An Algebraic Topological Perspective to Privacy

Alberto Speranzon   Shaunak D. Bopardikar

***Abstract*— We cast the classic problem of achieving k-anonymity for a given database as a problem in algebraic topology. Using techniques from this field of mathematics, we propose a framework for k-anonymity that brings new insights and algorithms to anonymize a database. We focus on the simpler case when the data lies in a metric space which is instrumental in introducing the main ideas and notation. Specifically, by mapping a database to the Euclidean space and by considering the distance between datapoints, we introduce a simplicial representation of the data and show how concepts from algebraic topology, such as the nerve complex and persistent homology, can be applied to efficiently obtain the entire spectrum of k-anonymity of the database for various values of k and levels of generalization. For this representation, we provide an analytic characterization of conditions under which a given representation of the dataset is k-anonymous. We introduce a weighted barcode diagram which, in this context, becomes a computational tool to tradeoff data anonymity with data loss expressed as level of generalization. Some simulations results are used to illustrate the main idea of the paper.**

***Index Terms*— Privacy, k-anonymity, persistent homology.**

I. INTRODUCTION

Recent times have seen a revolution in computing technologies. Third-party computing services, such as the Cloud, have been creating new paradigms for both, data storage and computation. Such technologies require one to repeatedly revisit the basic question of "How do we protect the data from a privacy perspective?". Although this problem originated in database theory and computer science applications, it has recently expanded to domains such as systems and control. As the world is becoming more connected via millions of sensors [1], privacy is becoming a top priority in the control community.

This paper considers static data collected within a database that, in the context of cyber-physical systems, could represent log/monitoring data that needs to be analyzed offline to determine overall system performance, enterprise level fault detection and propagation, forensic analysis, etc. Although several approaches have been proposed to address different aspects of the privatization of data within a database, this paper provides a novel framework and perspective to address the most classical version of the problem using concepts and tools from algebraic topology.

Among several methods that have been developed for database privacy, a classic and popular approach is *k-anonymity*, which is a mechanism for protecting privacy of individuals represented as entries in a database [2]. For a given value of k, the original database is modified so that at least k individuals in the database have identical quasi-identifiers. This is achieved by generalizing numeric or text attributes: for example, the ZIP code can be generalized so that a certain number of the least significant digits are suppressed as $46532 \rightarrow 465{*}{*}$, age could be generalized to intervals, $35 \rightarrow [30, 40]$, and the gender could be generalized as $\{M, F\} \rightarrow Person$. The problem of computing an optimal k-anonymous version of a database has been shown to be NP-hard [3]. However, efficient algorithms such as Incognito [4] and its variants or greedy clustering-based algorithms [5] have been proposed to achieve k-anonymity. A multi-dimensional extension of the greedy approaches has been addressed in [6], which results into a representation of the database that is reminiscent of classic grid-based paintings by Mondrian. In multi-dimensional settings, a data aggregation scheme based on Hilbert curves has been proposed in [7].

Algebraic topology is a branch of mathematics that leverages tools and concepts from abstract algebra to study topological spaces. For example, a simple model for sensor network is a set of points in a (multidimensional) space in which two points (or sensors) are *neighbours* if they are within a specified distance of each other. Then, concepts from algebraic topology, such as homology, have been used to detect *holes* in sensor networks [8], [9]. Distributed algorithms to localize holes in sensor networks using related concepts have been addressed in [10] and in [11]. Recently such methods have been also used for filtering and position estimation in [12], [13].

One concept of privacy which has been applied to several control problems is that of differential privacy [14]. Informally, this concept means that for a given database, if any single individual is removed from the database, then no output of a computation run on the database would become significantly more or less likely. This concept has been applied to achieve differential privacy of Kalman filtering and estimation problems [15], to ensure a level of truthfulness in electric vehicle charging applications [16], to achieve average consensus in a private manner [17], to name a few. While the concept of differential privacy is very general and is applicable to dynamic databases, the resulting mechanism relies very strongly on the type of function/query that needs to be computed on the database. In contrast, k-anonymity is a static concept but is independent of any computation to be carried out on the database and therefore suitable in the context of offline analysis. That said, there are situations when k-anonymity is not sufficient and individuals can be re-identified despite anonymization. Although approaches to address this gap have been considered (cf. [18]), k-anonymity

This work was supported by United Technologies Research Center.

Alberto Speranzon was with United Technologies Research Center at the time when this work was developed. He is now with Honeywell Aerospace – Advanced Technology, alberto.speranzon@honeywell.com. Shaunak D. Bopardikar is with United Technologies Research Center, Inc. bopardsd@utrc.utc.com.

is still widely used.

