies at 0.636. The last one lives until the end of evolution. At the right-most figure in the last row, what we mean by "Not die" is that, even though the birth-death pair of Point 1 shows up in the figure the actual death time of Point 1 is infinity.

$x \leq y \leq \infty\}$ (for a detailed description of such a way of vectorization, see Section 2 in Owada (2022)). By realizing a persistence diagram as a measure, it is possible to define the distance between two persistence diagrams by means of a distance between measures. One of the most popular choices is using the L^∞ Wasserstein distance of the measures, which is called the bottleneck distance. Specifically, let $\mathcal{P}, \mathcal{P}'$ be two persistence diagrams. Then the bottleneck distance between $\mathcal{P}$ and $\mathcal{P}'$ is defined as

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

where $\breve{\mathcal{P}}$ and $\breve{\mathcal{P}}'$ denote the persistence diagrams $\mathcal{P}$ and $\mathcal{P}'$ along with the copies of all points on the diagonal respectively; $g : \breve{\mathcal{P}} \leftrightarrow \breve{\mathcal{P}}'$ ranges over all bijections between $\breve{\mathcal{P}}$ and $\breve{\mathcal{P}}'$. In words, $\mathrm{d_B}(\mathcal{P}, \mathcal{P}')$ is the minimax cost of pairing the birth-death points in one diagram to the other diagram one-by-one in terms of ℓ_∞ distance. When two diagrams contain different numbers of birth-death points, then the remaining points in one diagram pair up with the points on the diagonal.

A key property of the bottleneck distance is the following stability property (for more details, see Cohen-Steiner et al. (2007); Chazal et al. (2016a)). Suppose that $\mathcal{P}_q(D)$ and $\mathcal{P}_q(D')$ are qth persistence diagrams constructed from the Čech complexes of two sets D and D' in a metric space $(\mathcal{X}, d)$, then

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

where

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

denotes the Hausdorff distance between D and D'. Analogously, if $\mathcal{P}_q(D)$ and $\mathcal{P}_q(D')$ are obtained from the filtrations of the sub-level sets of continuous tame functions f_D and $f_{D'}$, respectively. Then

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

The precise definition of tame functions is presented in Definition 24. For more comprehensive discussion, please refer to Cohen-Steiner et al. (2007). Intuitively, a $\mathbb{R}$-valued function f is said to be tame if the homology of its sub-level sets changes at most finitely many times.

2.3 Differential Privacy

DP is a mathematical framework designed to quantify the privacy leakage of a proposed randomized algorithm (called a mechanism), introduced by Dwork et al. (2006). The first step for measuring such privacy risk starts from specifying which databases are considered to "differ in one entry," which we refer to as *adjacent databases*. We say D and D' are adjacent if $d(D, D') \leq 1$, for some metric $d(\cdot, \cdot)$ between databases. In this paper, we use Hamming distance $H(\cdot, \cdot)$, which counts the number of entries that differ between D and D'. A privacy mechanism $\mathcal{M} : \mathcal{D}_n \to \mathcal{Y}$ returns a random variable $\mathcal{M}(D)$ for any $D \in \mathcal{D}_n$, and the privacy risk of the algorithm $\mathcal{M}$ can be evaluated by definition as follows:

Definition 1 (ϵ-Differential privacy (ϵ-DP): Dwork et al., 2006) *Given $\epsilon \geq 0$, a privacy mechanism $\mathcal{M}$ on the output space $\mathcal{Y}$ satisfies ϵ-DP if*

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

for every measurable set $S \subset \mathcal{Y}$ and all D and D' satisfying $H(D, D') \leq 1$.

This definition characterizes how much privacy leakage could occur via the privacy budget parameter ϵ when a single entity in D is not the same one in D'. The smaller that ϵ is, the harder it is to distinguish the probability distributions of $\mathcal{M}(D)$ and $\mathcal{M}(D')$, which accordingly makes it harder to identify whether data set D or D' was used in the analysis by $\mathcal{M}$ (Wasserman and Zhou, 2010).

For conceptual understanding, let us imagine that a data set D contains information of 100 people obtained by a survey. Let us call one person in D Person A, and let us assume that a ϵ-DP privacy mechanism $\mathcal{M}$ is employed so that a summary from $\mathcal{M}(D)$ is released, containing some useful information about D. Then, it is known that, for any other data set D' that shares 99 people with D, except for Person A, the probability distributions of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ cannot be easily distinguished. Thus, it is difficult to identify whether Person A is included in the actual data set D or not. Notice that Person A was chosen arbitrarily in D, so the privacy of each individual in D is protected.

While any ϵ-DP mechanism preserves privacy, not all mechanisms ensure good performance with respect to an underlying utility. It is straightforward to imagine that sanitized statistics can devastate the performance of the utility due to excessive noises for privacy. In contrast, the exponential mechanism is a general technique that ensures high utility while controlling the privacy leakage within the budget ϵ.

Proposition 2 (Exponential mechanism: McSherry and Talwar, 2007) *Let $n \in \mathbb{Z}_+$ and let $\{u_D : \mathcal{Y} \to \mathbb{R} : D \in \mathcal{D}_n\}$ be a collection of utility functions. Assume that the* sensitivity $\Delta(u)$ *is finite:*

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

where the supremum is over all adjacent D and D' and assume that $\int \exp(u_D(y))\, d\nu(y) < \infty$ for all $D \in \mathcal{D}$ where ν is a measure in $\mathcal{Y}$. If Δ satisfies $\Delta(u) \leq \Delta < \infty$, then the collection of mechanisms $\{\mathcal{M}(D) : D \in \mathcal{D}\}$, each of which has the probability density with respect to ν

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

satisfies ϵ-DP.

The exponential mechanism can be easily applied to a wide variety of problems having utility functions. One of the simplest examples is a count statistic. Let us define the count statistic count(D) of a data set D to be the number of data points in D having a certain property, and define a utility function $u_D(y) := -|y - \text{count}(D)|$, which has sensitivity 1. This utility puts higher values when y is close to count(D). Many machine learning and statistical inference problems can be privately handled using the exponential mechanism

when it is possible to define appropriate utility functions such as empirical risk or likelihood functions (Huang and Kannan, 2012; Awan et al., 2019; Cummings et al., 2019; Lu et al., 2022).

Proposition 3 (Utility of the exponential mechanism: Dwork and Roth, 2014) *Let* $\mathrm{OPT}_D = \max_{y\in\mathcal{Y}} u_D(y)$ *be the optimal value that can be achieved by the utility function over all outputs, given database* D*. Let* Y *be a random variable with the density given in (2.4). Then,*

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

for every $t \ge 0$*. Consequently,*

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

The utility function in the exponential mechanism must be carefully chosen to ensure that the error rate given in Proposition 3 translates to optimal rates for the private output. For example, Awan et al. (2019) showed that when the utility function has a quadratic Taylor expansion at its maximum, the randomness for privacy in the exponential mechanism often gives rise to $O_p(1/\sqrt{n})$ noise, which in general is of the same order as the non-private statistical estimation problems. On the other hand, Reimherr and Awan (2019) showed that for some utility functions which are locally approximated by the absolute value function, the randomness for privacy may be as low as $O_p(1/n)$.

With the goal of producing differentially private persistence diagrams, we propose using exponential mechanism whose utility function is the negative bottleneck distance between the private and non-private persistence diagrams.

3. Sensitivity of Persistence Diagrams in the Bottleneck Distance

Most DP algorithms require quantifying how much the value of a statistic is changed by changing a single point in a given data. The largest possible amount of that change in the statistic is called the sensitivity of the statistic. In this study, we regard a persistence diagram constructed from a data set D as a statistic that estimates the homological structure of the space underlying the data, and we use the bottleneck distance to define a metric on the space of persistence diagrams. Hence, to apply a DP mechanism to persistence diagrams, our first step should be estimating the sensitivity of persistence diagrams in terms of the bottleneck distance; namely, we are going to analyze how big the bottleneck distance $\mathrm{d_B}(\mathcal{P}_D, \mathcal{P}_{D'})$ can be, where the pair (D, D') denotes a pair of adjacent data sets. Note $\mathcal{P}_D$ and $\mathcal{P}_{D'}$ mean the persistence diagrams constructed from the data sets D and D' respectively under a given way of constructing persistent homology.

In differential privacy, to ensure consistent estimators, it is necessary that the sensitivity goes to 0 as the size of the data grows. Our key observation is that if a chosen way of constructing persistent homology is not outlier-robust, the sensitivity of the corresponding persistence diagrams may not tend to 0 even if the size of data, say n, grows.

We demonstrate that the sensitivity of the persistence diagrams of Čech complexes cannot converge to 0 even if the size of data grows to infinity. To overcome such an issue, we

propose using the notion called distance to a measure (DTM), which was thoroughly discussed by Chazal et al. (2018) to give birth to outlier-robust persistence diagrams. Moreover, among various versions of construction of DTM, we propose using L^1-DTM which gives the smallest sensitivity.

Before moving on to the main sensitivity analysis, we would like to clarify our terminology. In the introduction to this section, we have been using the word sensitivity to refer to two different quantities: sensitivity of the bottleneck distance and the sensitivity of utility functions of the exponential mechanism. To avoid a confusion, going forward we refer to the sensitivity of the bottleneck distance of persistence diagrams as the *base sensitivity*, which is the terminology introduced in Awan and Wang (2022). The base sensitivity of the bottleneck distance of qth persistence diagrams from Čech complexes is denoted by $\Delta_q^{\check{\text{C}}\text{ech}}$ and that from the L^1-DTM is denoted by Δ_q^{DTM}. The precise definition of them will be presented in each of the following subsections. Otherwise, going forward we will reserve the term "sensitivity" for the sensitivity of a given utility function of the exponential mechanism.

3.1 Sensitivity of the Persistence Diagrams of Čech Complexes

Let us illustrate how the construction of Čech complexes fails to have a decreasing base sensitivity. The situation of the following example is well illustrated in Figure 3. Note that Figure 3 draws figures by means of the Vietoris-Rips complex instead of the Čech complex. The Vietoris-Rips complex is a variant of the Čech complex which has a computational advantage than the Čech complex. In fact, the filtration of Vietoris-Rips complexes has essentially the same information with that of Čech complexes. The definition of the Vietoris-Rips complex and its relationship with the Čech complex is presented in Appendix A.3.

Example 1 *Let D be a set of n points in $\mathbb{R}^2$ that is tightly clustered into exactly two clusters. Write x to denote the point located at the midpoint of the clusters, and take D' to be the data set obtained by moving one point in D to x. Now, further imagine that n grows while the configuration of the points in D and D' remains the same, and derive the 0th dimensional persistence diagrams obtained from the Čech complexes of D and D'. Then, the connected components in D collapse into the two clusters quickly, while the isolated point $x \in D'$ produces an additional connected component that lives longer. Such a discrepancy between two persistence diagrams prohibits the bottleneck distance between them from going to 0. More precisely, the bottleneck distance between them remains as big as the distance of the point x from the clusters in D.*

The following lemma establishes that this phenomenon is widespread. We denote the qth persistence diagram constructed from the Čech complexes on the data set D by $\mathcal{P}_q^{\check{\text{C}}\text{ech}}(D)$.

