Preserving Privacy in High-Dimensional Data Publishing

Narges Alipourjeddi and Ali Miri
Department of Computer Science, Toronto Metropolitan University, Toronto, Canada

Keywords: High-Dimensional Data, Privacy Preservation, Persistent Homology, Differential Privacy, Data Publishing, Topological Data Analysis.

Abstract: As the era of big data unfolds, high-dimensional datasets with complex structures have become increasingly prevalent in various fields, including healthcare, finance, and social sciences. Extracting valuable insights from such data is essential for scientific discovery and decision-making. However, the publication of these datasets is full of privacy concerns, as they often contain sensitive and personally identifiable information. In this paper, we introduce a novel approach that addresses the delicate balance between data privacy and the exploration of high-dimensional data's underlying structure. We leverage the power of persistent homology, a topological data analysis method, to unveil hidden patterns and captures the persistent topological features of the data, allowing us to study its shape and structure across different scales. Adding noise into the low dimensional embedding and provide private persistence diagram with differential privacy, offers a rigorous and well-established framework to ensure that individuals' privacy in the dataset is protected. We synthetically generate high-dimensional data with a focus on differential privacy-preserved persistence diagrams, ensuring privacy in our publication of the synthesized dataset. We conduct extensive experiments on three real-world datasets and the experimental results demonstrate that our mechanism can significantly improve the data structure of the published data while satisfying differential privacy.

1 INTRODUCTION

In our data-driven era, high-dimensional datasets have become ubiquitous, permeating fields as diverse as healthcare, finance, and social sciences. The information encapsulated within these data sets holds the key to crucial scientific discoveries, informed decision-making, and innovation. However, sharing this data is not without its challenges, and among the most significant is the need to navigate the delicate balance between data publishing and data privacy.

The advent of big data has brought forth a pressing concern: how can we unlock the valuable insights hidden within high-dimensional datasets, while safeguarding the sensitive and personally identifiable information they contain? This question is central to our research as we delve into the intersection of data privacy and data publishing. Our approach, built upon the powerful foundations of persistent homology and differential privacy, seeks to address this fundamental question.

Privacy-Preserving Data Publishing (PPDP) has gained significant attentions in recent years as a promising approach for information sharing while preserving data privacy. There exists standard methods such as k-anonymity (Mahanan et al., 2021), l-diversity (Binjubeir et al., 2019) and t-closeness (Binjubeir et al., 2019) that data collectors (sometimes also referred to as curators) can apply to protect and anonymize datasets. However, these methods can still leak information when analysis involves additional datasets or auxiliary information from other sources. One also needs to able to formally measure information leakage and privacy protection. A commonly used methodology to provide a framework for preserving and measuring privacy is Differential Privacy (DP) (Dwork et al., 2014). DP can be used to provide privacy guarantees using an information theoretical approach. The main idea in this approach is that what can be learned from the published data is (approximately) the same, whether or not any particular individual was included in the input database. This model is mathematical foundation with a formal definition and rigorous proof while making the assumption that an attacker has the maximum background knowledge.

Nonetheless, ensuring differential privacy in the publication of high-dimensional data continues to be a powerful challenge, primarily due to the "Curse

Alipourjeddi, N. and Miri, A.
Preserving Privacy in High-Dimensional Data Publishing.
DOI: 10.5220/0012455600003648

In *Proceedings of the 10th International Conference on Information Systems Security and Privacy (ICISSP 2024)*, pages 845-852
ISBN: 978-989-758-683-5; ISSN: 2184-4356

of High-Dimensionality". This phenomenon signifies that as the dimensionality of the data grows, the complexity and computational cost of handling and analysing multidimensional data experience exponential growth.

One promising way to address high dimensionality is to disassemble the dataset into a group of lower dimensional datasets. One of the traditional approach for disassembling the dataset into a group of lower dimension dataset presented by Zhang et al. (Zhang et al., 2017). They used a Bayesian network to deal with high dimensionality. They assumed some correlations between attributes exist and, if these correlations can be modelled, the model can be used to generate a set of marginal datasets to simulate the distribution of the original dataset. The disadvantage of this solution is that it consumes too much of the privacy budget during network construction and, hence, makes the approximation of the distribution inaccurate.

