

# Shapley Homology: Topological Analysis of Sample Influence for Neural Networks

Kaixuan Zhang<sup>1</sup>, Qinglong Wang<sup>2</sup>, Xue Liu<sup>2</sup>, C. Lee Giles<sup>1</sup>

<sup>1</sup>Information Sciences and Technology, Pennsylvania State University.

<sup>2</sup>School of Computer Science, McGill University.

**Keywords:** Topological data analysis, homology, Shapley value, sample influence, deep learning.

## Abstract

Data samples collected for training machine learning models are typically assumed to be independent and identically distributed (iid). Recent research has demonstrated that this assumption can be problematic as it simplifies the manifold of structured data. This has motivated different research areas such as data poisoning, model improvement, and explanation of machine learning models. In this work, we study the influence of a sample on determining the intrinsic topological features of its underlying manifold. We propose the Shapley Homology framework, which provides a quantitative metric for the influence of a sample of the homology of a simplicial complex. More specifically, our proposed framework consists of two main parts: homology analysis where the Betti number of the target topological space, and Shapley value calculation where we decompose the topological features of a complex built from data points to individual points. By interpreting the influence as a probability measure, we further define an entropy

which reflects the complexity of the data manifold. Furthermore, we provide a preliminary discussion of the connection of the Shapley Homology to the Vapnik-Chervonenkis (VC) dimension. Empirical studies show that when the 0-dimensional Shapley homology is used on neighboring graphs, samples with higher influence scores have greater impact on the accuracy of neural networks that determine graph connectivity and on several regular grammars whose higher entropy values imply greater difficulty in being learned.

## 1 Introduction

In much machine learning research, it is common practice to assume that the training samples are independent and identically distributed (iid). As such, samples are implicitly regarded as having equal influence on determining the performance of a machine learning model. Recently, limitations of this assumption have been explored. For example, a recent study (Koh & Liang, 2017) showed that certain training samples can have significant influences over a model’s decisions for certain testing samples. This effect has motivated research on model interpretation (Gunning, 2017; Koh & Liang, 2017; Yeh et al., 2018), model & algorithm improvement (Lee & Yoo, 2019; T. Wang et al., 2018; Ren et al., 2018), and poisoning attacks (Y. Wang & Chaudhuri, 2018; Koh & Liang, 2017; X. Chen et al., 2017).

Many of the aforementioned methods study the sample influence problem by using neural networks. Specifically, it is common to adopt a neural network (either a target model or a surrogate) to identify samples that are deemed as more influential (Koh & Liang, 2017; T. Wang et al., 2018). However, as shown by X. Chen et al. (2017) in a poisoning attack scenario, a limited number of poisoning samples can be effective in applying a backdoor attack in a model-agnostic manner. This motivates work that studies the intrinsic properties of data so that one can develop countermeasures to methods that defeat learning models. Specifically, by representing the underlying space of data as a topological space, we study the sample influence problem from a topological perspec-

tive. The belief that topological features of the data space better represent its intrinsic properties has attracted recent research on establishing relationships between machine learning and topological data analysis (TDA) (Chazal & Michel, 2017; Hofer et al., 2017; Carlsson & Gabrielsson, 2018).

In this work, we consider TDA as a complementary approach to other approaches that study sample influence. We propose a framework for decomposing the topological features of a complex built with data points to individual points. We interpret the decomposed value for each point as representing its influence on affecting its topological features. We then calculate that influence as Shapley values (Narahari, 2012). Recent research (J. Chen et al., 2018; Lundberg & Lee, 2017; Datta et al., 2016) proposed similar approaches to quantify feature influence. By interpreting the influence as a probability measure defined on a data space, we also calculate the entropy which describes the complexity of this space. Under our framework, we devise an algorithm for calculating the influence of data samples and the entropy for any data set.

We perform both analytical and empirical studies on several sets of data samples from two families of structured data – graphs and strings. Specifically, we generate random graphs using the Erdos-Renyi model and binary strings with regular grammars. In both cases, we use neural networks as verification tools for our analysis of sample influence. Explicitly, we employ a feed-forward network trained to determine the connectivity of the generated graphs, and a recurrent network trained to recognize the generated strings. Our results show that samples identified by our method as having more significant influence indeed have more impact on the connectivity of their underlying graphs, and that grammars with higher complexity have more difficulty in being learned.

## 2 Related Work

### 2.1 Sample Influence

Recently, research has focused on how to explore and exploit influential samples. As an example, by analyzing how the performance of a model in the testing phase affects a small number of training samples motivated research in poisoning attacks (Y. Wang & Chaudhuri, 2018; Koh & Liang, 2017; X. Chen et al., 2017). In this case, a limited number of corrupted training examples are injected to degrade a target models performance. An alternative thread of research exploits this effect to improve the generalization accuracy of learning models and the efficiency of learning algorithms. Specifically, influential samples can be identified via learning and then used for enhancing a models' performance on imbalanced and corrupted data (Ren et al., 2018; Lee & Yoo, 2019). They can also be synthesized to represent a much larger set of samples, thus accelerating the learning process (T. Wang et al., 2018). In addition, influential samples can bring explainability to deep learning models by identifying representative samples or prototypes used in decision making (Yeh et al., 2018; Anirudh et al., 2017).