Lemma 4 *Let $D = \{x_1, \ldots, x_n\}$ be a subset of an Euclidean space $\mathbb{R}^d$. Let $\{d_1, \ldots, d_m\}$ be the set of distinct finite death times in $\mathcal{P}_0^{\check{\text{C}}\text{ech}}(D)$ with $0 < d_1 < \cdots < d_m < \infty$. Let $\delta = d_m - d_{m-1}$ (if $m = 1$, let $\delta = d_1$). Then, it is possible to take a set D' with $H(D, D') = 1$ satisfying that*

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

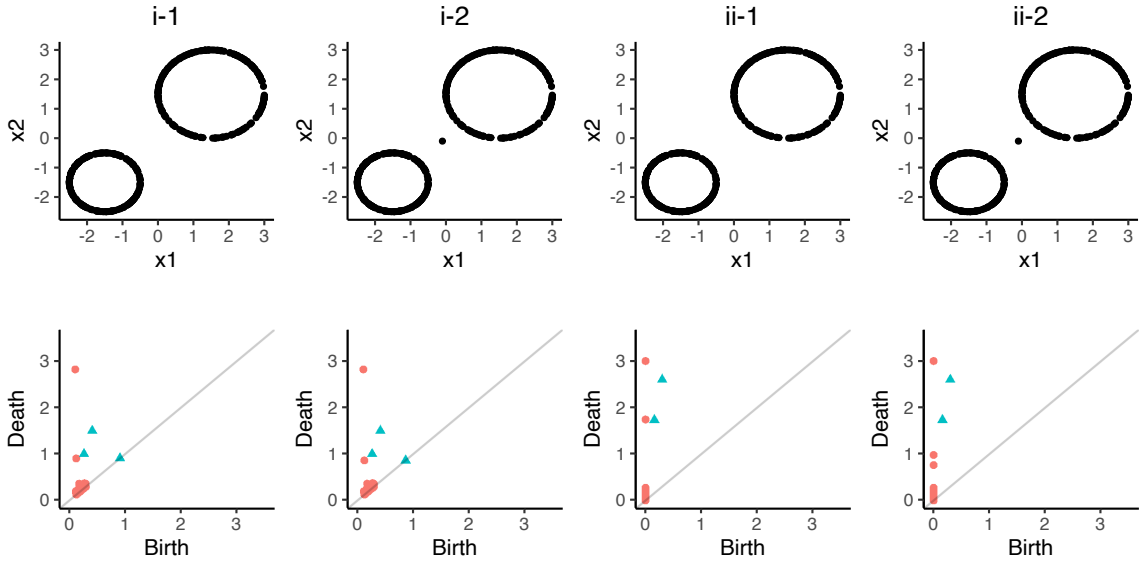


Figure 3: Persistence diagrams on D and D': the red circles and the green triangles are the connected components and the loops respectively. The columns (i-1) and (i-2) correspond to the results with L^1-DTM on D and D' respectively, and the two diagrams have 0.042 bottleneck distance in terms of the connected components. The columns (ii-1) and (ii-2) are from the Vietoris-Rips complex and the distance between the two diagrams have 0.762.

Roughly, the lemma can be proved by constructing a data set D' having an additional point at the middle of the most "significant connected components" in the filtration of Čech complexes of D, i.e., the connected components that die at time d_m. The detailed proof is presented in Appendix B.1.

From now on, we assume that all the data sets are supported in a bounded subset E of $\mathbb{R}^d$ unless there is any additional specification. We define the base sensitivity $\Delta_q^{\check{\text{C}}\text{ech}}$ concerning Čech complexes:

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

Note that the stability theorem (2.1) implies the following upper bound of the base sensitivity:

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

for every non-negative integer q. Lemma 4 provides the matching lower bound of the base sensitivity for $q = 0$. Moreover such upper and lower bounds show that the sensitivity of the utility function v_D defined as

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

has sensitivity of constant order:

Theorem 5 *Suppose that a given data-generating process is supported on a bounded subset E of a Euclidean space. Then, we have*

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

Moreover, the utility function v_D defined in (3.1) satisfies

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

Theorem 5 shows that why it is challenging to develop a privacy mechanism for Čech complexes: Čech complexes are so sensitive, in terms of the bottleneck distance of their persistence diagrams, that the sensitivity of the utility function $v_D(\cdot)$ remains constant regardless of the size n of the data set. This implies that the exponential mechanism using this utility function keeps adding a constant size of noise even if n gets bigger. This prevents the bottleneck distance from becoming small even in the case of huge n.

3.2 Sensitivity of the Persistence Diagrams of the DTM

The DTM, which was introduced by Chazal et al. (2011), provided a novel way to overcome the sensitivity to outliers. DTM proposes measuring how far each point is from the dense part of the support of the probability measure. By doing so, an outlier corresponds to a relatively large distance. Thus, when it comes to the filtration of the sub-level sets of a DTM, the topological features produced by the outlier would occur in the late period of the filtration, or it might not occur through the whole filtration. More thorough discussions on the DTM can be found in Chazal et al. (2018); Anai et al. (2020); Oudot (2015). By virtue of the properties of the DTM, it is very likely that DTM-based persistence diagrams give rise to a much smaller sensitivity, so it may provide us with a suitable TDA statistic to build our privatized mechanism upon. In fact, we show that DTM-based persistence diagrams achieve sensitivity converging to 0 as n grows to infinity, but the rate of decay depends on which class of DTM designs we use.

Basically, a DTM is defined to be a L^p norm of a certain function. The original version of DTM was defined to be a L^2-type quantity. We show that the L^2-type DTM produces persistence diagrams whose base sensitivity is bounded by $O(n^{-1/2})$, and we recognize that each L^p-DTM results in an analogous upper bound of the base sensitivity: $O(n^{-1/p})$. From this observation, we focus on the L^1-DTM that has the fastest decay rate in the base sensitivity. Furthermore, we also verify the base sensitivity of the persistence diagrams obtained from the L^1-DTM is bounded below by n^{-1} up to a constant. In other words, our sensitivity analysis for L^1-DTM is sharp up to constants.

We present the definition of the general L^p-DTM and its empirical realization. The key property to obtain upper bounds of the persistence diagrams is the so-called Wasserstein stability of a DTM, which was extensively discussed in the past literature; for instance, see Chazal et al. (2016b). As a result of the Wasserstein stability, we deduce the upper bound of rate $n^{-1/p}$ for the L^p-DTM. The matching lower bound of rate n^{-1} for the L^1-DTM is established by constructing a specific example that exactly gives the lower bound. All the proofs are presented in the appendix.

Definition 6 (Distance to a measure) *Let μ be a probability measure and X be a random variable whose probability distribution is μ. For the given μ, $0 < m < 1$, and $p \geq 1$, the L^p distance to the measure μ at resolution m is defined by*

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

where $G_x(t) = P\big[\parallel X - x \parallel \leq t\big]$. Here, $\parallel \cdot \parallel$ denotes the ℓ^2-norm in Euclidean spaces.

The hyperparameter m determines how much smoothing effect will be employed, which is reminiscent of the role of the bandwidth in a kernel density estimation. A natural empirical approximation would be the following:

Definition 7 (Empirical version of the DTM) *Let $X_1, \ldots, X_n$ be i.i.d. samples obtained from a probability distribution μ and μ_n the empirical probability measure defined on this sample, i.e.,*

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

The empirical L^p-DTM to μ at resolution m, denoted by $\hat{\delta}^{(p)}$, is defined to be the L^p-DTM to μ_n at resolution m; namely,

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

where $k = \lceil mn \rceil$ and $N_k(x)$ is the set containing the k nearest neighbors of x among $X_1, \ldots, X_n$. Here, the distance between data points is measured by the ℓ^2-norm in Euclidean space.

The key quantitative property of the L^p-DTM, which is called its Wasserstein stability, is the following: let μ and ν be probability measures defined on a common metric space, then

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

where $W_p(\mu,\nu)$ denotes the p-Wasserstein distance between μ and ν. For more details, see Chazal et al. (2016b). Let D and D' be adjacent data sets.

Let $\mathcal{P}_q^{\mathrm{DTM}_p}(D)$ denote the qth persistence diagram constructed from the filtration of sub-level sets of the L^p-DTM to the empirical distribution of the data set D. The base sensitivity of $\Delta_q^{\mathrm{DTM}_p}$ concerning the DTM is

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

By virtue of the stability theorem (2.2) and the Wasserstein stability (3.1), the following upper bound of the base sensitivity $\Delta_q^{\mathrm{DTM}_p}$ can be established by quantifying the p-Wasserstein distance between empirical distributions on adjacent data sets.

Theorem 8 (Sensitivity of the persistence diagrams via L^p-DTM) *Let D and D' be finite subsets of a bounded set E in $\mathbb{R}^d$ satisfying $|D| = |D'| = n$ and $H(D, D') = 1$. Then, for every non-negative integer q,*

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

In fact, as a result of the Wasserstein stability (3.1), the result of the theorem can be obtained by estimating the p-Wasserstein distance between the empirical distributions on D and D'. The detailed proof is presented in Appendix B.1.

According to Theorem 8, each L^p-DTM is guaranteed to have base sensitivity bounded above by $O(n^{-1/p})$. In particular, such a guaranteed upper bound becomes smallest when p is taken to be 1:

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

In fact, the upper bound of the L^1-DTM is sharp up to constants.

Proposition 9 (Lower bound of the sensitivity of the L^1-DTM) *Suppose that $m < 1/2$. Then, for every positive integer n, there exists a pair of sets D and D' that satisfies $|D| = |D'| = n$, $H(D, D') = 1$, and*

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

where $k = \lceil mn \rceil$.

The proof can be obtained by constructing a pair of adjacent data sets D and D' that achieve the proposed distance. In fact, the data sets illustrated in Figure 3 achieve it. For a detailed proof, see Appendix B.1.

Now, we introduce the utility function that we use to design our privacy mechanism. Let Pers denote the space of persistence diagrams, equipped with the bottleneck distance. For any given data set D and any non-negative integer q, we define the function $u_D^{(q)} : \mathsf{Pers} \to \mathbb{R}$ as follows:

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

Let ℓ be a chosen non-negative integer. Our utility function $u_D : \left(\mathsf{Pers}\right)^{\ell+1} \to \mathbb{R}$ is defined as follows:

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

As a result of the upper and lower bounds for the base sensitivity, we can establish the following upper and lower bounds of the sensitivity:

Corollary 10 *For a chosen $\ell \geq 0$, let the utility function u_D be defined as in (3.2). Then the following is satisfied:*

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

Remark 11 *The additive nature of the utility function u_D is what allows us to establish the upper and lower bounds in Corollary 10. Notice that the lower bound of the corollary is derived from the result of Proposition 9 which is only valid for the 0th persistence diagrams; but, the additive form of u_D allows it to be a lower bound for the sensitivity of the entire utility function.*

Remark 12 *Note that while the lower bound of L^1-DTM matches the rate of its upper bound, we do not at this time obtain such matching lower bounds of the other L^p-DTMs. Hence, it might be the case that the base sensitivity of the general L^p-DTM can be improved. For example, in the situation of Figure 3, we found empirical evidence that the bottleneck distance between L^2-DTM persistence diagrams of D and D' is also $O(n^{-1})$.*

4. Employment of the Exponential Mechanism with the L^1-DTM

In this section, we describe how to implement the exponential mechanism in order to generate privatized persistence diagrams.

Let $D = \{X_1, \ldots, X_n\}$ be a data set that consists of i.i.d. samples having a common probability distribution μ. For brevity, we denote by $\mathcal{P}_q(D)$ for each $q \geq 0$ the qth persistence diagram obtained from the L^1-DTM to the empirical measure μ_n, which was denoted by $\mathcal{P}_q^{\mathrm{DTM}_1}(D)$ in the previous section. And, we set $\mathcal{P}(D) := (\mathcal{P}_0(D), \ldots, \mathcal{P}_\ell(D))$ for the given ℓ. Let $\mathcal{P}_q(\mu)$ be the qth persistence diagram obtained from the L^1-DTM $\delta_\mu^{(1)}$ to the measure μ, and let us define $\mathcal{P}(\mu) := (\mathcal{P}_0(\mu), \ldots, \mathcal{P}_\ell(\mu))$ for the given ℓ. Also, let $\mathcal{P}_{\mathrm{DP}} = (\mathcal{P}_{0,\mathrm{DP}}, \ldots, \mathcal{P}_{\ell,\mathrm{DP}})$ be a tuple of privatized persistence diagrams generated from our algorithm.

We analyze the error of our privatized persistence diagrams from two different points of view. First, we estimate $\mathrm{d_B}(\mathcal{P}(\mu), \mathcal{P}_{\mathrm{DP}})$. This quantity represents the error of the privatized persistence diagrams from the persistence diagram of the original data-generating process. From a statistical perspective, $\mathcal{P}(\mu)$ can be regarded as a parameter characterizing the true data-generating process. Hence, the quantity $\mathrm{d_B}(\mathcal{P}(\mu), \mathcal{P}_{\mathrm{DP}})$ quantifies the amount of error in estimating the parameter $\mathcal{P}(\mu)$ by the privatized statistic $\mathcal{P}_{\mathrm{DP}}$ that is obtained by privatizing the actual statistic $\mathcal{P}(D)$. The second approach is to estimate the quantity $\mathrm{d_B}(\mathcal{P}(D), \mathcal{P}_{\mathrm{DP}})$ which quantifies how much the privatization process distorts the original non-privatized statistic.