In this work, we employ a non-linear dimensionality reduction method grounded in the manifold hypothesis, which posits that real-world data sets may reside on a non-linear, low-dimensional manifold embedded within a high-dimensional ambient vector space. In many real-world datasets, the characteristics of this underlying manifold are initially unknown. The process of manifold learning is employed to endeavour the extraction of this hidden manifold by mapping the data into a lower-dimensional space. One of the current tools in this era is Topological Data Analysis (TDA), utilized for the analysis of both geometric and topological information within datasets.

TDA represents an innovative field of data analysis that was developed to capture the underlying topological structures within data. Over the past few decades, TDA has undergone extensive research and exploration. This approach has proven invaluable in handling complex, high-dimensional datasets that challenge the capabilities of traditional data analysis methods.

Persistent homology is a powerful tool for dimensionality reduction from the field of TDA. In high-dimensional data analysis, the manifold hypothesis suggests that many datasets naturally lie on or near lower-dimensional manifolds. These manifolds represent the underlying structure of the data, even though the data is observed in a higher-dimensional space. Persistent homology detect topological features that represent the various components of the data, including the lower-dimensional manifolds. These features can include connected components (0-dimensional manifolds), loops (1-dimensional manifolds), voids (2-dimensional manifolds) and so on.

In this paper, we present a novel approach that obtain topological features for our datasets and captures how long these topological features persist privately. This makes it possible to generate and publish high dimensional data privately. Specifically, we make the following contributions:

1) We use persistent homology technique to analyse theoretical meaning behind our datasets and creating persistence diagram.
2) We implement differential privacy measures on the persistence diagram to make private features.
3) We generate synthetic dataset based on the private persistent diagram.

We commence with a preliminaries section, laying the groundwork with essential background information and the introduction of the notations we will use (Section 2). In Section 3, we delve into an examination of the related work in the field. Our framework is presented comprehensively in Section 4, while Section 5 showcases its practical capabilities. The paper concludes with a summary and insights in Section 6.

2 PRELIMINARIES

In this section we review some of the standard concepts from topology, algebraic topology and differential privacy. We want to use these methods to synthesize private high-dimensional datasets.

2.1 Differential Privacy Fundamentals

The protection of individuals' privacy in the context of data publishing and analysis has become a paramount concern with the increasing availability of large and sensitive datasets. Differential privacy offers a rigorous and effective approach to address this concern by ensuring that individual privacy is maintained while allowing for meaningful data publishing. This section introduces the core concepts and terminology related to differential privacy. Formally, differential privacy is defined as follows:

Definition 2.1 (ε-differential Privacy)**.** *A randomized mechanism M gives* ε*-differential privacy for every set of outputs* Ω*, and for any neighbouring datasets of D and* D'*, if M satisfies*

$$Pr[M(D) \in \Omega] \leq exp(\varepsilon) \cdot Pr[M(D') \in \Omega]$$

In other words, the probability of obtaining a specific outcome from the mechanism M is only slightly influenced by the inclusion or exclusion of any individual's data.

Two fundamental components of differential privacy are the sensitivity of a function and the privacy parameter ε. The sensitivity of a function f quantifies how much the function's output can change when a single data point is added or removed from the dataset. The parameter ε refers to the privacy budget, which controls the level of privacy guarantee achieved by mechanism M. A smaller ε represents a stronger privacy level. For a strong privacy guarantee, we need the privacy budget to be small with an ideal in the range of zero and one.

To achieve differential privacy, various privacy mechanisms introduce controlled randomness into data analysis. Common mechanisms include the Laplace mechanism and the exponential mechanism.

The Laplace mechanism (Dwork et al., 2016) means perturbing the output of a function with Laplace-distributed noise to achieve differential privacy. $Lap(b)$ to represent the noise sampled from a Laplace distribution with a scaling of b.