This paper introduces a novel perspective to data privacy based on algebraic topology. In particular, we address the case when the data lies in a metric space. By defining two datapoints that lie within a specified radius as neighbours, we show that the representation falls within the natural setting of a Čech (or in general, a Nerve) complex [19]. By increasing the radius (generalization), we show that the sequence of Čech complexes result into a *filtration*, i.e. nested complexes. This further implies that tools such as persistent homology can be applied to efficiently obtain the entire spectrum of k-anonymity of the database for various values of the generalization. The benefit of this approach is that once the family of complexes is built, for various generalization values, we can apply fast and scalable persistent homology algorithms, such as Perseus [20] to determine the tradeoffs. Furthermore, the persistent diagram not only provides the tradeoffs between a generalization and the value of k, but also show how many equivalent classes are formed for a given generalization achieving a certain k-anonymity, a metric that has an impact on the anonymized data quality [21]. For this representation, we provide an analytic characterization of conditions under which a given representation of the dataset is k-anonymous. We then briefly discuss how this method can be extended to the general case of a mix of categorical and metric data.

This paper is organized as follows. The problem formulation is presented in Section II. Background results and concepts from algebraic topology are reviewed in Section III. The proposed approach is presented in Section IV for numerical data along with some simulation results.

II. Problem formulation

Let us consider a database table $T(A_1, A_2, \ldots, A_m)$ consisting of N rows, where each $A_i \in \mathcal{D}$ are various attributes that in general can take the form of *numeric* and/or *categorical* values, i.e., the domain $\mathcal{D}$ can either be a set of discrete or continuous values. Without loss of generality, we can identify with $Q_T = \{A_1, A_2, \ldots, A_d\}$ a set of d attributes that we define as *quasi-identifiers*, namely attributes that can be joined with external information/databases so that private information can be obtained. Typical examples of private data that could be obtained are names of individuals, salaries, etc.

Another database table $\bar{T}(\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_m)$ consisting of N rows is said to be a *generalization* $\bar{T} = G(T)$ of the table T if, for every row T_j of T,

$$Q_{T_j} \subset Q_{\bar{T}_j} .$$

In this paper, as said previously, we will be focusing on the concept of k-anonymity for privacy, that is formally defined as follows.

Definition 2.1 (k-anonymity [2]): Consider a generalized database $\bar{T}$ and a quasi-identifier set $Q_{\bar{T}}$. The set $Q_{\bar{T}}$ is said to have the k-anonymity property if each unique tuple in the projection of $\bar{T}$ on $Q_{\bar{T}}$ occurs at least k times in $\bar{T}$. □

Given a database T, the problem of k-anonymity is thus to determine a generalization function $G(.)$ so that the resulting database $\bar{T} = G(T)$ is k-anonymous. Clearly, one may simply generalize every entry and find the smallest set that generalizes every row of T. However, this trivial method would completely destroy the information content in the original database. The problem is how to minimize such as *over-generalization* of the quasi-identifiers. This notion will be made precise in Section IV.

III. Background on Topological Methods

This section provides a summary of some concepts from algebraic topology (cf. [19], [22]) that will be used in our approach.

Definition 3.1 (Čech complex): Given a collection of points $\{x_i\} \in \mathbb{R}^n$, the *Čech complex* is the abstract simplicial complex whose k-simplices are determined by unordered $(k+1)-$tuples of points $\{x_i\}_0^k$ whose closed ϵ-ball neighborhoods have a point of common intersection.

A. Simplicial Homology

Homology is an algebraic characterization of "holes" in a topological space. The central notion is that of a boundary homomorphism, which in the context of simplicial complexes, encodes how simplicies are attached to their lower dimensional facets. To define (simplicial) homology, of a complex C, we choose an ordering of each simplex, in the same way directed graphs are ordered. Given such ordering we consider $\mathbb{R}$-vector spaces $\mathcal{C}_k(C)$ with basis the oriented k-simplicies. We thus have that $\mathcal{C}_\bullet$, forms a sequence of vector spaces, which we call chain complex. A *boundary homomorphism* is defined as the linear map $\partial_k : \mathcal{C}_k(C) \to \mathcal{C}_{k-1}(C)$ given by associating each basis element of $\mathcal{C}_k(C)$ to the formal sum of its (oriented) faces of dimension $k-1$. The boundary operator $\partial = \{\partial_k\}$ thus encodes the assembly instructions of C. It turns out that the k^{th} homology group of the complex C, $H_k(C)$ is given by

$$H_k(C) = Z_k/B_k = \ker \partial_k / \mathrm{im}\, \partial_{k+1} .$$

The group $Z_k = \ker \partial_k$ is called the k-th cycle group and its elements (chains) represent k-cycles. We have that $B_k = \mathrm{im}\, \partial_{k+1}$ is the k-th boundary group whose elements are k-boundaries. The quotient space $H_k(C)$ thus represents all the k-cycles that are not boundaries of $k+1$ simplices, namely cycles that represent k-dimensional "holes". The homology of the complex C is then $H_\bullet(C) = \{H_k(C)\}$.