Typical approaches measure the influence of a data point by decomposing the target functions values with respect to all of the data points of their contributions (Koh & Liang, 2017). More precisely, they observe the unit change of the value of the target function caused by this point. In most cases, the target function is the output of a trained model. We follow the same approach but change the decomposition rule and particularly the target function. In our case instead of building a model-dependent framework, we reveal the internal dependencies between data in a metric space.

### 2.2 Topological Data Analysis (TDA)

Currently, the most widely-used tool in TDA is persistent homology (Chazal & Michel, 2017), which is an algebraic method for measuring the topological features of shapes and functions. This method provides a multi-scale analysis of the underlying geomet-

ric structures represented by persistence barcodes or diagrams. Most of the previous research using persistent homologies has focused on the classification task by identifying manifolds constructed by data points (Carlsson et al., 2008; Li et al., 2014; Turner et al., 2014) and mapping the topological signatures to machine learning representations (Carlsson & Gabrielsson, 2018; Bubenik, 2015). Unlike previous work, our proposed Shapley homology emphasizes revealing the influence of a vertex on the topological features of a space from a microscopic viewpoint, and assembling individual influence from which to derive an entropy measure which reflects the macroscopic features of a space consisting of individual vertices.

### 3 Shapley Homology

Here, we introduce the relevant concepts that are necessary for our Shapley Homology framework. We then provide the definitions of sample influence and entropy.

#### 3.1 Čech Complex and Homology Group

**Čech Complex** In topological analysis, the study of a set of data points is typically based on simplicial complexes, which provide more abstract generalizations of neighboring graphs that describe the structural relationships between data points. Here, we are particularly interested in the Čech complex, which is an abstract simplicial complex. More formally, the definition of the Čech complex is as follows.

**Definition 1** (Čech complex). *The Čech complex  $C_r(X)$  is the intersection complex or nerve of the set of balls of radius  $r$  centered at points in  $X$ .*

In particular, given a finite point cloud  $X$  in a metric space and a radius  $r > 0$ , the Čech complex  $C_r(X)$  can be constructed by first taking the points in  $X$  as a vertex set of  $C_r(X)$ . Then for each set  $\sigma \subset X$ , let  $\sigma \in C_r(X)$  if the set of  $r$ -balls centered at points of  $\sigma$  have a nonempty intersection. Note, that the Čech complex has a property

showing that the simplicial complex constructed with a smaller  $r$  is a subcomplex of one constructed with a larger  $r$ . In this way we can obtain a filtration.

**Homology** The topological features of data space are typically formalized and studied through *Homology*, which is a classical concept in algebraic topology (Chazal & Michel, 2017). In the following, we briefly introduce homology to the extent that is necessary for understanding its role in our framework.

**Definition 2** (Homology Group). *A chain complex  $(\mathbf{A}, \mathbf{d})$  is a sequence of abelian groups or modules connected by homomorphisms  $d_n : A_n \rightarrow A_{n-1}$ , such that the composition of any two consecutive maps is the zero map. The  $k$ -th homology group is the group of cycles modulo boundaries in degree  $k$ , that is,*

$$H_k := \frac{\ker d_n}{\text{im } d_{n+1}}, \quad (1)$$

where  $\ker$  and  $\text{im}$  denotes the kernel and image of the homomorphism, respectively.

Generally speaking, the  $k$ -th homology is a quotient group that indicates the  $k$ -dimensional independent features of the space  $\mathcal{X}$ . Particularly, when  $k = 0$ , we have the following proposition (Hatcher, 2005):

**Proposition 1.** *For any space  $\mathcal{X}$ ,  $H_0(\mathcal{X})$  is a direct sum of abelian groups, one for each path-component of  $\mathcal{X}$ .*

Specially, when a complex takes the form of a neighboring graph, then the 0-homology denote the subgraphs that are connected in this graph. As for the 1-homology  $H_1$ , it represents the number of 1-dimensional holes in the given manifold.

### 3.2 Sample Influence and Entropy

Given a point cloud represented by a Čech complex, we study the influence of each data point on the topological features of this complex. This influence of a data point can be further interpreted as the probability that a unit change of the topological feature

is caused by this point. More formally, denote a data set containing  $n$  samples by  $X = \{x_1, \dots, x_n\}$ , we have the following definition of sample influence.

**Definition 3** (Sample Influence). *Given a discrete space  $(X, \Sigma)$ , the influence of any sample (or subset)  $\{x\} \subset X$  is a probability measure  $\mu$ .*

We then define an entropy that shares a similarity to the entropy in information theory; both reflect the number of internal configurations or arrangements of a system.

**Definition 4** (Entropy). *Given a probability measure  $\mu$ , the entropy of a dataset  $X$  is defined as  $H(X) = -\sum_{i=1}^n \mu(x_i) \log \mu(x_i)$ .*

When introducing new data points, these data points may change the number of boundaries needed to separate the data points in the given metric space. Moreover, each of the newly added data points will have a different influence on the resulting number. By interpreting the influence as a probability measure, our defined entropy quantifies the number of boundaries in a data set, and thus implies the capacity that a neural network needs to learn this data set.