4.1 Generating Privatized Persistence Diagrams

The design of an exponential mechanism is formulated by specifying an output space $\mathcal{Y}$ and a utility function $u_D : \mathcal{Y} \to \mathbb{R}$ for each data set D. Since our target to be privatized is a persistence diagram, it would be a natural choice to take the space of all possible persistence diagrams, which we denoted by Pers, as the output space. The first candidate for the utility function would be the function u_D defined in (3.2). However, its output space, Pers, contains persistence diagrams which have arbitrary many numbers of birth-death pairs; that is, to sample persistence diagrams from the whole Pers is inevitably an infinite-dimensional problem, which is technically difficult in computation. To bypass such an issue, we pre-specify a hyperparameter $M \in \mathbb{Z}_+$, a positive integer, and only take care of the space Pers_M of persistence diagrams having at most M birth-death pairs at each

dimension q. On each restricted space Pers_M, for any given data set D, we re-define the function $u_D^{(q)} : \mathsf{Pers}_M \to \mathbb{R}$ as follows:

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

The utility function u_D is also re-defined in the same way as in (3.2) by using the re-defined $u_D^{(q)}$s. Namely, the utility function $u_D : \big(\mathsf{Pers}_M\big)^{\ell+1} \to \mathbb{R}$ is defined as

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

Note that the upper and lower bounds established in Corollary 10 are still valid for the utility u_D defined in (4.1).

Under the choice of the utility function u_D in (4.1), the probability distribution from which privatized persistence diagrams are generated can be specified. Before describing the exponential mechanism, we introduce a discretized version of Pers_M that will make the analysis in Section 4.2 convenient. Note that each persistence diagram in Pers_M can be viewed as a family of at most M points in the upper-left triangle $\bar{\mathcal{T}} := \{(x, y) \,:\, 0 \le x \le y \le \mathrm{diam}E\}$. Instead of using $\bar{\mathcal{T}}$ directly, we discretize it with finitely many discrete points; for example, a set of equally-spaced finitely many points in $\bar{\mathcal{T}}$ can be a such discretization. By discretizing the set $\bar{\mathcal{T}}$ with N^2 discrete points, a discretization of Pers_M can be obtained; namely, the discretized version of Pers_M is the family of sets having at most M points in the discretized version of $\bar{\mathcal{T}}$. Note that such a discretization of Pers_M has cardinality at most N^{2M}. For a given positive integer N, we define $\widetilde{\mathsf{Pers}}_{M,N}$ to be the discretization of Pers_M obtained by discretizing $\bar{\mathcal{T}}$ with N^2 equally spaced discrete points. Therefore, our exponential mechanism is indeed carried out on the space $\widetilde{\mathsf{Pers}}_{M,N}$. The space $\widetilde{\mathsf{Pers}}_{M,N}$ is the actual output space where the private persistence diagrams generated by the following mechanism live.

Proposition 13 *Let $\ell \ge 0$ be fixed and the utility u_D defined in (4.1). Set $p(\cdot)$ to denote the probability density function characterized by the following expression:*

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

with respect to the uniform distribution on the set $(\widetilde{\mathsf{Pers}}_{M,N})^{\ell+1}$ as the base measure. In (4.2), Δ is defined as follows:

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

and $\mathcal{P}_{\mathrm{DP}} = (\mathcal{P}_{0,\mathrm{DP}}, \ldots, \mathcal{P}_{\ell,\mathrm{DP}})$. Then, the exponential mechanism characterized by the density (4.2) satisfies ϵ-DP.

Remark 14 *Note that the discretization is not necessary to establish Proposition 13, but is needed to derive the utility results in the following section, such as Proposition 15 and Theorem 17. It is possible that this discretization can be removed from our analysis using more sophisticated techniques.*

4.2 Analysis of Privatized Persistence Diagrams

Let $\ell \geq 0$ be determined. Recall that we have restricted the output space of our privatized persistence diagram in terms of M points for each dimension, and that these fall on a discretized version of the continuous persistence diagram space. To incorporate these limitations into our consideration for the error quantification, we define $\mathcal{P}_{\mathrm{OPT}}$ to be the closest persistence diagram from $\mathcal{P}(D) = (\mathcal{P}_0(D), \ldots, \mathcal{P}_\ell(D))$ that can be generated from the privacy algorithm. More precisely, for each q

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

where $\mathcal{P}$ ranges over all persistence diagrams having at most M elements, and

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

Similarly, the counterpart of $\mathcal{P}_{\mathrm{OPT}}$ on the discrete space $\widetilde{\mathsf{Pers}}_{M,N}$ is defined as follows. For every $q \geq 0$,

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

and

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

Moreover, for any pair of $(\ell+1)$-tuples of persistence diagrams $\mathcal{P} = (\mathcal{P}_0, \ldots, \mathcal{P}_\ell)$ and $\mathcal{P}' = (\mathcal{P}'_0, \ldots, \mathcal{P}'_\ell)$, we define

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

In general, in the literature on the exponential mechanism, there have been broad analyses with regard to the error-minimizing value in the space covered by the exponential mechanism. For instance, see Dwork and Roth (2014). One key result concerning $\mathcal{P}_{q,\mathrm{OPT}}$ is summarized in Proposition 2.5. Consequently, the following estimate can be established. Recall that the discretized space $\widetilde{\mathsf{Pers}}_{M,N}$ has been obtained by discretizing the upper-left triangle $\bar{\mathcal{T}}$ with N^2 equally-spaced discrete points.

Proposition 15 *Let $\mathcal{P}_{\mathrm{OPT}}$ be defined in the above and $\mathcal{P}_{\mathrm{DP}}$ the private persistence diagram obtained from the exponential mechanism summarized in Section 4.1. Suppose that the upper-left triangle $\bar{\mathcal{T}}$ is discretized into N^2 equally spaced points. Then the following holds:*

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

In particular, if we take $N = n$, it holds that

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

Remark 16 *In fact, the exponential mechanism itself only directly guarantees that privatized diagrams are concentrated at the optimal diagram in the discretized space. More specifically, we have*

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

On the other hand, as long as we employ fine enough discretization, it is trivial that the distance $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \widetilde{\mathcal{P}}_{\mathrm{OPT}})$ *is negligible compared to the error (4.3). For instance, taking* $N = N(n) = n$ *ensures that such an approximation error has order* $O_p(n^{-1})$ *and the term* $\log N$ *in (4.3) only adds* $\log n$ *amount of error. This guarantees the result in Proposition 15.*

To take advantage of the above result, we can estimate each of the two types of errors as follows.

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

and

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

Hence, the remaining part is to estimate $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(\mu))$ and $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(D))$, respectively.

Before stating our main theorem in this section, we would like to summarize the terminology that we use to call each of the error terms we are concerned with. First of all, we call $\mathrm{d_B}(\mathcal{P}(D), \mathcal{P}(\mu))$ the estimation error because $\mathcal{P}(D)$ can be viewed as a statistic estimating $\mathcal{P}(\mu)$. The term $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}_{\mathrm{OPT}})$ is called the privacy error, following the tradition in DP literature. On the other hand, we call the quantity $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(D))$ the approximation error because $\mathcal{P}_{\mathrm{OPT}}$ is the best approximation of $\mathcal{P}(D)$ in the space Pers_M. In contrast with the previous two terms, a type of quantity of the form $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(\mu))$ has not been analyzed in the literature before to our knowledge. As we noted, $\mathcal{P}(\mu)$ can be regarded as a population parameter describing the probability measure μ generating the data D. Concerning this perspective, we call the quantity $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(\mu))$ the constrained estimation error and call the corresponding quantity $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}(\mu))$ the privacy-estimation error in order to allude that this quantity would be interpreted as the amount of error in estimating the population parameter $\mathcal{P}(\mu)$ by the privatized statistic $\mathcal{P}_{\mathrm{DP}}$.

If we can choose M large enough, both terms $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(\mu))$ and $\mathrm{d_B}(\mathcal{P}_{\mathrm{OPT}}, \mathcal{P}(D))$ can be estimated through the convergence of the empirical distribution on D to the measure μ in terms of the Wasserstein distance W_1 (See Proposition 25). It turns out that both terms are bounded by $O_p((\ell+1)n^{-1/d})$. Hence, by taking $M = M(n)$ to be a slowly increasing sequence we can achieve such a bound without degrading the privacy error obtained in Proposition 15.

Theorem 17 (Upper bound for the privacy-estimation error) *Set* $M(n) = \log n$ *and* $N(n) = n$*. Then, for all large enough* n*, the following estimate holds:*

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

It is natural to wonder how sharp the obtained upper bounds are. As for the population-estimation error (and the estimation error), unfortunately, it is inevitable to get the rate $n^{-1/d}$ so long as the argument relies on the Wasserstein convergence of the empirical measure on D to the measure μ. This means that if a tighter rate is possible, it is required to use a different approach in order to examine the birth-and-death of homological features of the sub-level sets of the DTM more precisely. In the literature of TDA, there are some approaches that examined such features of Čech complexes by employing some toolkits from geometry. For example, see Bobrowski and Adler (2014). Such approaches may hint how to analyze persistence diagrams of the DTM more precisely.

As for the privacy error, we argue that it is sharp up to constants and logarithmic factors. Recall that $\mathcal{P}_{\mathrm{OPT}}$ is defined to be the persistence diagram in the range of our privacy algorithm which has the smallest distance from $\mathcal{P}(D)$. This definition lets us surmise that the distance $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}_{\mathrm{OPT}})$ could be smaller than the distance $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}(D))$ in a considerable probability. This means that if we are able to find a lower bound of $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}(D))$ matching the upper bound of $\mathrm{d_B}(\mathcal{P}_{\mathrm{DP}}, \mathcal{P}_{\mathrm{OPT}})$, it underpins that our estimate could be sharp. In the following theorem, we prove that, under some mild conditions, there is no ϵ-DP mechanism whose privacy error with respect to the persistence diagrams from L^1-DTM can be smaller than $1/(n\epsilon)$. For the following, we recall that $\mathcal{P}_0(D)$ denotes the 0th persistence diagram obtained by the L^1-DTM to the empirical measure on a given data set, as defined before.

Theorem 18 *Suppose that $m < 1/2$. Let n be a positive integer and $\mathcal{M}$ an arbitrary ϵ-DP mechanism that produces a privatized persistence diagram $\mathcal{M}(D)$ of a data set D. Assume that ϵ satisfies $1/n \le \epsilon \le 1$. Then it is not possible for $\mathcal{M}$ to achieve $\mathrm{d_B}(\mathcal{P}_0(D), \mathcal{M}(D)) = o_p\big(\frac{1}{n\epsilon}\big)$ for every database D with $|D| = n$.*

5. Simulation Study

In the following simulation studies and the real-world data analysis, we only consider the 0th and the 1st persistence diagrams; namely, the utility we use is given by taking $\ell = 1$, i.e., we set $u_D : \big(\mathsf{Pers}_M\big)^2 \to \mathbb{R}$ by

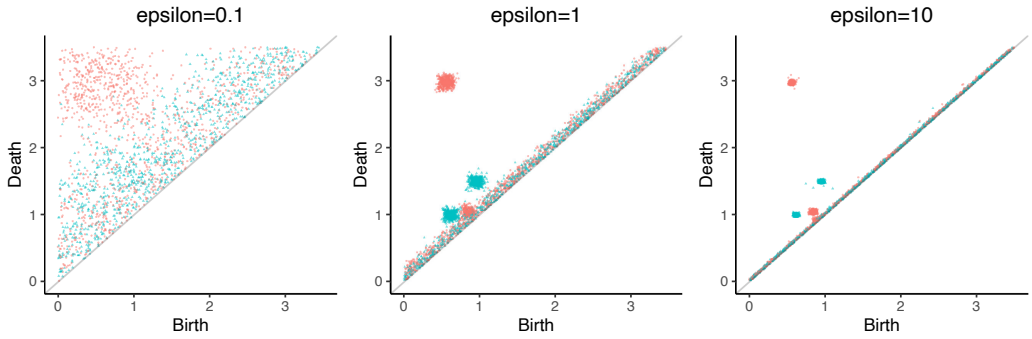
$$u_D(\mathcal{P}_0, \mathcal{P}_1) := u_D^{(0)}(\mathcal{P}_0) + u_D^{(1)}(\mathcal{P}_1).$$

The purpose of such restriction is only to present our algorithm succinctly; the algorithm can readily be extended to take the higher-dimensional features into consideration.