In this paper, we will use the notion of dimension of the k-th homology, that is the dimension of the vector space $H_k(C)$, $\mathrm{dim} H_k(C)$. In particular, the k-th homology group $H_k(C)$ is said to be *trivial* if $\mathrm{dim} H_k(C) = 0$.

B. Persistent Homology

Let us consider a sequence of complexes C^ϵ with $\epsilon = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_M\}$, more specifically the sequence of Čech complexes $\{C^{\epsilon_i}\}_{i=1}^N$, for increasing $\epsilon_i \in \mathbb{R}_{\geq 0}$, $\epsilon_i \leq \epsilon_j$ for $i \neq j$. There are clearly inclusion maps between such complexes

$$C^{\epsilon_1} \overset{\imath}{\hookrightarrow} C^{\epsilon_2} \overset{\imath}{\hookrightarrow} \cdots \overset{\imath}{\hookrightarrow} C^{\epsilon_{M-1}} \overset{\imath}{\hookrightarrow} C^{\epsilon_M} .$$

Rather than studying the homology of each complex for each value of the parameter ϵ, one can then study the homology

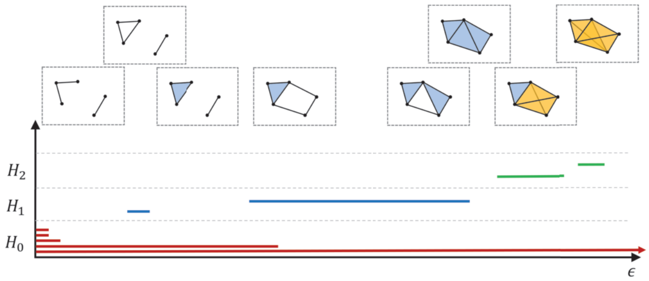

Fig. 1: Example of the barcodes.

Age	ZIP Code	Salary
25	47677	$47,000
22	47602	$32,000
24	47678	$52,000
43	47905	$151,000
52	47909	$145,000
38	47906	$98,000
47	47605	$110,000
36	47673	$92,000
32	47607	$115,000

(a) Example data set with $Q_T = \{\text{Age}, \text{ZIP Code}\}$.

Age	ZIP Code
[22-25]	[47602-47678]
[22-25]	[47602-47678]
[22-25]	[47602-47678]
[38-52]	[47905-47909]
[38-52]	[47905-47909]
[38-52]	[47905-47909]
[32-47]	[47605-47603]
[32-47]	[47605-47603]
[32-47]	[47605-47603]

(b) 3-anonymized table

TABLE I: Sample data set for illustrative purpose.

of the inclusions $\imath : H_\bullet(C^{\epsilon_i}) \rightarrow H_\bullet(C^{\epsilon_j})$ for $i < j$. Such maps are important as they capture topological features that persist over the parameter space.

The dimension of the homology groups as a function of the single parameter ϵ can be plotted in a diagram, called the *barcodes diagram*, see [22]. We show a simple example in Figure 1 where ϵ is the radius of a ball around each vertex, and where there is a k-simplex whenever $k+1$ circles have non-empty intersections. For small values of ϵ, we have that the number of connected components, $\dim H_0$, is the number of vertices (0-simplicies), and as ϵ increases, more vertices get connected, resulting in components to merge till a single connected component is obtained. For a small value of ϵ there are no high dimensional holes, however, at some point, before the first 2-simplex (blue triangle) gets filled in, such 2-simplex is not filled and generates a hole that quickly disappears. At a later value of ϵ, a large hole is formed at the time when the vertices form a single large connected component. As the ϵ parameter increases further at some point the "middle" hole gets filled and the dimension of the first homology, $\dim H_1$ becomes zero again. As ϵ further increases, tetrahedrons appear, first with an empty volume, namely a void, that disappear. Higher dimensional holes will likely occur, but we did not depict them. As ϵ further increases the complex will have trivial higher homology groups and only have a single connected component.

In this paper we will make use the persistent homology in the context of k-anonymity.