Definition 2.2. *For a function $f : D \to R$ over a dataset D, the mechanism Min provides the* ε*-differential privacy*

$$M(D) = f(D) + Lap(\frac{\Delta f}{\varepsilon})$$

In Definition 2.2, the parameter Δf refers to the global sensitivity, which determine how much perturbation is required for a particular query in a mechanism. This property is defined as the largest difference between the outputs of query q for any pair of neighbouring datasets which means that

$$\Delta_q = \max ||q(D) - q(D')||_1$$

where $||.||_1$ is the L_1 norm.

The Exponential Mechanism (McSherry and Talwar, 2007) is employed when you need to select an item from a set or make a decision based on data, and you want to ensure that the process is differentially private. This means that the probability of selecting one item over another should be adjusted to protect privacy while preserving the utility of the selection.

Definition 2.3. *Let $q(D,\phi)$ be a score function of dataset D that measures the quality of output ϕ, Δf represents the sensitivity of f. The exponential mechanism M satisfies* ε*-differential privacy if*

$$M(D) = (return \phi \propto exp(\frac{\varepsilon q(D,\Phi)}{2\Delta f})$$

2.2 Persistent Homology

As per the manifold distribution hypothesis (Goodfellow et al., 2016), real-world high-dimensional data is often situated on a lower-dimensional manifold hidden within the broader high-dimensional space. This underlying manifold is believed to possess a highly intricate non-linear structure, making its explicit definition challenging. Nonetheless, it is possible to scrutinize and analyze this manifold by considering its topological properties.

Topological Data Analysis (TDA) serves as a framework that integrates techniques from algebraic topology and statistical learning, providing a quantitative foundation for understanding these topological properties. Among the array of tools hailing from algebraic topology used in TDA, persistent homology stands out as a pivotal method. To compute the persistent homology of a space, it is necessary to initially express the space as a simplicial complex. Figure 1 shows example of simplices and one simplicial complex. A simplicial complex is essentially a collection of simplicial homology. Simplicial homology employs matrix reduction algorithms to assign K a family of groups, the homology groups. The d^{th} homology group $H_d(K)$ of K contains d-dimensional topological features, such as connected components $(d = 0)$, cycles/tunnels $(d = 1)$, and voids $(d = 2)$. Homology groups are typically summarised by their ranks, thereby obtaining a simple invariant "signature" of a manifold. For example, a circle in R^2 has one feature with $d = 1$ (a cycle), and one feature with $d = 0$ (a connected component).

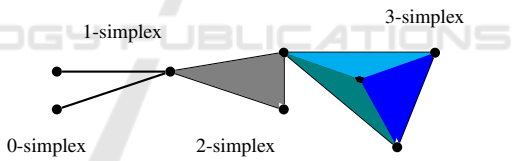


Figure 1: Example of one simplicial complex with different simplices. Two 0-simplex are vertex points, 1-simplex is a pair of vertex points which bound a line segment, a 2-simplex is a collection of vertex points which live on a triangle and a 3-dimensional simplex is a tetrahedron.

In practice scenarios, the underlying manifold M is often unknown and we are working with a point cloud $X := x_1, ..., x_n \subset R^d$ and a metric distance $X \times X \to R$ such as the Euclidean distance. Persistent homology adopts simplicial homology to this context. Instead of attempting to approximate M trough a single simplicial complex, which can be unstable due to the discrete nature of X, persistent homology monitors changes in homology groups across various scales of the metric. A distance function on the underlying space corresponds to a filtration of the simplicial complex. One common method of doing this is using the Vietoris-Rips construction. A *Vietoris-Rips* complex of parameter d is the simplicial com-

plex with finite set of points that has diameter at most d. The Vietoris–Rips complex of X at scale d contains all simplices of X whose elements $x_0, x_1, ...$ satisfy $dist(x_i, x_j) \leq d$ for all i, j.

We consider all distances d, then each homology appears at a particular value of d and disappear at another value of d. We represent the persistence of this hole as a pair, for example (d_1, d_2) and visualize this pair as a bar from d_1 to d_2. A collection of bars is a barcode. We can represent the persistent homology with a barcode or persistence diagram. A barcode represents each persistent generator with a horizontal line beginning at the first filtration level where it appears, and ending at the filtration level where it disappears.