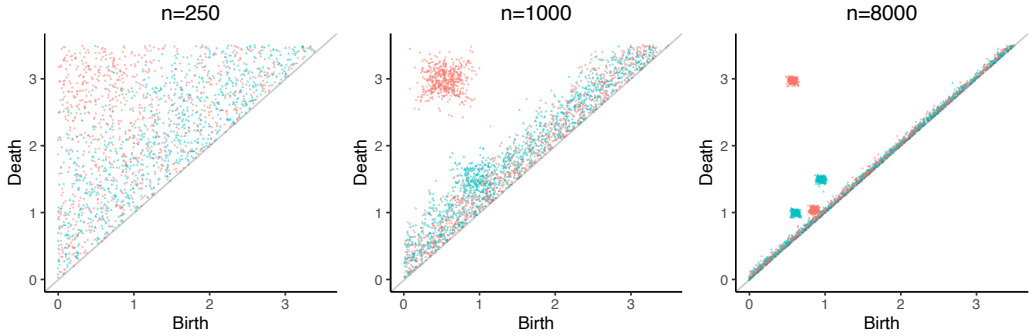
We produce the differentially private persistence diagrams and investigate the impact of the key hyper-parameters: the privacy budget ϵ and the sample size n, where the resolution of the DTM m is set to 0.2. For the exponential mechanism, the default parameters are $\epsilon = 1$, $m = 0.2$, $n = 4000$, and $M = 5$. These hyper-parameters were chosen by preliminary simulations. To sample from the exponential mechanism, we use a Markov chain Monte Carlo algorithm, specified in Appendix C.1. We choose the last iterate out of $T = 10000$ Monte Carlo diagrams as the reporting privatized diagram to be used for analysis*.

The simulation is based on the example in Figure 3, which has two circles at the origin $(1.5, 1.5)$ and $(-1.5, -1.5)$ whose radii are 1.5 and 1 respectively. Each circle consists of 200

*The `R` code is available at `https://github.com/jwsohn612/DPTDA`.



(a) Privatized persistence diagrams of different ϵ where $n = 4000$.



(b) Privatized persistence diagrams of different n where $\epsilon = 1$.

Figure 4: Privatized diagrams over 500 replicated data sets as described in Section 5: 0-dimensional connected components (orange) and 1-dimensional loops (green).

observations of uniformly generated points along the boundary of the circle. There is one more point in the middle of the circles for (i-2) and (ii-2). When inducing the Vietoris-Rips diagrams, the maximum filtration scale is specified as 3. All analyses are based on 500 sampled data sets, and we apply the privacy mechanism for each replication.

Figure 4 illustrates the outputs of the exponential mechanism as ϵ and n are varied. To reflect the variation of diagrams, we consider 500 replicated data sets. Our exponential algorithm is independently applied to each data set, and the algorithm reports the final diagram only. By repeating this procedure for all 500 replicates, we obtain the 500 reported private diagrams. Each panel in Figure 4 is drawn based on the 500 private diagrams that illustrate the shape of private diagrams' distribution. As expected, the variability in the privatized persistence diagrams becomes smaller as either ϵ or n becomes larger.

The overall tendency in terms of the bottleneck distance is exhibited in Figure 5. Note that the x-axis is written in the log scale. Gray areas in the panel show 95% point-wise quantile intervals of the bottleneck distance between the non-private diagram and its private one. Figure 5 depicts that both ϵ and n in the log scale have an approximately linear relationship to the log-bottleneck distance as shaded areas decently contain the red dotted lines e.g., $\log \mathrm{d_B}(\mathcal{P}(D), \cdot) \approx -\log n + c$ with some constant c. These results heuristically support that $d_B((P)_{DP}, \mathcal{P}(D)) = \tilde{O}_p(1/(n\epsilon))$ (considering ℓ to be fixed).

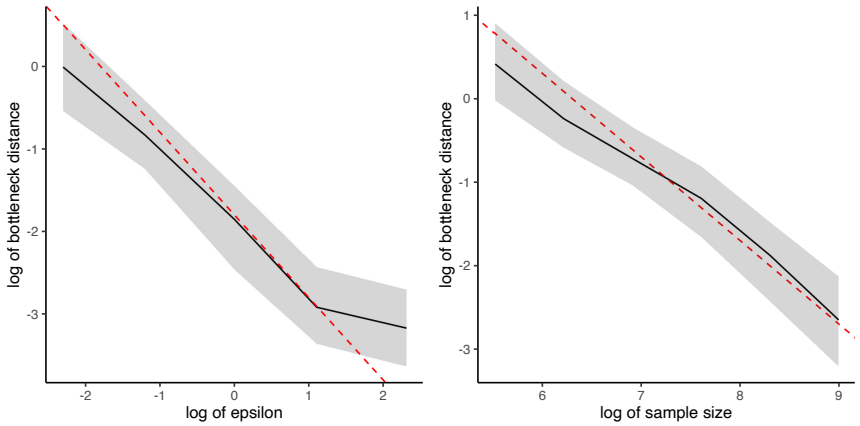


Figure 5: The 95% quantile intervals of $\mathrm{d_B}$ between $\mathcal{P}(D)$ and the corresponding private diagram over 500 replicates where ϵ (left) and n (right) increase respectively. Red dotted lines captured overall in the shaded areas have -1 slopes.

6. Real Data Analysis

In this section, we apply our mechanism to a real-world data set[†], which tracks the locations of three people walking around within a building, recorded on smartphones. We are going to call those people Walker A, B, and C. The 3-dimensional coordinates (x, y, z) of the location of each person were measured 20000 times over time so that the data set consists of 20000 location vectors (x, y, z) for each of Walker A, B, and C.

We calculate the persistence diagram corresponding to each of the walkers and apply our mechanism in order to privatize it. We would like to remark that we are not concerned with the privacy of individual walkers, but we consider the privacy of an individual's time stamps when they change. If a particular walker's persistence diagram changes significantly in accordance with a change of location at a certain timestamp, then the location information could be exposed to a risk of privacy leakage. Our privacy mechanism retains the topological structure of each walker's travel while protecting the information associated with each timestamp.

To obtain a privatized diagram, we carry out 50000 iterations in the MCMC procedure in our mechanism; the persistence diagram obtained at the last iteration is proposed as the reporting privatized diagram. We set the size of the sampling space $M = 5$, the resolution of the L^1-DTM $m = 0.05$, the privacy budget $\epsilon = 1$.

Figure 7 depicts the results of comparing the L^1-DTM persistence diagram corresponding to Walker C and its privatized diagram. One can see the diagrams look quite similar. In fact, the bottleneck distances $\mathrm{d_B}(\mathcal{P}_0(D), \mathcal{P}_{0,\mathrm{DP}})$ and $\mathrm{d_B}(\mathcal{P}_1(D), \mathcal{P}_{1,\mathrm{DP}})$ are both 0.01, which supports that our mechanism achieves high accuracy. Note that points near the diagonal are not considered significant, and do not substantially affect the bottleneck distance. In the right plot of Figure 7, we see that the bottleneck distance converges to a region $< .025$, and that the Markov chain seems to have converged after ≈ 5000 iterations.

[†]Data is provided at `http://bertrand.michel.perso.math.cnrs.fr/Enseignements/TDA/Tuto-Part1.html`

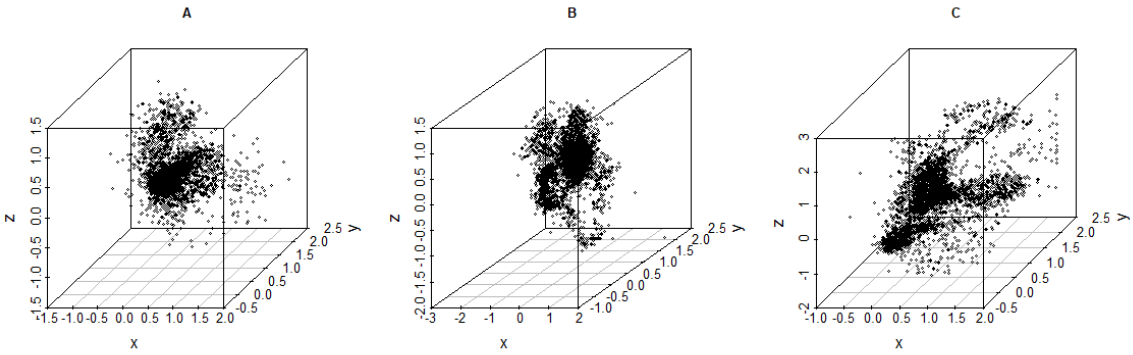


Figure 6: Scatter plots of the location information of Walker A,B, and C.

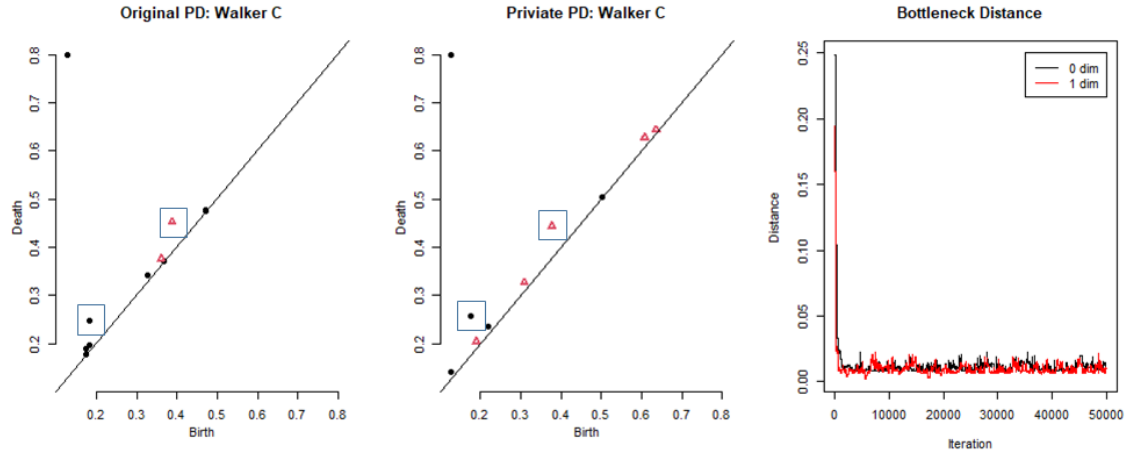


Figure 7: (Left) The true persistence diagram of Walker C; (Middle) a privatized persistence diagram of Walker C at the last iteration of a MCMC procedure, (Right) the bottleneck distances $\mathrm{d_B}(\mathcal{P}_0(D), \mathcal{P}_{0,\mathrm{DP}})$ (Black) and $\mathrm{d_B}(\mathcal{P}_1(D), \mathcal{P}_{1,\mathrm{DP}})$ (Red) of the true and a privatized diagram over MCMC iterations.

7. Discussion

In this paper, we propose the first mechanism for producing a differentially private persistence diagram, while highlighting the role of outlier-robustness in the sensitivity analysis. Even though our results offer significant contributions to private TDA, as well as a general understanding of the robust TDA measures, there are still some important weaknesses of our work as well as directions for future work:

As noted in other papers (e.g., Minami et al., 2016; Ganesh and Talwar, 2020; Seeman et al., 2021; Awan and Rao, 2023), MCMC implementations of the exponential mechanism can incur additional privacy costs, which should be incorporated into the analysis for more rigorous studies. The proposed techniques in the above papers could be applied to our instance of the exponential mechanism to ensure that the MCMC approximation is taken into account for end-to-end privacy protection. Another approach would be to directly sample the exponential mechanism on the discrete space of $O(N^{2(\ell+1)})$ points, but this could be computationally prohibitive with large N and ℓ.

While in this paper we recommended choosing M, the number of components in the persistence diagram, to be either a sufficiently large constant or an increasing function such as $M = \log n$, one could also consider M to be an unknown quantity that also needs to be learned in a private manner. As one of the anonymous reviewers suggested, one may be able to develop a reversible jump MCMC algorithm to sample from the exponential mechanism in this case. Some challenges of this approach would be 1) developing a base measure over the infinite-dimensional space $\prod_{M=1}^{\infty}(\widetilde{\mathsf{Pers}}_{M,N})^{\ell+1}$, which ensures that the exponential mechanism results in a valid probability distribution, 2) determine a reversible jump rule that allows for conversion between the dimensions, and 3) perform a customized utility analysis of this new mechanism. We leave it to future researchers to explore this direction.