IV. k-Anonymity via Persistent Homology: Numerical Attributes

In this section, we describe the algebraic topological approach toward achieving k-anonymity. In this section, we will restrict our attention to the case of the attributes in the quasi-identifier set Q_T being all numeric/continuous variables, such as Age, Salaries, Taxes paid in a year, etc. In this case, we can clearly represent the set Q_T as a $|Q_T|$-dimensional vector. We will assume that all the vectors are elements of a real vector space.

Table I (a) shows a sample dataset with two quasi-identifiers $Q_T = \{\texttt{Age}, \texttt{ZIP Code}\}$. Table I (b) shows a 3-anonymous version of the dataset, which we will compute in the remainder of this paper. In the following, we will often refer to entries of the table T as *points* due to the fact that they take real values. Also, note that without any loss of generality, we can consider the data to take values within the hypercube $\mathcal{M} = [0, 1]^{|Q_T|}$, since one can always normalize the data accordingly.

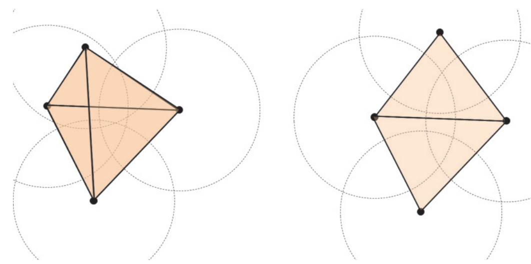

Fig. 2: In (a), we show a anonymity 4-simplex representing the fact that there exists an ϵ such that the 4 points can be anonymized. In (b), we show an example where there is no anonymity 4-simplex (indeed we have only two 2-simplices) as the selected ϵ is not large enough, or equivalently the generalization is high enough, to ensure 4-anonymity.

The first definition is a direct application of the Čech complex as summarized in Section III. In this paper, we use this structure to capture the k-anonymity property of the data.

Definition 4.1 (Anonymity Complex): Given a table T with N rows and a set of quasi-identifier Q_T, let us consider the N points $\{p_i\}_1^N \in \mathcal{M}^N$. We define an *anonymity complex* $\mathcal{C}(p)$ the simplicial complex whose k-simplicies are determined by $(k+1)$ points $\{p_{i_0}, p_{i_1}, \ldots, p_{i_k}\}_0^k$ such that closed ϵ-ball neighborhoods centered around these points have at least one intersection point. We call the radius ϵ the *global generalization* strategy. □

We now introduce an important building block.

Definition 4.2 (Anonymity k-simplex): Given a global generalization ϵ, we say that k points have the k-anonymity property if all the closed ϵ-ball neighborhoods of the k points all intersect in at least a point. In this case, we have that the k points form a k-simplex, which we term as an *anonymity k-simplex* and denote with S_k. □

Figure 2(a) shows an example of an anonymity 4-simplex for a given global generalization ϵ while Figure 2(b) shows an example where for the same value of ϵ, 4-anonymity cannot be achieved.

The following is a useful test to determine whether k-anonymity can be achieved for a given value of ϵ or not.

Lemma 4.1: Given a set of points $p = \{p_i\}_1^N$ corresponding to N rows of the table T, given a global generalization ϵ and the corresponding anonymity complex $C(p)$, we have that the points $\{p_i\}_1^N$ have the k-anonymity property if and only if

$$C(p) = \bigcup_i S_{\ell_i}$$

where S_{ℓ_i} is the i-th anonymity ℓ_i-simplex with $\ell_i \geq k$ for $i \in \mathbb{N}_0$. We say in this case that the anonymity complex achieves k-anonymity.

Proof: The points $\{p_i\}_1^N$ have the k-anonymity property if and only if we can sub-divide the set of points into subsets such that

$$\{p_i\}_1^N = \underbrace{\{p_i\}_1^{i_1}}_{\Sigma_{\ell_1}} \cup \underbrace{\{p_i\}_{i_1+1}^{i_2}}_{\Sigma_{\ell_2}} \cup \cdots \cup \underbrace{\{p_i\}_{i_{N-1}+1}^{N}}_{\Sigma_{\ell_N}},$$

and to each subset $\Sigma_{\ell_i}(p)$, we can associate an anonymity ℓ_i-simplex S_{ℓ_i} with $\ell_i \geq k$ for any i (for given fixed ϵ) and $S_{\ell_i} \cap S_{\ell_j} = \emptyset$ for $i \neq j$.

Given the previous set relations, we have that complex C associated with the set of points $\{p_i\}$ is then given by the union of the S_{ℓ_i}. ■

This result then establishes a natural connection between the properties of the anonymity complex, in terms of some of its subcomplexes, and the k-anonymity property.