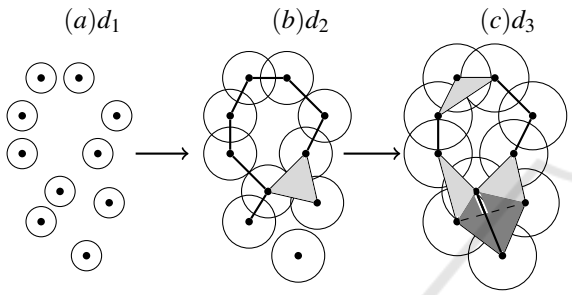


Figure 2: The three step filtration of Vietoris-Rips complex on the set of 10 points with increasing radius $0 < d_1 < d_2 < d_3$.

Figure 2 shows the The Vietoris–Rips complex of a point cloud X at different scales d_1, d_2 and d_3. As the distance threshold increases, the connectivity changes. The creation and destruction of d-dimensional topological features is recorded in the d^{th} persistence diagram which is showed in the Figure 3.

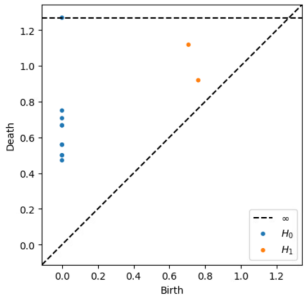


Figure 3: The persistent diagram corresponding to the filtration in the figure on top. Blue points represent persistent homology groups of dimension 0, and the orange ones of dimension 1.

A persistence diagram $P = (b_i, d_i)$ 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 "vectoriz" 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 P by the point measure $\sum_{i=1}^{m} \delta_{(b_i,d_i)}$ which is a measure on the set $T := \{(x;y) : 0 \leq x \leq y \leq \infty\}$ (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 P, P' be two persistence diagrams. Then the bottleneck distance between P and P' is defined as

$$W_{\infty}(P,P') := \inf_{\eta:P\to P} sup_{t\in P} ||t - \eta(t)||_{\infty}$$

where η ranges over bijections between P and P'.

A small perturbation in the input filtration leads to a small perturbation of its persistence diagram in the bottleneck distance. It means that for our work a key property of the bottleneck distance is stability property (Chazal et al., 2016). In this paper, our objective is to generate a differentially private persistence diagram for our dataset and subsequently generalize synthetic data based on its insights.

3 RELATED WORK

The field of publishing high-dimensional data has garnered significant attention from researchers seeking effective methods to balance the disclosure of information with the imperative to preserve privacy. Researchers have investigated the application of differential privacy mechanisms for publishing high-dimensional data. Dimensionality reduction is a pivotal step in managing high-dimensional datasets. A powerful approach of dimensionality reduction is the Bayesian network model proposed in (Zhang et al., 2017), in which Zhang developed a differentially private scheme PrivBayes for publishing high dimensional data. PrivBayes first constructs a Bayesian network to approximate the distribution of the original dataset. It adds noise into each marginal of the Bayesian network to guarantee differential privacy. It constructs an approximate distribution of the original dataset, and samples the tuples from the approximate distribution to construct a synthetic dataset. DP2-Pub algorithm (Jiang et al., 2023) is another method which is based on the Bayesian network and propose an invariant post randomization method (PRAM) to apply it to each attribute cluster. Another approach involves analyzing attribute correlations and utilizing a dependency graph to generate synthetic data that aligns with the joint distribution. (Chen et al., 2015). These solutions have a drawback as it significantly consume too

much of the privacy budget during network construction.

Computational topology and persistent homology (PH) have started gaining traction in several areas of data analysis. In (Alipourjeddi and Miri, 2023), PH is employed to assess synthetic datasets, enhancing accuracy. The integration of PH into graph analysis, as indicated in (Alipourjeddi and Miri, 2022) contributes to more precise synthetic datasets. Further, the fusion of PH with neural networks (Moor et al., 2020) facilitates dimensionality reduction. Additionally, incorporating differential privacy with PH enables differentially private Topological Data Analysis (Kang et al., 2023). These studies collectively concentrate on the manifold hypothesis and preserving topological structures of the input space.