On the side of TDA, we would like to mention that some other outlier-robust TDA methods could be discussed for privacy protection. For instance, a kernel distance which was also discussed by Chazal et al. (2018) may be a good candidate.

Besides TDA, the overall framework of how we propose a privacy mechanism can be applied to any other statistics that take their values in a metric space. A utility function concerned with a metric space-valued statistic can be defined similarly as we do with a persistence diagram and the bottleneck distance; this was already recognized in Dwork et al. (2006). However, the utility analysis for each different problem requires unique understanding of the specific structure and properties of the statistic and metric space at hand.

A theoretical limitation of our method is its scaling with the number of dimensions, denoted by $\ell + 1$, we consider. It is well known that the error in ϵ-DP mechanisms typically scales linearly with the dimension, and our instance of the exponential mechanism is no different. Since this is a limitation of ϵ-DP, it can only be addressed by using a different privacy framework. Future researchers may consider developing DP-TDA methods in the frameworks of approximate DP (Dwork and Roth, 2014), zero-concentrated DP (Bun and Steinke, 2016) or Gaussian DP (Dong et al., 2022), which often allow for the magnitude of the privacy noise to scale only in the square-root of the dimension.

Another limitation of our work is that our utility analysis depends on an artificial discretization of the persistence diagram space. This limitation is caused by the use of Proposition 3, and could be potentially addressed by using more advanced techniques.

Finally, Dong et al. (2020) proposed an alternative quantity to sensitivity for the exponential mechanism, which may improve the finite-sample performance.

Acknowledgments

The authors are thankful to the anonymous reviewers for their helpful comments that significantly improved the presentation of this manuscript. Taegyu Kang's research was partially supported by the AFOSR grant FA9550-22-0238 at Purdue University. Sehwan Kim started working on this project while he was at Purdue University. Jordan Awan's research is supported in part by NSF grant SES-2150615.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

Hirokazu Anai, Frédéric Chazal, Marc Glisse, Yuichi Ike, Hiroya Inakoshi, Raphaël Tinarrage, and Yuhei Umeda. DTM-based filtrations. In *Topological Data Analysis*, volume 15, pages 33–66. Springer, 2020.

Marco Avella-Medina. Privacy-preserving parametric inference: a case for robust statistics. *Journal of the American Statistical Association*, 116(534):969–983, 2021.

Jordan Awan and Vinayak Rao. Privacy-aware rejection sampling. *Journal of Machine Learning Research*, 24(74):1–32, 2023.

Jordan Awan and Yue Wang. Differentially private Kolmogorov-Smirnov-type tests. arXiv:2208.06236, 2022.

Jordan Awan, Ana Kenney, Matthew Reimherr, and Aleksandra Slavković. Benefits and pitfalls of the exponential mechanism with applications to Hilbert spaces and functional pca. In *International Conference on Machine Learning*, pages 374–384. PMLR, 2019.

Leo Betthauser, Urszula Chajewska, Maurice Diesendruck, and Rohith Pesala. Discovering distribution shifts using latent space representations. arXiv:2202.02339, 2022.

Omer Bobrowski and Robert J. Adler. Distance functions, critical points, and the topology of random čech complexes. *Homology, Homotopy and Applications*, 16(2):311–344, 2014.

Glen E. Bredon. *Topology and geometry*. Springer-Verlag, New York, 1997.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658. Springer, 2016.

Luigi Caputi, Anna Pidnebesna, and Jaroslav Hlinka. Promises and pitfalls of topological data analysis for brain connectivity analysis. *NeuroImage*, 238:118245, 2021.

Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46 (2):255–308, 2009.

Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11:733–751, 2011.

Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. *The structure and stability of persistence modules*. Springer, 2016a.

Frédéric Chazal, Pascal Massart, and Bertrand Michel. Rates of convergence for robust geometric inference. *Electronic Journal of Statistics*, 10:2243–2286, 2016b.

Frédéric Chazal, Brittany Fasy, Fabrizio Lecci, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Robust topological inference: Distance to a measure and kernel distance. *Journal of Machine Learning Research*, 18:1–40, 2018.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete and Computational Geometry*, 37:103–120, 2007.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*, pages 309–315, 2019.

Meryll Dindin, Yuhei Umeda, and Frederic Chazal. Topological data analysis for arrhythmia detection through modular neural networks. In *Advances in Artificial Intelligence: 33rd Canadian Conference on Artificial Intelligence, Canadian AI 2020, Ottawa, ON, Canada, May 13–15, 2020, Proceedings 33*, pages 177–188. Springer, 2020.

Jinshuo Dong, David Durfee, and Ryan Rogers. Optimal differential privacy composition for exponential mechanisms. In *International Conference on Machine Learning*, pages 2597–2606. PMLR, 2020.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society Series B*, 84(1):3–37, 2022.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC'09—Proceedings of the 2009 ACM International Symposium on Theory of Computing*, pages 371–380, 2009.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Now Publishers Inc., 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

Herbert Edelsbrunner and John Harer. Persistent homology - a survey. In *Surveys on discrete and computational geometry*, Contemporary Mathematics. American Mathematical Society, 2008.

Herbert Edelsbrunner and John Harer. *Computational Topology : An Introduction*. American Mathematical Society, 2009.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(4):707–738, 2015.

Arun Ganesh and Kunal Talwar. Faster differentially private samplers via Rényi divergence analysis of discretized Langevin MCMC. *Advances in Neural Information Processing Systems*, 33:7222–7233, 2020.

Rob Hall, Alessandro Rinaldo, and Larry Wasserman. Differential privacy for functions and functional data. *The Journal of Machine Learning Research*, 14(1):703–727, 2013.

Jonathan Hehir, Siddharth Vishwanath, Aleksandra Slavković, and Xiaoyue Niu. Problems on random graphs under local differential privacy, 2022. Paper presented at JMM 2022.

Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:52, 2021.

Christoph Hofer, Roland Kwitt, Marc Niethammer, and Mandar Dixit. Connectivity-optimized representation learning via persistent homology. In *International conference on machine learning*, pages 2751–2760. PMLR, 2019.

Zhiyi Huang and Sampath Kannan. The exponential mechanism for social welfare: Private, truthful, and nearly optimal. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 140–149. IEEE, 2012.

Vishesh Karwa and Aleksandra Slavković. Inference using noisy degrees: Differentially private β-model and synthetic graphs. *The Annals of Statistics*, 44(1):87–112, 2016.

Vishesh Karwa, Pavel N Krivitsky, and Aleksandra B Slavković. Sharing social network data: differentially private estimation of exponential family random-graph models. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, pages 481–500, 2017.

Firas A Khasawneh and Elizabeth Munch. Chatter detection in turning using persistent homology. *Mechanical Systems and Signal Processing*, 70:527–541, 2016.

Hyekyoung Lee, Moo K Chung, Hyejin Kang, Bung-Nyun Kim, and Dong Soo Lee. Discriminative persistent homology of brain networks. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 841–844, 2011.

Ephy R. Love, Benjamin Filippenko, Vasileios Maroulas, and Gunnar Carlsson. Topological convolutional layers for deep learning. *Journal of Machine Learning Research*, 24:1–35, 2023.

Zhigang Lu, Hassan Jameel Asghar, Mohamed Ali Kaafar, Darren Webb, and Peter Dickinson. A differentially private framework for deep learning with convexified loss functions. *IEEE Transactions on Information Forensics and Security*, 17:2151–2165, 2022.

Melissa R. McGuirl, Alexandria Volkening, and Björn Sandstede. Topological data analysis of zebrafish patterns. *Proceedings of the National Academy of Sciences*, 117(10):5113–5124, 2020.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103, 2007.

Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. *Advances in Neural Information Processing Systems*, 29, 2016.

Ardalan Mirshani, Matthew Reimherr, and Aleksandra Slavković. Formal privacy for functional data with gaussian perturbations. In *International Conference on Machine Learning*, pages 4595–4604. PMLR, 2019.

James R. Munkres. *Elements of algebraic topology*. Addison-Wesley Publishing Company, 1984.

Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011.

Partha Niyogi, Stephen Smale, and Shumuel Weinberger. A topological view of unsupervised learning from noisy data. *SIAM Journal on Computing*, 40(3):646–663, 2011.

Steve Y. Oudot. *Persistence Theory: From Quiver Representations to Data Analysis*. American Mathematical Society, 2015.

Takashi Owada. Convergence of persistence diagram in the sparse regime. *The Annals of Applied Probability*, 32(6):4706–4736, 2022.

Matthew Reimherr and Jordan Awan. Kng: The k-norm gradient mechanism. *Advances in Neural Information Processing Systems*, 32, 2019.

Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nicholas Turk-Browne, and Smita Krishnaswamy. Uncovering the topology of time-varying fmri data using cubical persistence. *Advances in neural information processing systems*, 33:6900–6912, 2020.

Jeremy Seeman, Matthew Reimherr, and Aleksandra Slavković. Exact privacy guarantees for markov chain implementations of the exponential mechanism with artificial atoms. *Advances in Neural Information Processing Systems*, 34:13125–13136, 2021.

Daniel Shnier, Mircea Voineagu, and Irina Voineagu. Persistent homology analysis of brain transcriptome data in autism. *Journal of the Royal Society Interface*, 16, 2019.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015.

Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5(1):501–532, 2018.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010. doi: 10.1198/jasa.2009.tm08651.

Xiaoqi Xu, Nicolas Drougard, and Raphaëlle N. Roy. Topological data analysis as a new tool for EEG processing. *Frontiers in Neuroscience*, 15, 2021.

Xin Xu, Jessi Cisewski-Kehe, Sheridan B. Green, and Daisuke Nagai. Finding cosmic voids and filament loops using topological data analysis. *Astronomy and Computing*, 27:34–52, 2019.

Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete Computational Geometry*, 33:249–274, 2005.

Appendix A. More on Persistent Homology

This part is devoted to providing more detailed background information about how to construct persistent homology and the corresponding persistence diagram. We start with the definition of simplicial complexes and simplicial homology, and then we introduce how to construct persistent homology.

A.1 Simplicial homology

Let us start with the definition of simplicial complexes. Most of the contents of this subsection is based on Munkres (1984).

Definition 19 (Simplicial complexes) *An (abstract) simplicial complex is a collection K of finite non-empty sets, such that if σ is an element of K, so is every non-empty subset of σ.*

Each element σ of a simplicial complex K is called a simplex of K. The dimension of the simplex σ is defined to be $|\sigma| - 1$, i.e., the number of elements in σ minus one. When σ is a q-dimensional simplex, we simply say that σ is a q-simplex. The dimension $\dim K$ of the simplicial complex K is defined to be the maximum dimension of the simplices in K, i.e.,

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

If the set $\{\dim \sigma \,:\, \sigma \in K\}$ is not bounded, set $\dim K = \infty$. Each non-empty subset of σ is called a face of σ.

Let K be a simplicial complex. For each simplex $\sigma = \{v_0, \ldots, v_q\}$ in K, one can consider ordered tuples of vertices in σ. Namely, for every permutation α on $\{0, \ldots, q\}$, there exists an ordered tuple $(v_{\alpha(0)}, \ldots, v_{\alpha(1)})$. Such an ordered tuple is called a ordered simplex of σ. The collection of all ordered simplices of every simplex in K is called the ordered simplicial complex of K, and denoted by K_{ord}.

Let K_{ord} be an ordered simplicial complex of a simplicial complex K. Let $v^\alpha = (v_{\alpha(0)}, \ldots, v_{\alpha(q)})$ and $v^\beta = (v_{\beta(0)}, \ldots, v_{\beta(q)})$ be two ordered q-simplices of a common q-simplex $\sigma = \{v_0, \ldots, v_q\}$. Declare $v^\alpha \sim v^\beta$ if and only if α and β have the same sign, i.e., α and β differ only by even numbers of transpositions. Notice that this relation defines an equivalence relation on the set of ordered simplices of every simplex σ. Let $[v_0, \ldots, v_q]$ denote the equivalence class of the ordered simplex $(v_0, \ldots, v_q)$. Such an equivalence class is called an oriented q-simplex. Namely, every q-simplex with $q \geq 1$ induces two oriented q-simplices. Let K_{ori} denote the set of all oriented simplices of every simplex in K. When there is no confusion, we use the symbol σ to denote both a simplex and an oriented simplex.

For every natural number $q \geq 0$, set K^q_{ori} be the set of all oriented q-simplices of K. Define $C_q(K)$ be the set of all functions $c : K^q_{\text{ori}} \to \mathbb{Z}$ satisfying the following.