### 3.3 The Framework of Shapley Homology

Here we propose a framework (depicted in Figure 1) of Shapley Homology in order to study sample influence. Specifically, in Figure 1, we provide an example for investigating a specific topological feature – the Betti number (Rote & Vegter, 2006) – of a topological space. As will become clear from the following, the Betti number can be used to quantify the homology group of topological space according to its connectivity.

**Definition 5** (Betti Number). *Given a non-negative integer  $k$ , the  $k$ -th Betti number  $\beta_k(\mathcal{X})$  of the space  $\mathcal{X}$  is defined as the rank of the abelian group  $H_k(\mathcal{X})$ , that is,*

$$\beta_k(\mathcal{X}) = \text{rank}(H_k(\mathcal{X})). \quad (2)$$

Following the definition, as there are no special structures (such as real projective space in the complex), the Betti number  $\beta_k(\mathcal{X})$  indicates the number of direct sum of

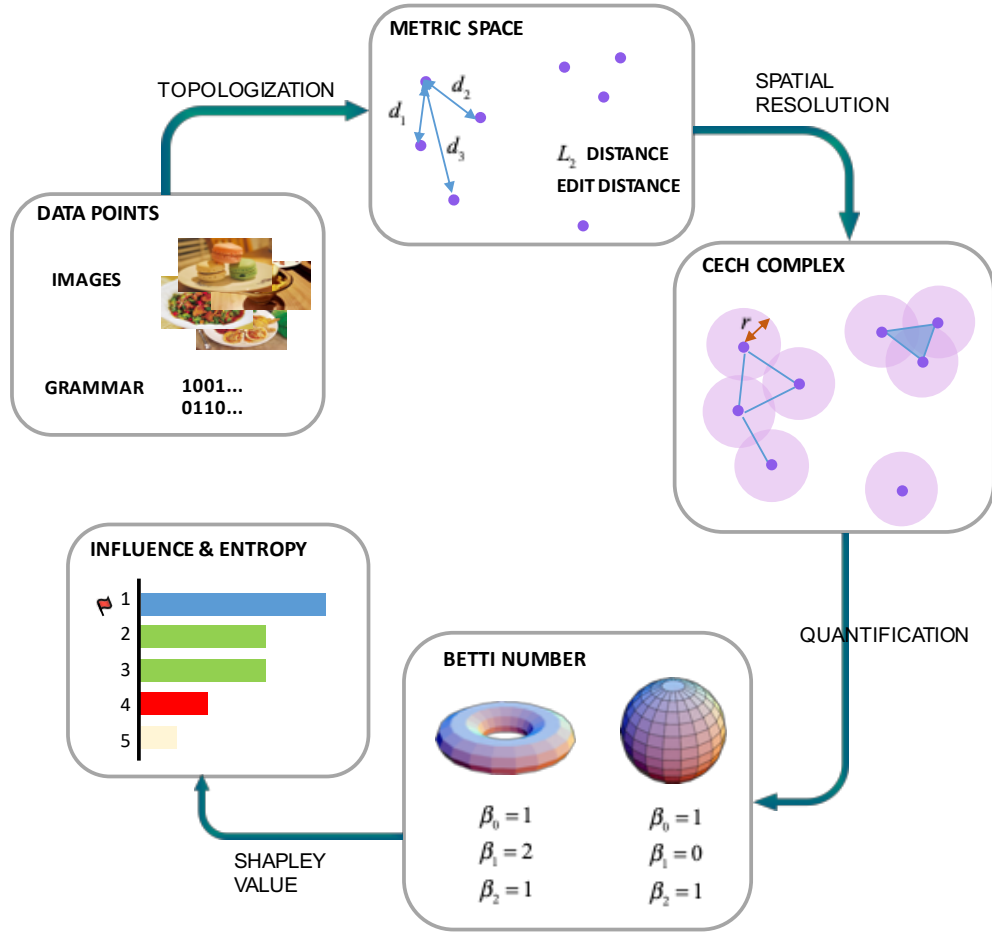


Figure 1: Framework of Shapley Homology for influence analysis

abelian groups of the  $k$ -th homology group. In other words, the  $k$ -th Betti number refers to the number of  $k$ -dimensional voids on a topological surface.

While the Betti number only indicates an overall feature of a complex, we need to further distribute this number to each vertex of this complex as its influence score. Recent research (J. Chen et al., 2018) on interpreting neural networks has introduced Shapley value from cooperative game theory to decompose a classification score of a neural network made for a specific sample to individual features as their importance or contribution to rendering this classification result. Inspired by this line of work, we also employ the Shapley value as the influence score for each vertex. However, it should be noted that for a fixed  $k$ , the Betti number  $\beta_k$  does not satisfy the monotonicity property of Shapley values. That is, the Betti number of  $X_1$  is not necessarily larger than that of



$X_2$  when  $X_1$  is a subcomplex of  $X_2$ . As a result, we cannot adopt the formulation for calculating Shapley values directly. Here we use the following variant formulation for calculating Shapley values for a Čech complex:

$$s(x_i) = \sum_{C \subseteq X \setminus x_i} \frac{|C|!(n - |C| - 1)!}{n!} |\beta(C \cup x_i) - \beta(C)|. \quad (3)$$

It is important to note that in our formulation (3), we use the absolute value to resolve the monotonicity issue.