In light of the above analysis, we present the novel method to generate synthetic data with differentially private persistence diagram. To the best of our knowledge, our work is the first attempt of publishing high dimensional dataset privately with topological approach.

4 METHODOLOGY

The well-known Manifold Hypothesis (Cao et al., 2020) states that in high dimensional data such as census data are concentrated on a low dimensional manifold in a Euclidean space embedded in the high-dimensional background space. Based on this hypothesis, we focus the following problem in this paper: We have a high-dimensional dataset with r attributes, and our strategy involves publishing and releasing the dataset to the public while satisfying differential privacy. We consider persistent homology to preserves the homology structure of our dataset accurately.

First, we propose how to add differential privacy into the persistence diagram of our dataset. In this step we need to consider the method which is sensitive to outlier. Due to the differential privacy principle, the specific data of any one individual should not have a significant effect on the outcome of the analysis to achieve privacy protection sensitivity (Avella-Medina, 2021). We examine the sensitivity of the bottleneck distance of persistence diagrams, which is the most widely used presentation of persistent homology. Because the magnitude of outlier-robustness affects the rate of sensitivity of the bottleneck distance, We use L^1-DTM in order to achieve a minimal sensitivity (Kang et al., 2023). We apply the exponential mechanism which utility function is defined in terms of the bottleneck distance of L^1-DTM persistence diagrams in order to produce differentially privatized persistence diagrams.

Second, we generate the synthetic dataset from the private persistence diagram. In this step, we choose randomly an initial hole from our persistence diagram or persistence barcode and sampling the attributes. We terminate this process when all attributes have been sampled.

4.1 Differentially Private Persistence Diagram Construction

In the realm of differential privacy algorithms, it is commonplace to quantify the extent to which the value of a statistic changes when altering a single point within a given dataset. This maximal potential change in the statistic is commonly referred to as the sensitivity of the statistic. It is necessary that the sensitivity goes to 0 as the size of the data grows.

In our work, we use a persistence diagram constructed from a dataset D as a statistic that provides an estimation of the homological structure underlying the data. To measure distances between persistence diagrams, we employ the bottleneck distance, defining a metric on the space of these diagrams. Consequently, when applying a differential privacy mechanism to persistence diagrams, our initial step involves estimating the sensitivity of persistence diagrams in terms of the bottleneck distance. Specifically, we need to analyse the maximum potential magnitude of the bottleneck distance, where the pair (D, D') denotes an adjacent pair of datasets. The sensitivity of the persistence diagrams of VietorisRips complexes cannot converge to 0 even if the size of data grows to infinity (Kang et al., 2023). Weighted VietorisRips filtration can be useful to highlight topological features against outliers and noise. In this regard, Chazal propose using the notion called distance to a measure (DTM), to get outlier-robust persistence diagrams(Chazal et al., 2017; Anai et al., 2020).

Definition 4.1. *Given a probability measure P , for $0 < m < 1$, the distance-to-measure (DTM) at resolution m is defined by*

$$\delta(x) = \delta_{P,m}(x) = \sqrt{\frac{1}{m}\int_0^m (G_x^{-1}(u))^2 du}$$

where $G_x(t) = P(||X - x|| \leq t)$.

The definition is L^2 type of DTM where the sensitivity is bounded by $O(n^{-1/2})$. We focus on L^1 type DTM for getting fastest decrease rate for sensitivity which is bounded by $O(n^{-1})$ (Kang et al., 2023).

To generate differential private persistence diagram, employ exponential mechanism with utility

function with the bottleneck distance

$$u_D(P_0,...,P_l) = \sum_{q=0}^{l} u_D^q P_q$$

where

$$u_D^q(P) = -d_B(P, P_q(D))$$

More specifically, we use negative bottleneck distance between private and non-private persistence diagrams as a utility function.