- $c(\sigma) = -c(\sigma')$ if σ and σ' are opposite orientations of the same simplex.
- $c(\sigma) = 0$ for all but finitely many oridented q-simplices σ, i.e., each c is finitely supported.

One can equip a group structure on $C_q(K)$ by defining the group operation to be elementwise addition. Notice that $C_q(K)$ is an abelian group with that group structure. Moreover, it is straightforward that $C_q(K)$ is a free abelian group whose basis can be constructed by choosing exactly one oriented simplex for every simplex σ. One can represent every element c in $C_q(K)$ by the finite $\mathbb{Z}$-linear combinations of oriented q-simplices of K, i.e., each c can be written as

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

where k is finite, $n_i \in \mathbb{Z}$ and $\sigma_i \in K^q_{\text{ori}}$ for all $1 \leq i \leq k$. Each function c is called a q-chain of K and $C_q(K)$ is called the group of oriented q-chains of K. We set $C_q(K) = 0$ if $q < 0$ or $q > \dim K$.

Now, we define the boundary operator of oriented chain complexes.

Definition 20 (Boundary operator) *Let K be a simplicial complex. For every integer q, define*

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

by assigning

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

where $[v_0, \ldots, \hat{v}_i, \ldots, v_q]$ is the $(q-1)$-oriented simplex obtained by deleting v_i from $[v_0, \ldots, v_q]$. Since $C_q(K)$ is a free abelian group, the map ∂_q can be extended into a unique group homomorphism $\partial_q : C_q(K) \to C_{q-1}(K)$. This homomorphism is called the boundary operator.

The key property of the boundary operator is the following:

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

In other words, the sequence $(C_q(K), \partial_q)_{q\in\mathbb{Z}}$ of abelian groups and group homomorphisms form a chain complex. This property can be rephrased as follows.

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

where Ker and Im mean the kernel and the image of a homomorphism, respectively. Since the sequence of groups of oriented chain complexes form a chain complex, it is possible to define the homology groups of it. Moreover, the kernel $\operatorname{Ker}\partial_q$ is usually written as $Z_q(K)$ and each of its elements is called a q-cycle; and, the image $\operatorname{Im}\partial_{q-1}$ is usually written as $B_q(K)$ and each of its elements is called a q-boundary.

Definition 21 (Simplicial homology) *Let K be a simplicial complex. For every integer q, the qth simplicial homology group is defined to be the following quotient group:*

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

and denoted by $H_q(K)$

Remark 22 *Instead of constructing $C_q(K)$ to be an abelian group, one can consider the free R-module on the same basis where R is a commutative ring. The boundary operator can be defined by the same, and now it can be uniquely extended to be an R-module homomorphism $\partial_q : C_q(K) \to C_{q-1}(K)$. The resulting sequence $(C_q(K), \partial_q)$ of R-modules and R-module homomorphisms form a chain complex of R-modules, so the simplicial homology of it can be defined by the same way; in this case, each homology group $H_q(K)$ becomes an R-module as well. In such a case, we denote the qth simplicial homology module of K by $H_q(K; R)$ and call it the qth simplicial homology of K with coefficients in R.*

A.2 Persistent homology

Let $\{K_r\}_{r\in\mathbb{R}}$ be a collection of simplicial complexes satisfying $K_{r_1} \subseteq K_{r_2}$ if $r_1 \leq r_2$. Such a collection is called a filtration of simplicial complexes (parametrized by $\mathbb{R}$). For each simplicial complex K_r in the filtration, it is possible to construct the chain complex $(C_q(K_r), \partial_q)_{q\in\mathbb{Z}}$ and the corresponding homology groups $(H_q(K_r))_{q\in\mathbb{Z}}$. In addition, each inclusion map $\iota_{r_1,r_2} : K_{r_1} \to K_{r_2}$ $(r_1 \leq r_2)$, induces a group homomorphism $C_q(K_{r_1}) \to C_q(K_{r_2})$, which is actually the inclusion map $C_q(K_{r_1}) \hookrightarrow C_q(K_{r_2})$ for every integer q; and, all such homomorphisms (inclusions) commute with boundary operators, i.e., each inclusion induces a chain map between chain complexes of oriented chains. Thus, each inclusion ι_{r_1,r_2} induces a homomorphism $\iota^q_{r_1,r_2} : H_q(K_{r_1}) \to H_q(K_{r_2})$ between homology groups for every q. This produces a collection $\{H_q(K_r)\}_{r\in\mathbb{R}}$ of simplicial homology groups accompanied with a group homomorphism $\iota^q_{r_1,r_2} : H_q(K_{r_1}) \to H_q(K_{r_2})$ for every q and every pair $r_1 \leq r_2$.

For each pair $r_1 \leq r_2$ and each q, the image of $\iota^q_{r_1,r_2} : H_q(K_{r_1}) \to H_q(K_{r_2})$, denoted by $\operatorname{Im}\iota^q_{r_1,r_2}$, is called the qth persistent homology group that persists from r_1 to r_2. The rank of the group $\operatorname{Im}\iota^q_{r_1,r_2}$ is called the qth persistent Betti number that persists from r_1 to r_2 and denoted by $\beta^q_{r_1,r_2}$. Intuitively, the Betti number $\beta^q_{r_1,r_2}$ represents the number of independent q-cycles that were born before the parameter r_1 and have not been dead until the parameter r_2 in the filtration. Furthermore, for each q-cycle in the filtration, it is possible to consider

the parameter values at which the cycle shows up at first (birth) and disappears (death), respectively.

Let σ be a q-cycle that shows up in the filtration at some point, i.e., σ is an element of $\mathrm{Ker}\,\partial_q(K_r)$ for some r. Then, it is possible to consider the birth and death times (parameter values) of it. By bringing together all birth-death pairs of all q-cycles in the filtration, one can form a multiset of points of the form (b, d) with $b \leq d \leq \infty$. That multiset is called the qth persistence diagram of the filtration. The formal construction of the persistence diagram is involved with the structure theorem of finitely generated graded modules over a principal ideal domain, which is a theorem from abstract algebra. Please refer to Carlsson (2009) and Edelsbrunner and Harer (2008) for more formal and comprehensive discussion.

A.3 Some ways to construct a filtration of simplicial complexes

Now, we introduce several ways to obtain a filtration of simplicial complexes that play a role in the main discussion of this paper. The contents of this subsection can be found in Edelsbrunner and Harer (2009).

Let $D = \{x_1, \ldots, x_n\}$ be a finite subset of a metric space $(\mathcal{X}, d)$. For every non-negative real number $r \geq 0$, consider the ball $B(x_i; r) := \{y \in \mathcal{X} \ : \ d(y, x_i) < r)$ centered at each $i \in \{1, \ldots, n\}$. The Čech complex $\check{\mathcal{C}}(D; r)$ on D with radius r is the simplicial complex defined as follows. A subset $\sigma = \{x_{i_0}, \ldots, x_{i_q}\}$ of D is a member of $\check{\mathcal{C}}(D; r)$ if and only if $\cap_{j=0}^{q} B(x_{i_j}; r) \neq \emptyset$. Notice that $\check{\mathcal{C}}(D; r_1) \subseteq \check{\mathcal{C}}(D; r_2)$ for every pair $r_1 \leq r_2$. Hence, the collection $\{\check{\mathcal{C}}(D; r)\}_{r \geq 0}$ of Čech complexes forms a filtration of simplicial complexes.

There are several variants of the Čech complex. One of such variants is the Vietoris-Rips complex. The Vietoris-Rips complex $\mathrm{VR}(D; r)$ on D with radius r is defined as follows. A subset $\sigma = \{x_{i_0}, \ldots, x_{i_q}\}$ of D is a member of $\mathrm{VR}(D; r)$ if and only if $B(x_{i_{j_k}}; r) \cap B(x_{i_{j_l}}; r) \neq \emptyset$ for every $k, l \in \{0, \ldots, q\}$, i.e., The balls $B(x_{i_0}; r), \ldots B(x_{i_q}; r)$ pairwise intersect with one another. It is also obvious that $\mathrm{VR}(D; r_1) \subseteq \mathrm{VR}(D; r_2)$ whenever $r_1 \leq r_2$. Hence, the collection $\{\mathrm{VR}(D; r)\}_{r \geq 0}$ of Vietoris-Rips complexes forms a filtration of simplicial complexes. The following relationship between the Čech complex and the Vietoris-Rips complex indicates that, on a finite subset in an Euclidean space, the filtration of Čech complexes and that of Vietoris-Rips complexes have essentially the same information.

Proposition 23 *Let $D = \{x_1, \ldots, x_n\}$ be a finite subset of a Euclidean space equipped with the metric induced by the ℓ^2-norm on it. Then, for every $r \geq 0$, the following holds.*

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

The last way of construction is obtained from a real-valued function defined on a metric space. Let $(\mathcal{X}, d)$ be triangulable a metric space and $f : \mathcal{X} \to \mathbb{R}$ be a real-valued continuous function. For each $r \in \mathbb{R}$, consider the sub-level set $L_r := f^{-1}\big((-\infty, r]\big)$, which is a subset of $\mathcal{X}$. Notice that $L_{r_1} \subseteq L_{r_2}$ whenever $r_1 \leq r_2$. Moreover, since $\mathcal{X}$ is triangulable, all sub-level sets can be triangulized while respecting the inclusion relationships. Hence, the collection of such triangulizations of the collection $\{L_r\}_{r \in \mathbb{R}}$ of sub-level sets produces a filtration of simplicial complexes.

Before closing this section, we introduce a certain condition on a continuous function $f : \mathcal{X} \to \mathbb{R}$ that ensures that f does not behave too wildly.

Definition 24 (Tame functions) *Let $(\mathcal{X}, d)$ be a triangulable metric space and $f : \mathcal{X} \to \mathbb{R}$ a real-valued continuous function. Set X_r to be the triangulization of the sub-level set $f^{-1}\big((-\infty, r]\big)$. Let $\iota^q_{r_1,r_2} : H_q(X_{r_1}) \to H_q(X_{r_2})$ be the group homomorphism induced by the inclusion map $\iota_{r_1,r_2} : X_{r_1} \to X_{r_2}$ for every pair $r_1 \leq r_2$. We call $r \in \mathbb{R}$ a homological critical value if there is no positive number $\epsilon > 0$ for which $\iota^q_{r-\epsilon,r+\epsilon}$ is an isomorphism for each dimension q. The function f is said to be tame if f produces only finitely many homological critical values and all homology groups of all sub-level sets of it have finite rank.*

Appendix B. Proofs of the Main Results

In this part, we present the detailed proofs of the theorems in Section 3 and Section 4. Throughout this section, unless there is no further specification, the symbol $\|\cdot\|$ denotes the ℓ^2-norm in the Euclidean space where the data points discussed in each proof live.

B.1 Proofs of the Results in Section 3

Proof [Proof of Lemma 4] Fix $r > 0$ so that $d_{m-1} < r < d_m$, and let $G(D; r)$ be the geometric graph with vertex set $\mathcal{X}$ and connecting threshold r; i.e., D is the vertex set of $G(D; r)$, and each pair $\{x_i, x_j\}$ of vertices is an edge of it if $d(x_i, x_j) \leq r$. Since $(0, d_m)$ is an element of the diagram $\mathcal{P}_0^{\check{\text{C}}\text{ech}}(D)$, there are at least two connected components $\mathcal{Y}_1$ and $\mathcal{Y}_2$ in $G(D; r)$ which satisfy

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

Let $\mathcal{Y}_1$ and $\mathcal{Y}_2$ be such connected components, and let $x_1 \in \mathcal{Y}_1$ and $x_2 \in \mathcal{Y}_2$ be the points attaining the minimum, i.e., $d(x_1, x_2) = d_m$. Set D' to be the set obtained by adding one point, say z, at the mid-point of x_1 and x_2. It is obvious that the death time of $\mathcal{Y}_1$ (or equivalently, $\mathcal{Y}_2$) is cut in half. Notice that the death times of the other connected components in the filtration of D cannot be bigger by adding the point z. Thus, we can write

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

with $d_j' \leq d_{m-1}$ for all $j = 1, \ldots, t$. Here, the element $(0, d_m/2)$ has multiplicity at least 2.