It is clear that our formulation (3) satisfies the symmetry axiom, whereas, it does not satisfy other Shapley axioms, including the linearity and carrier axioms (Narahari, 2012). Nonetheless, our formulation still measures the marginal effect of a vertex. Besides, as we mainly focus on measuring the topological feature, both the decrease and increase in Betti number of a vertex are crucial for determining the influence. Furthermore, since our entropy is symmetry-invariant, its value will remain the same under the group action on the vertices (Conrad, 2008).

The above discussion indicates that the influence of a data sample can be regarded as a function of the radius  $r$ , the Betti number  $\beta_k$ , the size of the data set containing this sample, and the topological space constructed upon the chosen metric. Unlike persistent homology (Edelsbrunner & Harer, 2008), which studies the topological features of data space at varying “resolution”  $r$ , our analysis takes a more static view of the topological features of a complex built with a fixed  $r$ . As such, our analysis can be viewed as taking a slice from the filtration of used for persistent homology. In the following section, we propose an algorithm for calculating the influence scores of data points in a Čech complex constructed with certain specified  $r$  and  $k$ .

## 4 Algorithm Design and Case Study

With Proposition 1 and Definition 5 and when the Betti number  $\beta_0$  is regarded as a quantitative indicator, it equals the number of connected components in a complex. In this case, the task of calculating  $\beta_0$  of a Čech complex is equivalent to calculating the

number of connected components of a graph. This enables us to compute the Laplacian matrix  $L$ <sup>1</sup> of a graph then apply the following proposition (Marsden, 2013).

**Proposition 2.** *A graph  $G$  has  $m$  connected components if and only if the algebraic multiplicity of 0 in the Laplacian is  $m$ .*

With the above proposition, we can see that the Betti number  $\beta_0(X)$  is equal to the number of zeros in the spectrum of the corresponding Laplacian matrix  $L_X$ . As such, we propose the Algorithm 1 to calculate the influence score and the entropy under the setting of  $k = 0$  and  $r$  as some constant. It is evident that our proposed Algorithm 1 always produces a probability measure. A more detailed discussion of the choice of  $k$  and  $r$  is provided in Section 5.

Note that this algorithm is specifically designed for the zero dimensional homology group, which is a special case in our Shapley Homology framework. In this case, the metric computed in Algorithm 1 can be interpreted as the betweenness centrality (Newman, 2005) of a graph. However, it is important to note that betweenness centrality does not have a natural correspondence to any topological features. On the contrary, the metric derived for the zero dimensional homology group precisely matches with the Betti number. Furthermore, while it provides conceptual convenience in relating our designed metric with betweenness centrality, this connection should not dim the generality of our framework to other homology groups with higher dimensions.

Here we provide a simple time complexity analysis for our algorithm and propose several possible solutions to accelerate the algorithm for future work. Our algorithm can be decomposed into three parts: (1) complex construction; (2) calculation of the graph spectrum; (3) assignment of the Shapley value. For the first step, we need to calculate pairwise distance. From the adjacent matrix, it is clear that the complexity is  $O(n^2)$ . Second, we need to compute the spectrum of all subcomplexes. As such, in total, the complexity is  $O(n^3 2^n)$ , where the complexity of each Laplacian decomposition

---

<sup>1</sup>The Laplacian matrix  $L$  is defined to be  $D - A$  where  $D$  and  $A$  denote the degree matrix and adjacency matrix respectively.

---

**Algorithm 1**

---

**Input:**

Data  $X = \{x_1, \dots, x_n\}$ ; Metric  $d(\cdot, \cdot)$ ; Resolution  $r$ ;

**Output:**

Influence  $\mu(x_1), \dots, \mu(x_n)$ ; Entropy  $H(D)$ ;

- 1: **for** each pair of data sample  $x_i, x_j$  **do**
  - 2:   Calculate the pairwise distance  $d(x_i, x_j)$ ;
  - 3: **end for**
  - 4: Build the Complex  $\Delta$  according to the resolution  $r$ ;
  - 5: **for** each subcomplex  $C \subseteq \Delta$  **do**
  - 6:   Calculate the Laplacian matrix  $L$  of  $C$ ;
  - 7:   Calculate the multiplicity of eigenvalue 0 of  $L$ ;
  - 8: **end for**
  - 9: **for**  $i = 1$  to  $n$  **do**
  - 10:   Calculate  $s(x_i)$  according to (3);
  - 11: **end for**
  - 12: Set  $\mu(x_i)$  as the normalized  $s(x_i)$ , then  $H(X) = -\sum_{i=1}^n \mu(x_i) \log \mu(x_i)$ ;
- 

is  $O(n^3)$ . For the third step, we sum all the marginal utility for one sample, which results in a complexity of  $O(2^n)$ . Therefore, the complexity of computing the influence scores for all samples is  $O(n2^n)$ . Based on the above analysis, we obtain the overall complexity of our algorithm as  $O(n^32^n)$ . Clearly, the second and third steps contribute most to the complexity. In order to alleviate the computational burden caused by the second step, future work will consider various approaches (e.g., (Cohen-Steiner et al., 2018)) that approximates the spectrum of a graph. For the third step, several existing approximation algorithms, e.g., C-Shapley and L-Shapley (J. Chen et al., 2018), can be considered for approximating the Shapley value using local topological properties.