We further explore how topological properties of the anonymity complex are related to the k-anonymity property and how we can leverage that to find an "optimal" generalization. We first establish some topological properties of C and then define what we mean with "optimal" generalization.

Proposition 4.1: An anonymity complex C, for a given ϵ, has the k-anonymity property if and only if its homology groups $H_n(C)$ are trivial for any $n > 0$, and every connected components is an ℓ-simplex with $\ell \geq k$. Furthermore, when this is the case the number of equivalence classes generated by using the ϵ generalization is given by the dimension of the zero-th homology, $\dim H_0(C)$. □

Proof: (If) From Lemma 4.1, we know that we can decompose C into a finite number of disjoint anonymity k-simplices. It is known [19] that in this case $H_n(C) = \bigoplus_i H_n(S_{\ell_i})$, namely the n-th homology of C is given by the direct sum of the n-th homology of the anonymity k-simplices. As the k-simplices are simply connected spaces and contractible, they have trivial high order ($n > 0$) homology groups and $H_0(S_{\ell_i}) \approx \mathbb{Z}$ for every i.

(Only if) Let us assume that C has $H_n(C) = \{0\}$ for $n > 0$, and $H_0(C)$ is non-trivial, and in particular let us assume that $H_0(C) \approx \mathbb{R}^s$. This means that the complex C has s connected components. From the hypothesis that the connected components are ℓ-simplices with $\ell \geq k$, we know that each component is a anonymity simplex. Thus the anonymity complex C has the k-anonymity property. ■

We are now able to connect topological properties of the anonymity complex with the k-anonymity property. Of course, as it can be seen from Proposition 4.1, the result still depends on ϵ, namely the generalization.

When anonymizing a dataset, one is typically interested in "corrupting" the data by the least amount. Indeed, if one carefully thinks about the k-anonymity problem, it is always possible to find a large enough k that makes the data anonymous, i.e., if one makes the extreme choice of $k = n$, then the entire data will be in one equivalence class and thus, k-anonymity will be achieved. The issue with this is that the information contained in the data will be completely lost.

Algorithm 1 k-anonymity via Persistence Homology

> **Inputs:** $p = \{p_i\}_1^N \in \mathbb{R}^{d\times N}$, Parameter: k, Radii: $\{\epsilon_1, \ldots, \epsilon_M\}$
> Re-scale the dataset into the unit cube $[0,1]^d \subset \mathbb{R}^{d\times N}$.
> **for** every value of $\epsilon \in \{\epsilon_1, \ldots, \epsilon_M\}$ **do**
>   Construct the anonymity complex $C^\epsilon(p)$
> **end for**
> Compute *weighted* persistent homology
> **return** Complete bar code diagram

In the context of this paper, as the generalization is parametrized by ϵ, we are interested for find the smallest value of ϵ that gives k-anonymity. We then have the following definition.

Definition 4.3 (Minimal Anonymity Complex): Given an anonymity complex $C(p)$ associated to a set of points, lets us denote with C^ϵ the anonymity complex for a given generalization ϵ.

We define as *minimal anonymity complex* the following object

$$C^{\epsilon^*} = \min_\epsilon C^\epsilon,$$

such that C^{ϵ^*} achieves k-anonymity. □

Even without minimization over ϵ, the k-anonymity problem known to be an NP-hard problem, and so it is clear that we cannot easily find the minimal anonymity complex. To find an approximate solution to the problem, we instead study the *persistent homology of* C^ϵ. In particular, in this paper, we adapt the idea of persistent homology as a tool to provide the full spectrum of k-anonymization one can obtain. Our approach is summarized in Algorithm 1.

Note that the parameter ϵ induces a family of complexes such that $C^{\epsilon_1} \overset{\imath}{\hookrightarrow} C^{\epsilon_2} \overset{\imath}{\hookrightarrow} \cdots \overset{\imath}{\hookrightarrow} C^{\epsilon_M}$, where $\epsilon_i \leq \epsilon_j$, for any $i < j$, and thus we recover the same setting as in the persistent homology. Formally we have a ϵ-based filtration [22]. The idea here is to leverage the barcodes or persistent diagram to extract regimes of interests, namely anonymity strategies – values of ϵ – that lead to k-anonymity for different values of k.

Such a persistent diagram has two specific features:

- Each bar in H_0 diagram has a *weight* corresponding to the number of elements in the connected components;
- Given a value k, we only consider bars that have at least k elements.