To calculate the bottleneck distance between $\mathcal{P}_0^{\check{\text{C}}\text{ech}}(D)$ and $\mathcal{P}_0^{\check{\text{C}}\text{ech}}(D')$ we have to consider all possible bijections between $\mathcal{P}_0^{\check{\text{C}}\text{ech}}(D)$ and $\mathcal{P}_0^{\check{\text{C}}\text{ech}}(D')$. All such bijections can be classified into three categories. First, $(0, d_m) \in \mathcal{P}_0^{\check{\text{C}}\text{ech}}(D)$ is associated with element $(0, d_j') \in \mathcal{P}_0^{\check{\text{C}}\text{ech}}(D')$ for some $j \in \{1, \ldots, t\}$. Second, $(0, d_m) \in \mathcal{P}_0^{\check{\text{C}}\text{ech}}(D)$ is associated with $(0, d_m/2) \in \mathcal{P}_0^{\check{\text{C}}\text{ech}}(D')$. Third, $(0, d_m) \in \mathcal{P}_0^{\check{\text{C}}\text{ech}}(D)$ is associated with a point in the diagonal line. In the first case, the possible minimum distance concerning $(0, d_m)$ cannot be smaller than δ. In the second case, the distance between $(0, d_m)$ and $(0, d_m/2)$ is obviously $d_m/2$. In the last case, the distance between $(0, d_m)$ and the diagonal line is $d_m/\sqrt{2}$. Since the bottleneck distance is defined by taking the minimum over all such bijections, the desired result follows. ■

Proof [Proof of Theorem 5] Suppose that n is even. Let a and b be two points in the set E with $|a-b| = \mathrm{diam}E$, and D consist of $n/2$ copies of a and $n/2$ copies of b. Let D' be obtained by moving one of as to the mid-point of a and b, say c. Then, it is obvious that

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

and

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

This proves that the bottleneck distance between these two diagrams is lower bounded by $\mathrm{diam}E/4$, which implies the desired result. When n is odd, one can take D to have $(n-1)/2$ copies of a and $(n+1)/2$ copies of b, and the result does not change.

As for the second result, the proposed upper bound can be established by applying the reverse triangle inequality. To establish the lower bound, notice that for any pair of sets D and D',

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

The supremum of the last expression over all adjacent pairs D and D' is lower bounded by $\mathrm{diam}E/4$ as a consequence of the first result. This completes the proof. ■

Proof [Proof of Theorem 8] Let $\delta_D^{(p)}$ and $\delta_{D'}^{(p)}$ be L^p-DTM to the empirical distributions of D and D', respectively. By the stability theorem (2.2) and the Wasserstein stability (3.1) of the DTM, we have

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

where $\hat{\mu}_D$ and $\hat{\mu}_{D'}$ represent the empirical distributions on D and D', respectively. We are going to establish a upper bound of the right-hand side of the inequality (A.1.2).

Assume that $H(D, D') = 1$. Let x be the element that is in D but not in D', and z be the element that is in D' but not in D. Let π be the coupling of $\hat{\mu}_D$ and $\hat{\mu}_{D'}$ defined as follows: For every $y \in D$, set

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

and

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

It is straightforward to verify that π is indeed a coupling of $\hat{\mu}_D$ and $\hat{\mu}_{D'}$. With this π we have

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

By the definition of the Wasserstein distance W_p, we have

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

where ν ranges over all couplings of P_n and P_n'. Therefore, we obtain the following:

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

which implies the desired result. ■

Proof [Proof of Proposition 9] Let D be a data set whose points are split into 50% and 50% at two ends a, b of E respectively. More specifically, $\| a - b \| = \text{diam}\, E$ and the set D contains $2/n$ copies of a and $2/n$ copies of b. Let c be the mid-point of a and b; that is $\| a - c \| = \| b - c \| = \text{diam}\, E/2$. Construct D' by moving one a in D to c; namely, D' has $n/2 - 1$ copies of a, $n/2$ copies of b, and one c. Let δ_D be the L^1-DTM to the empirical distribution on D with resolution m and $\delta_{D'}$ likewise. Then, we have

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

On the other hand,

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

Recall that $k = \lceil mn \rceil$. Notice that any point x on the line segment $\overline{ab}$ satisfies $\delta_D(x) \leq \delta_D(c)$ and $\delta_{D'}(x) \leq \delta_{D'}(c)$. Hence, the 0th persistence diagram $\mathcal{P}_0^{\text{DTM}_1}(D)$ of D is obtained as follows:

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

Similarly, $\mathcal{P}_0^{\text{DTM}_1}(D')$ is obtained as follows:

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

The bottleneck distance between the two diagrams above is calculated as follows:

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

■

Proof [Proof of Corollary 10] The upper bound is obtained by applying the reverse triangle inequality.

As for the lower bound, notice that

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

The last expression is bounded below by $\text{diam}E/(2\lceil mn \rceil)$ as a result of Proposition 9. ■

B.2 Proofs of the Results in Section 4

Proof [Proof of Proposition 15] Notice that the inequality (2.5) gives

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

for every $t \geq 0$. The reverse triangle inequality yields

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

Combining those two yields

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

Recall that Δ is chosen to be $\Delta = (\ell+1)\mathrm{diam}E/(mn)$ and $|\widetilde{\mathsf{Pers}}_{M,N}| = N^M$. Thus, we have

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

Now, recall that the upper-left triangle $\bar{\mathcal{T}}$ is discretized by $N^2 = N^2(n) = n^2$ equally-spaced points; the length of each spacing is bounded by $C\mathrm{diam}E/n$ for some constant C that only depends on the chosen Euclidean distance. Hence, with all large enough n, the error in approximating $\mathcal{P}_{\mathrm{OPT}}$ by $\widetilde{\mathcal{P}}_{\mathrm{OPT}}$ satisfies

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

which completes the proof. ■

Theorem 17 can be proved by establishing the following result.

Proposition 25 *Let $M = M(n)$ be a non-decreasing sequence of positive integers satisfying $M(n) \geq |\mathcal{P}_q(\mu)|$ for all large enough n. Then, for every $q \geq 0$, we have*

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

Moreover, we also have

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

Proof [Proof of Proposition 25] With large enough n, we can assume that $M \geq |\mathcal{P}_q(\mu)|$. In other words, $\mathcal{P}_q(\mu)$ belongs to the space of persistence diagrams having at most M elements. Hence, by the definition of $\mathcal{P}_{q,\mathrm{OPT}}$,

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

As for $\mathrm{d_B}(\mathcal{P}_q(\mu), \mathcal{P}_{q,\mathrm{OPT}})$, the triangle inequality gives

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

According to Theorem 2 in Fournier and Guillin (2015) along with the stability theorem (2.2) of the bottleneck distance and the Wasserstein stability (3.1) of the DTM, it is straightforward to deduce that

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

for every $q \geq 0$. ■

Proof [Proof of Theorem 17] According to Proposition 25, we have

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

for every $q \geq 0$. This estimate, together with the estimate given in Proposition 15, establishes the desired result. ■

The proof of Theorem 18 is achieved by establishing the following three lemmas. The first lemma is rather technical.

Lemma 26 *Set $\mathcal{P} = \mathcal{P}_0^{\mathrm{DTM}_1}$. Assume that the resolution m of the DTM is chosen to satisfy $m \leq 1/2$. For any pair of positive integers K and n with $1 \leq K \leq n$, there exists a pair of data sets X_n and Y_n satisfying $|X_n| = |Y_n| = n$, $H(X_n, Y_n) = K$, and $\mathrm{d_B}(\mathcal{P}(X_n), \mathcal{P}(Y_n)) \geq \frac{CK}{n}$ for some constant C independent of K and n, where $H(X_n, Y_n)$ denotes the Hamming distance between X_n and Y_n.*

Proof

Recall that $k = \lceil mn \rceil$. The whole situation will be broken down into three cases: (i) $1 \leq K \leq \min\{k, n/2 - k\}$, (ii) $\min\{k, n/2 - k\} < K < \max\{k, n/2 - k\}$, and (iii) $K \geq \max\{k, n/2 - k\}$.

First, let us assume that $1 \leq K \leq \min\{k, n/2 - k\}$. Choose two points a and b satisfying $\| a - b \| = \operatorname{diam} E$; for instance, in the case of $E = [0,1]^d$, one may choose $a = (0, \ldots, 0)$ and $b = (1, \ldots, 1)$. Choose the data set X_n that consists of $n/2$ copies of a and $n/2$ copies of b (If n is odd, take $(n-1)/2$ copies of a and $(n+1)/2$ copies of b; the results will be the same). On the other hand, choose the data set Y_n constructed by moving K copies of a to the mid-point of a and b, say c, i.e., as multisets, X_n and Y_n can be expressed as follows:

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

Since $M \leqslant n/2 - k$, the point a still has more than k numbers of points in the data set Y_n. Thus, we have

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

Let $x(t)$ be the point in the line segment $\overline{ac}$ that divides $\overline{ac}$ into the ratio $t : (1-t)$ with $t \in [0,1]$. Then, we have

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

Now, let us further decompose the situation into two cases: (i-1) $(k-2K) \geq 0 \Leftrightarrow K \leq k/2$ and (i-2) $(k-2K) < 0 \Leftrightarrow K > k/2$. In the case (i-1), $\delta_{Y_n}(x(t))$ is increasing in t. Hence, $\mathcal{P}(Y_n)$ is obtained to be

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

Notice that $\mathcal{P}(X_n)$ is obtained to be

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

Therefore,

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

In the case (i-2), $\delta_{Y_n}((x(t))$ decreases from $t = 1/2$ to $t = 1$. Thus, $\mathcal{P}(Y_n)$ is given to be

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

the bottleneck distance between $\mathcal{P}(X_n)$ and $\mathcal{P}(Y_n)$ can be derived by comparing the two different scenarios. First case corresponds $(0, \operatorname{diam}/4)$ in $\mathcal{P}(Y_n)$ to $(0, \operatorname{diam} E/2)$ in $\mathcal{P}(X_n)$. The distance obtained from this case must be greater than or equal to $\operatorname{diam} E/4$. The other case corresponds $(0, \operatorname{diam} E/2)$ in $\mathcal{P}(X_n)$ to $(\delta_{Y_n}(c), \operatorname{diam} E/4)$ in $\mathcal{P}(Y_n)$. Consequently, $(0, \operatorname{diam} E/4)$ in $\mathcal{P}(Y_n)$ must correspond to a point in the diagonal. Thus, the distance obtained in this case must be greater than or equal to $\operatorname{diam} E/(4\sqrt{2})$. Therefore,

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

Now, let us turn our attention to the case (ii), which assumes that $\min\{k, n/2-k\} < K < \max\{k, n/2-k\}$. First, consider the case $k < n/2-k$, so that $k < K < n/2-k$. In this case, both a and c have at least k points, so

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

The above result gives us

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

Thus,

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

Second, consider the case $k > n/2-k$, so that $n/2-k < K < k$. In this case, both a and c have less than k points. Thus,

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

Using the similar argument we utilized in the case (i), it is possible to demonstrate that the desired result is true in this case too.

Finally, let us consider the case (iii) where $K \geq \max\{k, n/2 - k\}$. In this case, $\mathcal{P}(Y_n)$ has at least one element $(0, \operatorname{diam}/4)$. Hence its bottleneck distance from $\mathcal{P}(X_n)$ is always greater than or equal to $\operatorname{diam} E/4$. This completes the proof of the lemma.

■

The next two lemmas address the concept of DP in terms of a hypothesis testing framework. Lemma 27 is a well-known folklore result in the DP literature. It can be easily derived using the f-DP framework (Dong et al., 2022). We give a direct proof that does not require using f-DP.

Lemma 27 *Let X and X' be adjacent data sets, and $\mathcal{M}$ be any ϵ-DP mechanism. For a hypothesis test $H_0 : \mathcal{M}(X)$ versus $H_1 : \mathcal{M}(X')$,*

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

Proof Call $\mathcal{Y}$ the probability space that $\mathcal{M}(X)$ lives in. Call μ_X the probability measures on $\mathcal{Y}$ for $\mathcal{M}(X)$. By Awan et al. (2019, Proposition 2.3), there exists a base measure ν, which dominates μ_X for all databases X. Call f_X the density of μ_X with respect to ν, which by Awan et al. (2019, Proposition 2.3) satisfies $f_X \leq e^{\epsilon} f_{X'}$ almost everywhere ν, for adjacent databases X and X'.