4.2 Synthetic Data by Differential Privacy Persistence Diagram

Our approach focuses on leveraging persistent diagrams to generate synthetic data that preserves the essential topological features of the original high-dimensional dataset. Analysing the differentially private persistent diagrams provide valuable insights into the homological characteristics inherent in the data privately.

In the first step, we translate and understand the persistent points corresponding to connected components. A clear trend emerges, showcasing the birth and death of connected components across different scales. Such persistence indicates the robustness of specific structural elements in the original dataset. Focusing on loops and voids, we identify regions of sustained persistence, signifying the presence of consistent topological patterns. Peaks and valleys in the diagrams provide valuable insights into the lifetimes of these features, aiding in the understanding of their relevance and stability. All analysing help us to formulate synthesis rules for generating synthetic data. For our datasets, we prioritize components with long persistence (we define a threshold for determining the persistence) and simulate the birth and death events of topological features. We need to ensure that the distribution of synthetic points are aligned with the topological structure. These results allowing us to recreate the topological patterns in a low-dimensional space. Secondly, we transform the synthetic points from the low dimensional space to match the dimensionality of the original dataset. To generate synthetic data, we apply the topological autoencoders method (TopoAE) (Moor et al., 2020). this paper evaluates the topological loss in term of distance matrix for each persistence diagram $A^X[\pi^X]$. Hence, $L_t = L_{X\to Z} + L_{Z\to X}$ (Moor et al., 2020) where,

$$L_{X\to Z} = \frac{1}{2}||A^X[\pi^X]] - A^Z[\pi^X]||^2$$

and

$$L_{Z\to X} = \frac{1}{2}||A^Z[\pi^Z]] - A^X[\pi^Z]||^2$$

The key idea for both terms is to align and preserve topologically relevant distances from both spaces.

5 EXPERIMENTAL EVALUATION

In this part, we carry out extensive experiments to demonstrate the performance of our mechanism and compare it with two benchmark approaches, PrivBayes (Zhang et al., 2017) and DP2-Pub (Jiang et al., 2023). Note that our comparative study focuses on PrivBayes and DP2-Pub because these methods share a common approach of decomposing high-dimensional data into a set of low-dimensional representations. The evaluation is based on the three real high-dimension datasets: the Adult dataset (Asuncion and Newman, 2007), the Poker-Hand dataset (Asuncion and Newman, 2007) and the Cleveland dataset (Asuncion and Newman, 2007).

Adult dataset contains personal information such as gender, salary, and education level of 45222 records extracted from the 1994 US Census, where each record has 15 attributes. Each record of Poker-Hand dataset is an example of a hand consisting of five playing cards drawn from a standard deck of 52. Each card is described using two attributes (suit and rank), for a total of 11 predictive attributes. There is one Class attribute that describes the"Poker Hand". Cleveland Heart Disease dataset presents the heart disease in the patient and contains 14 attributes.

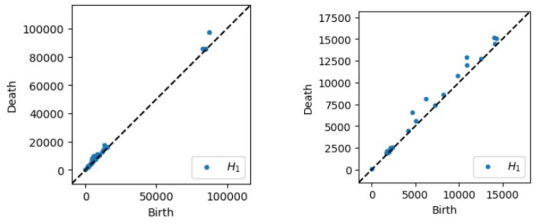
Initially, we assess the persistent homology of our datasets. Table 2 illustrates the count of homology in various dimensions of these datasets, presented on persistent diagrams. Notably, our analysis focuses on a subset of both the Adult and Poker-Hand datasets.

Table 1: Persintent barcodes in different dimensions.