In the following, we provide a set of case studies on several different graphs with representative topological structures. In particular, we study four types of graphs repre-

$g$	Description	Entropy
1	$1^*$	0.00
2	even number of 0s and even number of 1s.	$\log M$
3	$1^* + 0^*(1 + 0)$	$3 \log 2/2$
	an odd number of consecutive 1s is	$\approx 2.30$
4	always followed by an even number of consecutive 0s.	$(N = 4)$

Table 1: Example grammars and their associated entropy values.

senting the space of binary strings generated by four regular grammars. We select these grammars due to their simplicity for demonstration. A brief introduction of the selected grammars <sup>2</sup> is provided in Table 1. Since we deal with binary string, we specify the distance metric used in these studies as the edit distance (De la Higuera, 2010) and set the radius  $r = 1$ . Also, we set the length  $N$  of generated strings to fixed values as specified in Table 1.

Furthermore, we generalize our proposed framework to another six special sets of graphs and provide the analytical results of their vertices' influence and the entropy values. These graphs are selected as they represent a set of simple complexes that can be used for building up more complicated topological structures.

## 4.1 A Case Study on Regular Grammars

**Simple Examples** Here we calculate and analyze the first three grammars shown in Table 1 as they have simple yet different topological features. As we can see from Table 1, for the first grammar, given a fixed  $N$ , there exists only one string defined by this grammar. As a result, the influence of this sample is 1, and the entropy  $H(g_1) = 0$ . For the second grammar, it is clear to see that any pair of valid strings have an edit

<sup>2</sup>Grammar 1, 2 and 4 are selected from the set of Tomita grammars (Tomita, 1982), of which their indices in the Tomita grammars are 1, 5 and 3, respectively.

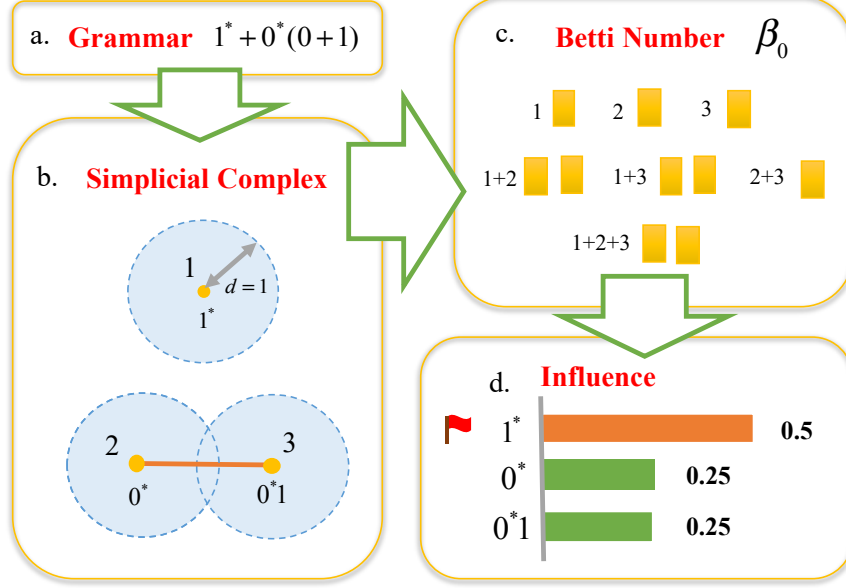


Figure 2: An simple illustration of the Shapley Homology framework on grammar  $1^* + 0^*(0 + 1)$ .

distance larger than 1. In this case, the complex formed with the radius  $r = 1$  consists of disjoint points. Assuming that there exist  $M$  valid strings defined by this grammar with the length of  $N$ , then all strings have the same influence  $1/M$  and the entropy  $H(g_2) = \log M$ . For the third grammar, when the length  $N$  of its associated strings is larger than 3, then the set  $g_3$  of these strings can be expressed as  $g_3 = \{1^N, 0^N, 0^{N-1}1\}$ , denoted as  $g_3 = \{1, 2, 3\}$  for notation simplicity. We depict the complex for the case when  $r = 1$  in Figure 2. According to Proposition 2, we then have the following Betti number  $\beta_0$  of each subcomplex.

$$\beta_0(G_1) = \beta_0(G_2) = \beta_0(G_3) = \beta_0(G_{2,3}) = 1,$$

$$\beta_0(G_{1,2}) = \beta_0(G_{1,3}) = \beta_0(G_{1,2,3}) = 2.$$

where  $G_S$  ( $S \subseteq \{1, 2, 3\}$ ) denotes the subgraph of  $G_{1,2,3}$  formed by the vertex set  $S$ . According to (3) and Algorithm 1, we have  $\mu(1) = 0.5, \mu(2) = \mu(3) = 0.25$ , and finally the entropy is  $H(g_3) = \frac{3}{2} \log 2$ .