Let X and X' be adjacent databases, and let $\phi : \mathcal{Y} \to [0, 1]$ be a test. Then the type I and type II errors are $\mathrm{I} = \mathbb{E}\phi(\mathcal{M}(X))$ and $\mathrm{II} = 1 - \mathbb{E}\phi(\mathcal{M}(X'))$, respectively. Then

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

Repeating the argument using $\phi' = 1 - \phi$ and swapping the roles of X and X', we get

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

Then,

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

which implies that $\mathrm{I} + \mathrm{II} \geq \frac{2}{1+e^{\epsilon}}$.

■

Lemma 28 *Let $(\epsilon_n)_{n=1}^{\infty}$ be a sequence of positive numbers satisfying $1/n \leq \epsilon_n \leq 1$ for every n. Set $K_n = \lfloor 1/\epsilon_n \rfloor$. For each n, For a given sequence of positive numbers $(\Delta_n)_{n=1}^{\infty}$, let $\{(X_n, Y_n)\}_{n=1}^{\infty}$ be a sequence of finite data sets satisfying, for each n, $H(X_n, Y_n) = K_n$ and $\mathrm{d_B}(\mathcal{P}(X_n), \mathcal{P}(Y_n)) \geq K_n \Delta_n$. Here, H denotes the Hamming distance between sets and $\mathcal{P}$ means an arbitrary persistence diagram. Then for any $\epsilon(n)$-DP mechanism $\mathcal{M}$ that produces a privatized persistence diagram, it is not possible for both $\mathrm{d_B}(P_{X_n}, \mathcal{M}(X_n)) = o_p(\Delta_n/\epsilon_n)$ and $\mathrm{d_B}(P_{Y_n}, \mathcal{M}(Y_n)) = o_p(\Delta_n/\epsilon_n)$.*

Proof For simplicity of notation, we will suppress the dependence of X, Y, ϵ, Δ, and K on n. We will construct a hypothesis test for $H_0 : \mathcal{M}(X)$ versus $H_1 : \mathcal{M}(Y)$. Note that since $\mathcal{M}$ is ϵ-DP for groups of size 1, it is $K\epsilon$-DP for groups of size K (Dwork and Roth, 2014, Theorem 2.2).

Define the sets S_X and S_Y as follows:

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

and define our test to be $\phi(\mathcal{M}(\cdot)) = I(\mathcal{M}(\cdot) \in S_Y)$, which is the indicator function on the event $\mathcal{M}(\cdot) \in S_Y$. Then

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

As a result of Lemma 27, we have that

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

since $k\epsilon \leq 1$, which implies that either

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

or

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

This rules out the possibility that both are $o_p(K\Delta) \leq o_p(\Delta/\epsilon)$. ■

Proof [Proof of Theorem 18] According to Lemma 26, it is guaranteed that the 0th persistence diagram of the L^1-DTM filtration $\mathcal{P}_0^{\mathrm{DTM}_1}$ satisfies the conditions stated in Lemma 28 with $\Delta_n = \frac{C}{mn}$ for some constant C independent of n. For brevity, set $\mathcal{P} = \mathcal{P}_0^{\mathrm{DTM}_1}$. Then, for a sequence (X_n, Y_n) of data sets satisfying the condition, Lemma 28 tells us that either

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

or

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

holds, which rules out the possibility that they are $o_p(K/n) \leq o_p(1/(n\epsilon))$. This completes the proof. ■

Appendix C. Supplements of the Simulation and the Real Data Analysis

C.1 More Detailed Description of Our Algorithm

Here, the algorithm of our privacy mechanism, which is introduced in Section 4, is explained in detail. For a given data set D, Let S denote the maximum value of the L^1-DTM function on the data set and M a specified positive integer.

To get an initial diagram $\mathcal{P}_{\mathrm{DP}}^{(0)} = \big(\mathcal{P}_{0,\mathrm{DP}}^{(0)}, \mathcal{P}_{1,\mathrm{DP}}^{(0)}\big)$, generate independently and identically distributed sample $x_1, \ldots, x_M, z_1, \ldots, z_M$ from the uniform distribution on the closed interval $[0, S]$, i.e., $x_1, \ldots, x_n, z_1, \ldots, z_M \overset{i.i.d.}{\sim} \mathrm{Unif}[0, S]$, symbolically. For each $i \in \{1, \ldots, M\}$, set

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

the diagram $\mathcal{P}_{0,\mathrm{DP}}^{(0)}$ is constructed as follows:

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

The other initial diagram $\mathcal{P}_{1,\mathrm{DP}}^{(0)}$ is generated in the same way independent of $\mathcal{P}_{0,\mathrm{DP}}^{(0)}$. Notice that each initial diagram consists of M points uniformly distributed on the upper-left triangle $\{(x, y) \,:\, x, y \in [0, S] \text{ and } y \geqslant x\}$.

For generated $\mathcal{P}_{\mathrm{DP}}^{(t)} = (\mathcal{P}_{0,\mathrm{DP}}^{(t)}, \mathcal{P}_{1,\mathrm{DP}}^{(t)})$, the next candidate $\mathcal{P}_{\mathrm{DP}}^{(t+1)}$ by adding Gaussian noise to each element in each diagram in tth step. To be precise, write

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

Generate i.i.d. sample $Z_1^{(t)}, \ldots, Z_M^{(t)}$ from the 2-dimensional Gaussian distribution with mean $(0, 0)$ and covariance matrix $\sigma^2 I_2$, where σ is a pre-specified positive number and I_2 is the 2 by 2 identity matrix. Set

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

$\mathcal{P}_1^{'}$ is constructed in the same with independent of $\mathcal{P}_0^{'}$, and set $\mathcal{P}^{'} = (\mathcal{P}_0^{'}, \mathcal{P}_1^{'})$. Then, calculate the accept/reject probability in the Metropolis-Hastings sampler p defined in terms of the bottleneck distance:

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

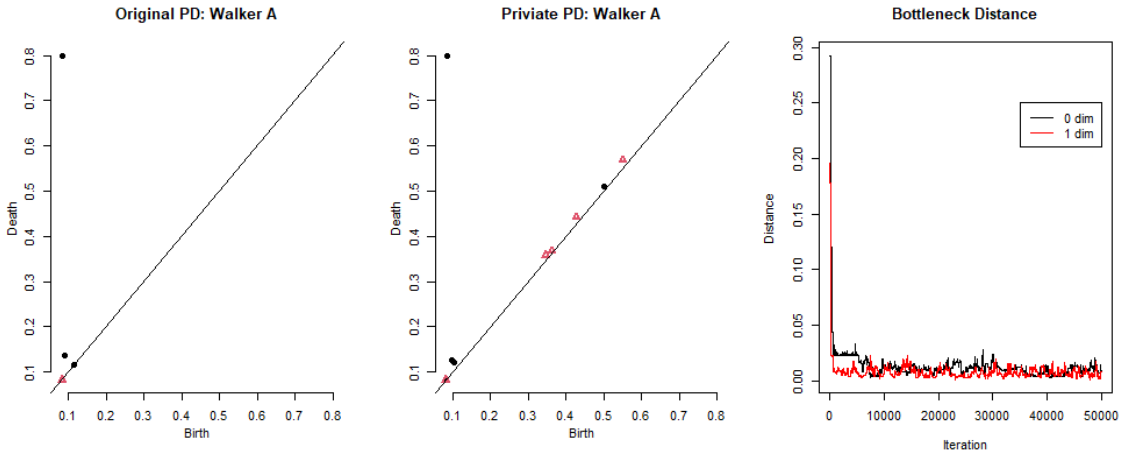
where

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

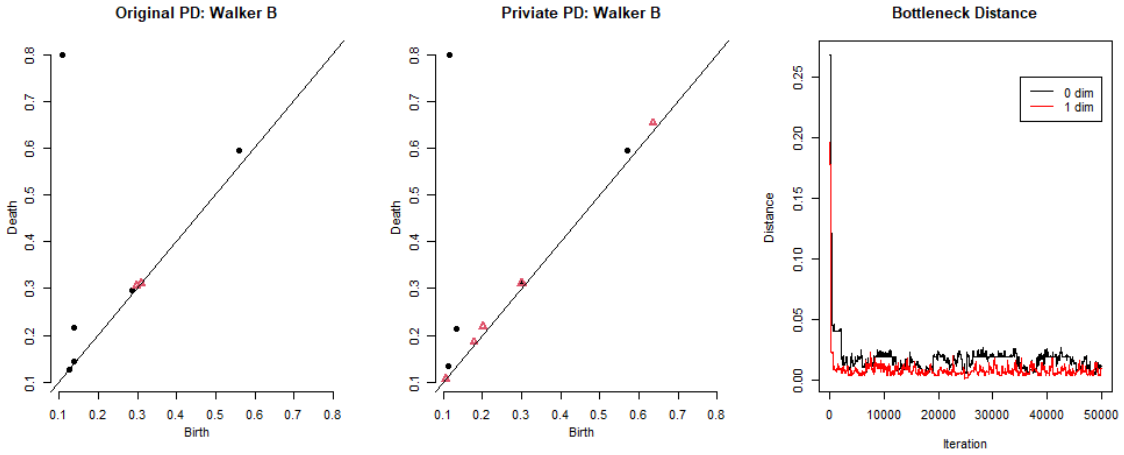
Generate $U \sim \mathrm{Unif}(0, 1)$. If $\log U \leq p$, take $\mathcal{P}_{\mathrm{DP}}^{(t+1)} = \mathcal{P}^{'}$; otherwise, take $\mathcal{P}_{\mathrm{DP}}^{(t+1)} = \mathcal{P}_{\mathrm{DP}}^{(t)}$. This procedure is carried out repeatedly again, and the $\mathcal{P}_{\mathrm{DP}}^{(t)}$ at the final iteration is proposed as a privatized persistence diagram. The whole procedure is summarized in Algorithm 1.

C.2 Additional Results in the Real Data Analysis

The following illustration, Figure 8, depicts the accuracy of privatized persistence diagrams for Walker A and B. The procedure of implementing the mechanism is the same with that for Walker C described in Section 6. One can see that we also obtain quite accurate privatized diagrams in this case as well.



(a) Walker A



(b) Walker B

Figure 8: Figure 8-(a) presents the L^1-DTM persistence diagram of the data of Walker A and its privatized diagram. Also, the change of the bottleneck distance between the true and privatized diagram over the MCMC iterations is depicted. At the final iteration, we have $\mathrm{d_B}(P_0(D), P_{0,\mathrm{DP}}) = 0.01$ and $\mathrm{d_B}(P_1(D), P_{1,\mathrm{DP}}) = 0.009$. Figure 8-(b) presents the same kind of information about Walker B. Here, at the final iteration, we obtain $\mathrm{d_B}(P_0(D), P_{0,\mathrm{DP}}) = 0.011$ and $\mathrm{d_B}(P_1(D), P_{1,\mathrm{DP}}) = 0.009$.

Algorithm 1 MCMC implementation of the exponential mechanism

Input: $\mathcal{P}_0(D), \mathcal{P}_1(D)$, and a positive integer M
Initialization: $\mathcal{P}_{0,\text{DP}}^{(0)}, \mathcal{P}_{1,\text{DP}}^{(0)} \sim \text{Unif}(\mathsf{Pers}_M)$
for $i = 1, 2, \ldots,$ **do**
  $\mathcal{P}_0' = \mathcal{P}_{0,\text{DP}}^{(t-1)} + \text{N}(0, \sigma^2 I_2)$, $\mathcal{P}_1' = \mathcal{P}_{1,\text{DP}}^{(t-1)} + \text{N}(0, \sigma^2 I_2)$
  $\mathcal{P}' = (\mathcal{P}_0', \mathcal{P}_1')$
  $p = \min\left\{0, -\frac{\epsilon}{2\Delta}\left(u_D(\mathcal{P}^{(t-1)}) - u_D(\mathcal{P}')\right)\right\}$
  $U \sim \text{Unif}(0, 1)$
  if $\log U \leq p$ **then**
    $\mathcal{P}_{0,\text{DP}}^{(t)} = \mathcal{P}_0', \mathcal{P}_{1,\text{DP}}^{(t)} = \mathcal{P}_1'$
  else
    $\mathcal{P}_{0,\text{DP}}^{(t)} = \mathcal{P}_{0,\text{DP}}^{(t-1)}, \mathcal{P}_{1,\text{DP}}^{(t)} = \mathcal{P}_{1,\text{DP}}^{(t-1)}$
  end if
end for