To illustrate the proposed approach, consider the sample dataset shown in part (a) of Table I. For the case of 3-anonymity in particular, the lower center figure depicts H_0 while the one in top center shows the H_1. We can see that a *hole* gets created for the values of the radius approximately in $[0.11, 0.13]$ and then it gets *filled* thereby making the dataset 3-anonymous. But again in the interval $[0.17, 0.19]$, a hole gets created around when the complex $[1, 2, 3, 7, 8, 9]$ gets formed. After ϵ increases more, this hole disappears and we are left with an anonymity 3-simplex $[4, 5, 6]$ and an anonymity 6-simplex $[1, 2, 3, 7, 8, 9]$.

We can clearly see three regimes where 3-anonymity is possible. The first one (as indicated) has three classes with three elements in each. The second regime corresponds to the values of radius in the interval $[0.19, 0.4]$ which has only two equivalence classes, and the third one (radius greater than 0.4) being the trivial solution where there is only one class with nine elements.

For minimal data quality loss, the interval $[0.17, 0.19]$ is the best solution as 3-anonymity can be reached with largest number of classes. Mapping the first one back to the dataset yields a 3-anonymous version of the dataset shown in part (b) of Table I.

In Figure 3 shows that 2-anonymity can be achieved for generalizations $\epsilon > 0.08$. For generalizations with $\epsilon \leq 0.08$, we have only one anonymity 2-simplex and the rest of the data will be just points, thus 2-anonymity cannot be achieved. The rightmost plot shows that 4-anonymity can be reached for $\epsilon > 0.4$. Note that even if there is an anonymity 6-simplex in the interval $[0.19, 0.4]$, there is no way for the $[4, 5, 6]$ complex to achieve 4-anonymity and thus the generalizations in the interval $[0.19, 0.4]$ will not yield 4-anonymity. For $\epsilon > 0.4]$, we can clearly achieve 4-anonymity, but as everything gets into a single class, all the data will be generalized to the same record and thus data quality is compromised.

Remark 4.1 (Advantages of proposed approach): The main advantage of the method proposed is that it enables us not only to find a k-anonymization for a fixed k (if it exists), but also to provide alternative regimes that can be especially useful when k-anonymity cannot be achieved for the given k. This generally is not something that other algorithms, such as [4], [6], [7] directly provide. Indeed, one would need to run the same algorithms for various values of k to obtain the same tradeoff picture as we obtain. The persistent diagram we consider in this paper is instead computed in *one shot* from the filtration $\{C^\epsilon\}$. Furthermore, we leverage very scalable algorithms for such computation based on discrete Morse theory, see [23], [20]. Also, note that not only the persistent diagram allows us to determine the right regime that gives the desired k-anonymity, but we can also look at the other important tradeoff parameter such as the number of classes. For a given k, it is indeed possible to find various generalizations ϵ that meet the k-anonymity requirement, however some might lead to equivalence classes with many more than k elements that is in general not desirable.

Remark 4.2 (Extension to Categorical/Mixed Data): The methodology we have described in the previous section has nice properties but also limitations: it is restricted to numerical attributes where the notion of a radius is well defined and "intersecting balls" (or polytopic approximations) in high dimensional spaces can be computationally challenging. We discuss, briefly, here ideas on how to extent the previous setting to scenarios where attributes are categorical (such gender, country of citizenship, phone numbers).

The starting point in this case is a *generalization* tree, that given a set of categorical attributes provides a common generalization. For example, given the attributes $\{$*Male, Female*$\}$ as leaves of a tree, their generalization could be $\{$*Person*$\}$, appearing as root in a tree. For countries, such as $\{$*USA, Canada, Mexico Brazil, Argentina*$\}$ we could have $\{$*USA, Canada, Mexico*$\}$ generalizing to $\{$*North America*$\}$, $\{$*Brazil, Argentina*$\}$ to $\{$*South America*$\}$ and in turn $\{$*North America, South America*$\}$ generalizing to $\{$*America*$\}$. In this context, $\{$*USA, Canada, Mexico*$\}$ forms a generalized anonymity 3-simplex when generalized as $\{$*North America*$\}$. This simple example shows it is possible to build anonymity complexes based on generalization trees.

What changes, compared to the previous discussion, is the inclusion relations among the complexes for different generalizations. As one can imagine the discussion here is more complicated and requires to consider and extension of the persistent homology, called *zig-zag* homology [24]. It suffice to say, here, that it is possible to build (zig-zag) persistent diagrams that are similar to the one described in the previous sections (see Figure 3) thus enabling to determine various anonimity regimes. For a more detailed discussion please see the authors' paper [25].