Datasets	H_0	H_1	H_2
Adult	1000	27	0
Poker-Hand	800	863	564
Cleveland	303	199	34

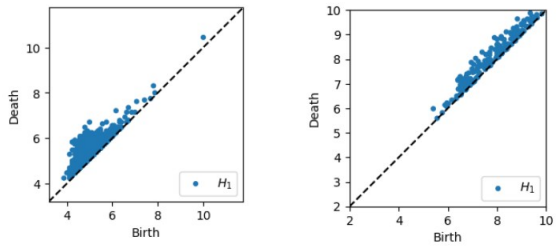
To obtain the private persistence diagrams for each dataset, we set the resolution of the L^1-DTM as $m = 0.05$, the privacy budget $\varepsilon = 1$ and we consider 1-dimension of topological feature in figures. Figure 4, Figure 5 and Figure 6 show the results of comparing the L^1-DTM persistence diagram corresponding to Adult, Poker-Hand and Cleveland datasets and their differentially private diagrams respectively.

After getting the differential private persistence diagram, we generate the synthetic data. We use Monte carlo (MC) method to align the distributions. The threshold for persistence of component H_0 are



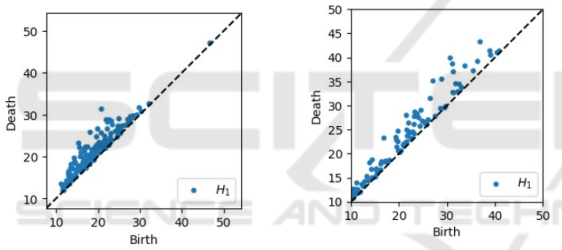
(a) Original PD. (b) Private PD.

Figure 4: Persistent diagrams(PDs) of Original and private Adult dataset.



(a) Original PD. (b) Private PD.

Figure 5: Persistent diagrams(PDs) of Original and private Poker-Hand dataset.



(a) Original PD. (b) Private PD.

Figure 6: Persistent diagrams(PDs) of Original and private Cleveland dataset.

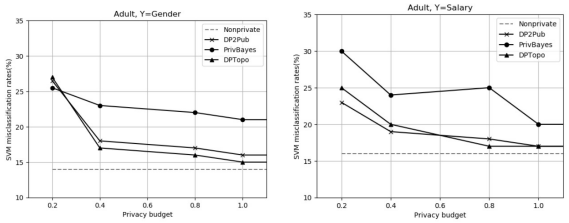


Figure 7: Multiple SVM classifiers on Adult dataset.

vary for each datasets. We set the threshold for Adult, $T_{Adult} = 50000$, for Poker-Hand $T_{PK} = 5$, and for Cleveland $T_{Celeland} = 25$. Also, we use our differentially private persistence diagram in the TopoAE algorithm and generate the synthetic dataset.

For the second task, we evaluate the performance of PrivBayes, DP2-Pub1, our work DPTopo, and Non-Private (no DP is considered) for SVM classification. Figure 7 , Figure 8 and Figure 9 show the

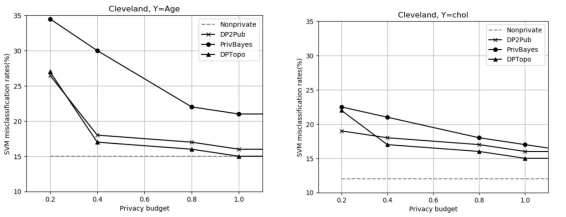


Figure 8: Multiple SVM classifiers on Cleveland dataset.

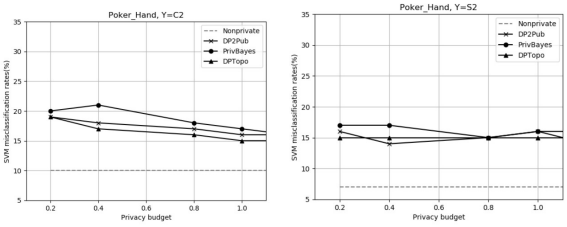


Figure 9: Multiple SVM classifiers on Poker-Hand dataset.

misclassification rate of three datasets at different levels of privacy protection or privacy budgets. The misclassification rate of the original dataset (denoted as Non Private) stands for the best performance we can achieve. It means that the lowest misclassification rate is better result. We observe that our result outperforms compared to others. Moreover, A lower privacy budget typically means stronger privacy protection but may lead to a higher misclassification rate. Another notable observation is that the misclassification rate of SVM decreases with the increase in the privacy budget. This finding is in line with the theoretical expectation that as the privacy budget expands, privacy protection weakens, leading to an increase in the availability of data and a consequent reduction in misclassification rates.