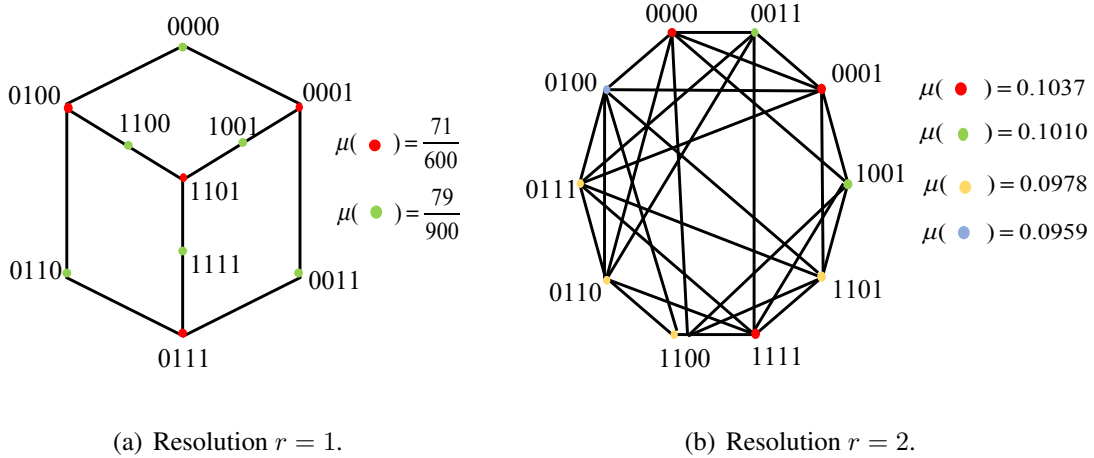


Figure 3: The Čech complex, influence scores and entropy of the fourth grammar with  $N = 4$ .

**A Complicated Example** The fourth grammar  $g_4$  shown in Table 1 is more complicated than the three grammars mentioned above. In particular, let  $N = 4$ ,  $r = 1$  or  $2$ , then we illustrate the results in Figure 3, and the entropy is  $H(g_4) = 2.292$  when  $r = 1$  and  $H(g_4) = 2.302$  when  $r = 2$ . Besides the analytical results presented here, we further demonstrate the difference between this grammar and the first two grammars in a set of empirical studies in Section 6.2.

## 4.2 A Case Study on Special Graphs

In this part, we apply our framework to six special families of graphs, which are shown with examples in Table 2. Due to space constraint, here we omit the detailed calculation of the influence scores and entropy values for these graphs. With the analytical results shown in Table 2, it is easy to derive the following two corollaries:

**Corollary 1.**  $H(K_n) = H(C_n) > H(W_n) > H(S_{n-1})$  for  $n > 5$ , where  $K_n$  denotes the complete graph,  $C_n$  denotes the cycle graph,  $W_n$  denotes wheel graph and  $S_n$  denotes the path graph.

**Corollary 2.** Suppose that  $m + n$  is constant, then the value of  $H(K_{m,n})$  decreases as  $|m - n|$  increases, where  $K_{m,n}$  denotes the complete bipartite graph.







	Example	Shapley Value	Influence Score
Complete Graph $K_n$		$1/n$	$1/n$
Cycles $C_n$		$2/3 - 1/n$	$1/n$
Wheel Graph $W_n$		Periphery: $1/3 - 1/n(n-1)$ Center: $(n^2 - 7n + 18)/6n$	Periphery: $\frac{2(n^2-n-3)}{3(n-1)(n^2-3n+4)}$ Center: $\frac{(n^2-7n+18)}{3(n^2-3n+4)}$
Star $S_{n-1}$		Periphery: $1/2$ Center: $(n^2 - 3n + 4)/2n$	Periphery: $n/2(n^2 - 2n + 2)$ Center: $\frac{n^2-3n+4}{2(n^2-2n+2)}$
Path Graph $P_n$		Ends: $1/2$ Middle: $2/3$	Ends: $3/2(2n - 1)$ Middle: $2/(2n - 1)$
Complete Bipartite Graph $K_{m,n}$		$m$ side: $\frac{n(n-1)}{m(m+1)(m+n)} + \frac{1}{n+1}$ $n$ side: $\frac{m(m-1)}{n(n+1)(m+n)} + \frac{1}{m+1}$	$m$ side: $\frac{m^3+n^3+m^2n+mn+m^2-n}{m(m+n)(2m^2+2n^2+m+n-mn-1)}$ $n$ side: $\frac{m^3+n^3+n^2m+mn+n^2-m}{n(m+n)(2m^2+2n^2+m+n-mn-1)}$

Table 2: Six special sets of graphs and their values of influence for the vertices.

## 5 Extended Discussion of Shapley Homology

As previously introduced in Section 3.3, our proposed influence score can be viewed as a function of the parameter  $k$  and  $r$ . Here we extend our study from the previous case of  $k = 0$  to cover other cases when  $k > 0$ , and show that adopting the 0-dimensional homology offers the merits of having more meaningful and practical implication. Also, we discuss the effect of selecting different radii  $r$  on the topological features we study. More importantly, we show that our framework provides a more general view of sample influence and can be further adapted to capture the dynamical features of a topological space.