V. Conclusion

This paper introduced a new perspective to k-anonymity in data privacy based on algebraic topology. In particular, we addressed the case when the data lies in a metric space. We demonstrated how tools such as persistence homology can be applied to efficiently obtain the entire spectrum of k-anonymity of the database for various values of the radius of proximity. For this representation, we provided an analytic characterization of conditions under which a given representation of the dataset is k-anonymous. Finally, we discussed how this method can be extended to address the general case of a mix of categorical and metric data.

In future, it would be interesting to investigate other notions of privacy using these tools. In particular, applicability of such techniques to dynamic databases would be an interesting case which naturally arise in control applications.

VI. Acknowledgments

The authors would like to thank Vidit Nanda for discussions on zig-zag persistent homology and the Perseus software used in this paper to compute persistent homology.

References

[1] "Design technology for the trillion-device future," http://youtu.be/ViJ3SH5t4Ys, presented at the DARPA Wait, What? conference, St. Louis, MO, 2015.

[2] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[3] A. Meyerson and R. Williams, "On the complexity of optimal k-anonymity," in *ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2004, pp. 223–228.

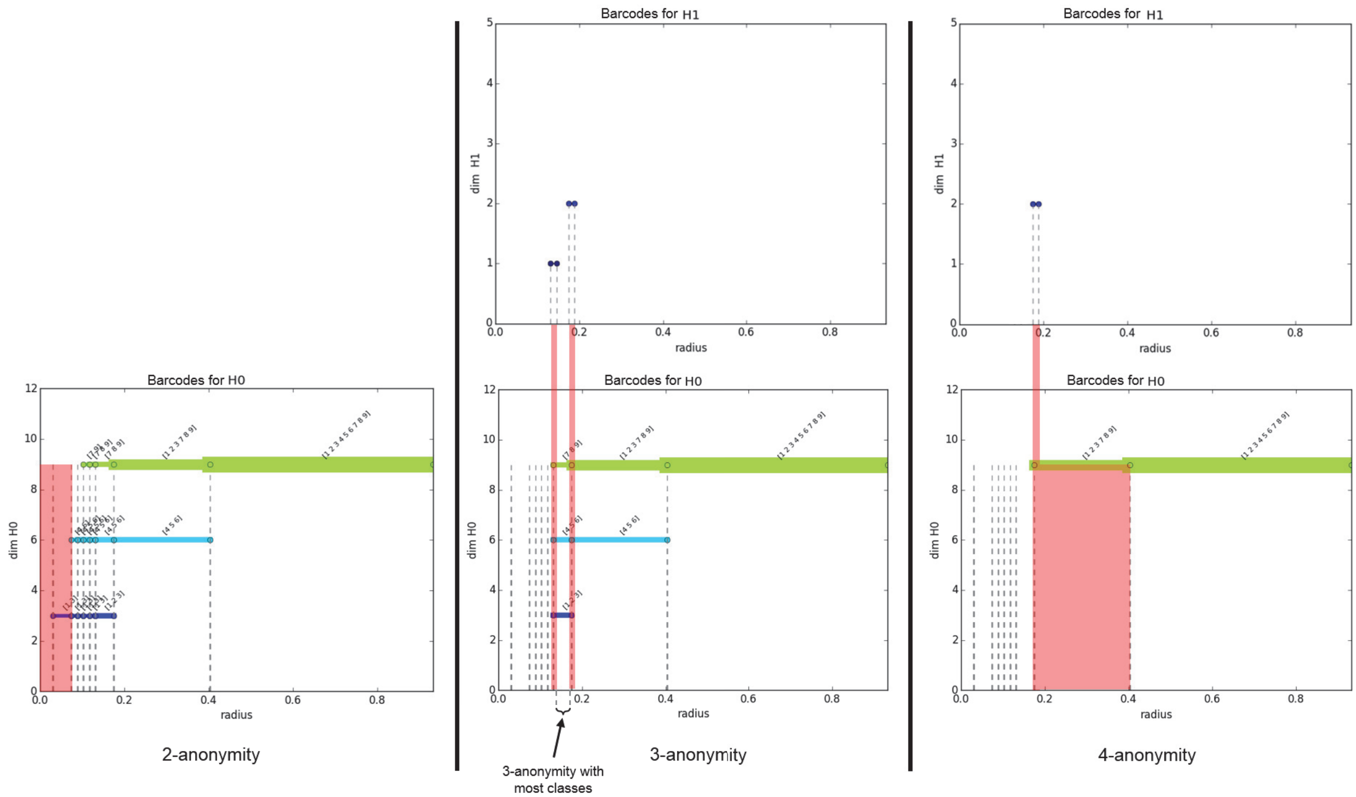

Fig. 3: Weighted barcode showing the full spectrum of anonymity regimes. Although the barcode is only a single diagram, we split it here into three diagrams where we only show the simplices that meet the requirements of the k-anonymity indicated underneath each figure. With red bars we show regimes that cannot achieve 2,3,4-anonymity. These correspond to either situations where not all the k-simplices admit trivial higher order homology groups, as is the case for 3-anonymity between $[0.11, 0.13]$ and $[0.17, 0.19]$, not all the simplicies are ℓ-simplices with $\ell \geq k$. This happens for example in 2-anonymity in the interval $[0, 0.08]$ and between $[0.19, 0.4]$ for 4-anonymity. We only show here $\dim H_0$ and $\dim H_1$ as higher order homology groups are trivial for this simple example.

[4] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient full-domain k-anonymity," in *ACM SIGMOD Int. Conf. on Management of Data*, 2005, pp. 49–60.

[5] J. Byun, A. Kamra, E. Bertino, and N. Li, "Efficient k-anonymization using clustering techniques," in *Advances in Databases: Concepts, Systems and Applications*, e. a. Kotagiri, R., Ed. Springer Berlin Heidelberg, 2007.

[6] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in *Int. Conf. on Data Engineering*, 2006.