6 CONCLUSIONS

In this paper, we presented a novel approach to generating private synthetic data leveraging insights from persistent homology. Our methodology successfully replicated the essential topological features observed in the high-dimensional original dataset. By applying the weighted Vietoris-Rips complex algorithm, we computed persistent homology and extracted meaningful diagrams. We produced differential private persistence diagrams by applying exponential mechanism. We used a negative bottleneck distance between private and non-private persistence diagram as a utility function. we used L^1-DTM to achieve minimal sensitivity. For generating synthetic data based on differentially private persistence diagram, we kept similar birth and death events for persistent points with the same distribution. We

transformed the low-dimensional space with synthetic points to high-dimensional space by topological autoencoders method. Our research highlights the efficacy of persistent homology-inspired synthesis in producing differential private synthetic data with significant topological structures. As the field of Topological Data Analysis (TDA) progresses, exploring alternative metrics for computing the persistence diagram, such as the persistence landscape, becomes crucial. Adopting an alternative privacy framework like zero-concentrated Differential Privacy has also shown to yield lower errors in the privacy mechanism.

ACKNOWLEDGEMENTS

This work was supported in parts by funds from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery and the Canada First Research Excellence Fund (CFREF) Bridging Divides programs.

REFERENCES

Alipourjeddi, N. and Miri, A. (2022). Publishing private high-dimensional datasets: A topological approach. In *the 2022 International Wireless Communications and Mobile Computing (IWCMC)*, pages 1142–1147. IEEE.

Alipourjeddi, N. and Miri, A. (2023). Evaluating generative adversarial networks: A topological approach. In *the 2023 International Conference on Computing, Networking and Communications (ICNC)*, pages 202–206. IEEE.

Anai, H., Chazal, F., Glisse, M., Ike, Y., Inakoshi, H., Tinarrage, R., and Umeda, Y. (2020). Dtm-based filtrations. In *Topological Data Analysis: The Abel Symposium 2018*, pages 33–66. Springer.

Asuncion, A. and Newman, D. (2007). UCI machine learning repository.

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

Binjubeir, M., Ahmed, A. A., Ismail, M. A. B., Sadiq, A. S., and Khan, M. K. (2019). Comprehensive survey on big data privacy protection. *IEEE Access*, 8:20067–20079.

Cao, W., Yan, Z., He, Z., and He, Z. (2020). A comprehensive survey on geometric deep learning. *IEEE Access*, 8:35929–35949.

Chazal, F., De Silva, V., Glisse, M., and Oudot, S. (2016). *The structure and stability of persistence modules*. Springer.

Chazal, F., Fasy, B., Lecci, F., Michel, B., Rinaldo, A., Rinaldo, A., and Wasserman, L. (2017). Robust topological inference: Distance to a measure and kernel distance. *The Journal of Machine Learning Research*, 18(1):5845–5884.

Chen, R., Xiao, Q., Zhang, Y., and Xu, J. (2015). Differentially private high-dimensional data publication via sampling-based inference. In *the proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 129–138. ACM.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2016). Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3):17–51.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Jiang, H., Yu, H., Cheng, X., Pei, J., Pless, R., and Yu, J. (2023). DP2-Pub: Differentially private high-dimensional data publication with invariant post randomization. *IEEE Transactions on Knowledge and Data Engineering*.

Kang, T., Kim, S., Sohn, J., and Awan, J. (2023). Differentially private topological data analysis. *arXiv preprint arXiv:2305.03609*.

Mahanan, W., Chaovalitwongse, W. A., and Natwichai, J. (2021). Data privacy preservation algorithm with *k*-anonymity. *World Wide Web*, 24:1551–1561.

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

Moor, M., Horn, M., Rieck, B., and Borgwardt, K. (2020). Topological autoencoders. In *the proceeding of International conference on machine learning*, pages 7045–7054. PMLR.

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

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2017). PrivBayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)*, 42(4):1–41.