### 5.1 Discussion of the Choice of $k$

We first consider the case when  $k = 1$ . In this case, the Betti number  $\beta_1$  represents the number of 1-dimensional holes in a given manifold. Take the third grammar introduced in Section 4.1 as an example. It is clear to see that when using the 1-dimensional homology, all subcomplex has  $\beta_1$  equal to 0. This is precisely the case when our algorithm cannot generate a probability measure. In a more general case, it is easy to show by the homology theory (Schapira, 2001) that once all subcomplex gets Betti

number  $\beta_{\hat{K}} = 0$  for some  $\hat{K}$ , then for any other  $k$ -dimensional homology with  $k > \hat{K}$ , our framework cannot produce a probability measure. When adopting  $k$ -dimensional homology with  $k$  larger than 2, our framework will highlight the effect of having the voids constructed or destroyed in the manifold by introducing a data point. However, the voids in a high-dimensional manifold can be very abstract and cause difficulty for intuitive understanding.

Another difficulty of adopting homology with higher dimensions is due to practical concerns. More specifically, since it is challenging to calculate the homology when  $k$  is large, one may need to apply tools such as the Euler characteristic or Mayer-Vietoris Sequences (Hatcher, 2005), which is out of the scope of this paper.

## 5.2 Discussion of the Choice of $r$

The radius  $r$  plays a critical role in building a manifold. For example, when  $r$  is sufficiently small, the resulting complex contains only discrete data points with equal influence in our framework. When  $r$  is sufficiently large, on the other hand, the complex is contractible, which indicates that there is no difference between the influence of data points contained in this complex. Both aforementioned extreme cases produce equal importance for the data points in a data set, and correspond to the i.i.d. assumption of sample distribution that is commonly assumed in current machine learning practice. In this sense, our framework provides a general abstraction of sample influence.

In practice, selecting a proper  $r$  can be very difficult. This explains the usage of persistent homology (Edelsbrunner & Harer, 2008), which studies the evolution of topological features of a space as  $r$  dynamically changes. It is not difficult to extend our framework to dynamically generate a series of probability measures for the influence and the corresponding entropy values for a data set. However, a sequence of probability measures with various radii has an unclear physical meaning. In this way, the radius should depend on the specific application and should be pre-defined as the resolution of the problem of interest.



### 5.3 Remarks on Related Definitions of Entropy

Different notions of entropy have been previously proposed. Here we briefly revisit several representative definitions and compare with our entropy definition, which quantifies the number of boundaries in a data set. As such, it determines the capacity of a neural network that is needed to learn this data set (Khoury & Hadfield-Menell, 2018).

**Graph Entropy** An important property of the graph entropy (Rezaei, 2013) is monotonicity since it is defined based on mutual information. Specifically, it describes that the entropy of a subgraph is smaller than that of the whole graph on the same vertex set. In our case, by considering the complex as a graph, our entropy is defined to capture the geometric properties (such as the symmetry invariant property mentioned in Section 3.3) of a graph. More specifically, our entropy measure focuses more on the variation of the topological features when a graph is changed. As such, our definition also covers variations that may violate the monotonicity property and the subadditivity property.

**Entropy for Grammatical Learning** This entropy (Q. Wang, Zhang, Ororbia, et al., 2018) is defined to reflect the balance between the population of strings accepted and rejected by a particular grammar. It shows that the entropy of a certain grammar is equal to that of the complement of that grammar. This raises a contradiction in the intuition that a grammar with a higher cardinality of strings is more likely to have a higher entropy value. Our entropy, on the other hand, is defined to capture the intrinsic properties of a set of samples instead of reflecting the difference between different sets of samples. In this sense, our entropy is more likely to assign a higher entropy value to a set of samples with larger cardinality.

**Entropy in Symbolic Dynamics** This type of entropy (Williams, 2004) is defined to reflect the cardinality of a shift space, which can be regarded as a more general definition of regular grammar. It implicitly assumes that any shift contained in a shift space

has equal influence. This is contrary to our case in that we define the entropy to describe the complexity of a topological space that contains vertices with different influence. As such, our entropy provides a more fine-grained description of a topological space.

## 5.4 Connection to the VC Dimension

Here we provide a preliminary discussion on the connection between our proposed Shapley Homology and Vapnik-Chervonenkis (VC) dimension (Vapnik, 2000). Recall this essentially reflects the complexity of a space of functions by measuring the cardinality of the largest set of samples that can be “shattered” by functions in this space. From a similar point of view, we expect the topology of data space is also critical in evaluating the complexity of real-world learning algorithms in learning a given set of data.

Note that the proposed approach has a close connection to statistical learning theory. An obvious interpretation of our complexity is that it can be viewed as analogous to sample complexity (Hanneke, 2016), which is closely related to the VC dimension. However, here we point out the limitation of complexity specified by the VC dimension, which is part of the motivation of this work. Specifically, given a certain data space, only the hypothesis space of models with sufficiently large VC dimension can shatter this data space. For this hypothesis space, we can further use sample complexity to specify the number of samples required to achieve certain PAC learning criteria. However, we argue that different shattering of the data space leads to different levels of complexity (or different entropy values in our terms). Instead of focusing only on the maximally shattered data space, we argue that in practice, when building a learning model, a different shattering should be treated differently. To better explain this effect, we take regular grammars as an example case. One can consider a certain binary regular grammar as a certain configuration in the hypothesis space. In other words, a binary regular grammar explicitly split all the  $2^N$  strings into the set of accept strings and the set of rejected strings, given that the strings have a fixed length of  $N$ . Since this

grammar is equivalent to a DFA, and if we regard this DFA as a classification model, it itself has a certain VC dimension (Ishigami & Tani, 1997). Indeed, this effect is shown in the experiments in Section 6 on grammar learning. In particular in our experiments, the demonstrated different levels of difficulty of learning different regular grammars indicate that different grammars (or different ways of shattering the data space) should not be taken as equal.