[7] Y.-K. Kim, H. Lee, M. Yoon, and J.-W. Chang, "Hilbert-curve based data aggregation scheme to enforce data privacy and data integrity for wireless sensor networks," *Int. J. of Distr. Sensor Networks*, 2013.

[8] R. Ghrist and A. Muhammad, "Coverage and hole-detection in sensor networks via homology," in *Int. Symp. on Information Processing in Sensor Networks*, 2005.

[9] V. De Silva and R. Ghrist, "Homological sensor networks," *Notices of the American Mathematical Society*, vol. 54, no. 01, pp. 10–17, 2007.

[10] A. Muhammad and M. Egerstedt, "Control using higher order Laplacians in network topologies," in *Int. Symp. of Mathematical Theory of Networks and Systems*, 2006, pp. 1024–1038.

[11] A. Tahbaz-Salehi and A. Jadbabaie, "Distributed coverage verification in sensor networks without location information," *IEEE Trans. on Aut. Control*, vol. 55, no. 8, pp. 1837–1849, 2010.

[12] J. Derenick, A. Speranzon, and R. Ghrist, "Homological sensing for mobile robot localization," in *IEEE Int. Conf. on Robotics and Automation*, 2013, pp. 572–579.

[13] R. Ghrist, D. Lipsky, J. Derenick, and A. Speranzon, "Topological landmark-based navigation and mapping," 2012. [Online]. Available: http://www.math.upenn.edu/~ghrist/preprints/landmarkvisibility.pdf

[14] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

[15] J. Le Ny and G. J. Pappas, "Differentially private filtering," *IEEE Trans. on Aut. Control*, vol. 59, no. 2, pp. 341–354, 2014.

[16] S. Han, U. Topcu, and G. J. Pappas, "Approximately truthful mechanism for electric vehicle charging via joint differential privacy," in *American Control Conference*, 2015.

[17] Y. Mo and R. Murray, "Privacy preserving average consensus," in *IEEE International Conference on Decision and Control*, 2014.

[18] N. Li, L. T., and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," in *IEEE Int. Conf. on Data Engineering*, 2007, pp. 106–115.

[19] R. Ghrist, *Elementary Algebraic Topology*, 1st ed. Createspace.

[20] "Perseus, the persistent homology software," http://www.sas.upenn.edu/~vnanda/perseus, accessed 2015-09-27.

[21] R. Dewri, I. Ray, and D. Whitley, "On the optimal selection of k in the k-anonymity problem," in *IEEE Int. Conf. on Data Engineering*, 2008, pp. 1364–1366.

[22] R. Ghrist, "Barcodes: the persistent topology of data," *Bulletin of the American Mathematical Society*, vol. 45, no. 1, pp. 61–75, 2008.

[23] K. Mischaikow and V. Nanda, "Morse theory for filtrations and efficient computation of persistent homology," *Discrete & Computational Geometry*, vol. 50, no. 2, pp. 330–353, 2013.

[24] G. Carlsson, V. de Silva, and D. Morozov, "Zigzag persistent homology and real-valued functions," in *Annual Symposium on Computational Geometry*. ACM, 2009, pp. 247–256.

[25] A. Speranzon and S. D. Bopardikar, "An algebraic topological perspective to privacy: Numerical and categorical data," 2016. [Online]. Available: http://http://arxiv.org/abs/1602.06643