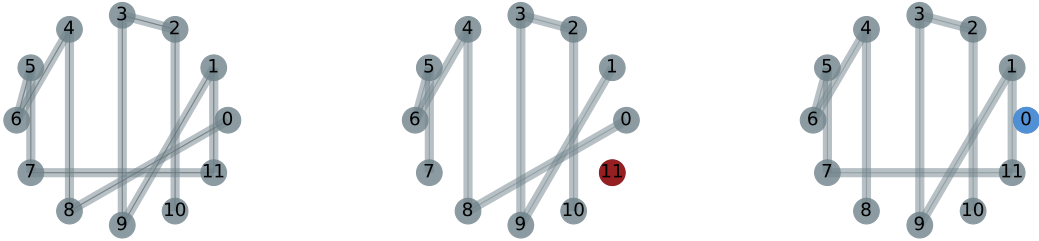
## 6 Experiments

In this section, we first demonstrate the results of using our Algorithm 1 to identify influential nodes in random graphs. Then we evaluate this on several data sets generated by regular grammars to determine if data sets assigned by our algorithm with higher entropy values cause more challenges for neural networks to be able to learn these grammars. The settings of all parameters in the experiments are provided in the supplementary file.

### 6.1 Graph Classification

For these experiments, we first constructed the learning models and data sets of random graphs following Dai et al. (2018). We adopted this similar setting for the purpose of evaluating and verifying the influence of individual nodes in a graph. To avoid the evaluation results from being biased by a particular synthetic data set, we performed 20 experimental trials by varying the probability used in the Erdos-Renyi random graph model to generate the same number of random graphs. More specifically, in each trial, we generated a set of 6,000 undirected random graphs from three classes (2000 for each class), and split the generated set into a training and a testing set with the ratio of 9 to 1.

We constructed the learning models based on `structure2vec` (Dai et al., 2016) for the task of determining the number of connected components (up to 3) of a given graph. The constructed models function as verification tools for the fidelity test. The idea is to first mask nodes in a graph assigned by our algorithm with different influence scores,



(a) The original graph                      (b) Graph manipulated by  
masking the most influential  
node.    (c) Graph manipulated by  
masking the least influential  
node.

Figure 4: Example graphs with classification results for each graph from left to right : 1, 2 and 1. This indicates the number of connected components recognized by a neural network on these graphs.

and examine how much the classification results for this graph is affected. Similar fidelity tests have been widely adopted in previous research on studying feature influence (Ribeiro et al., 2016; Lundberg & Lee, 2017; J. Chen et al., 2018). In each trial, we first trained a neural network on the training set with random initialization. For each graph in the testing set, we computed the influence score for each node in this graph. Then we generated two data sets,  $D_{top}$  and  $D_{bottom}$  with graphs masked out the top- $J$  nodes and bottom- $J$  nodes (both the top- $J$  nodes and bottom- $J$  nodes are identified by their influence scores). We show in Figure 4 the examples of a clean graph and its top- $I$  and bottom- $I$  masked versions. We also constructed a third data set  $D_{rand}$  by randomly masking  $L$  nodes for all testing graphs. The evaluation results were obtained by comparing the classification performance of our trained models on these three data sets, and that achieved on the original testing sets.

Results in Figure 5 show all trials that fit the shaded area for each line plot. The accuracy value indexed by  $L = 0$  is the averaged accuracy of our models obtained on the clean testing sets from all trials. It is clear from Figure 5 that the influence score calculated by our algorithm effectively indicates the impact of a node on determining the connectivity of the graph containing this node. In addition, as we increase  $L$ , the

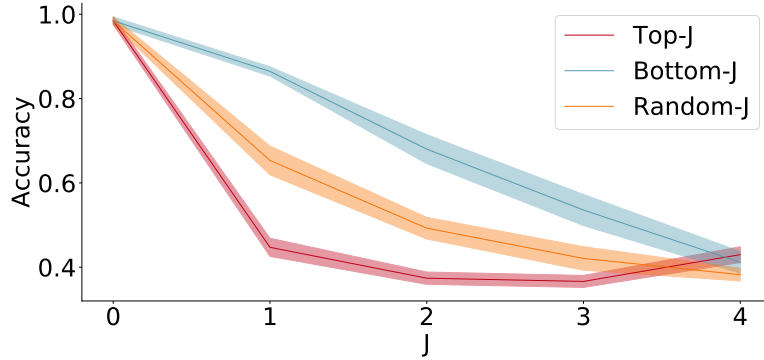


Figure 5: Accuracy of neural networks obtained on data sets with varying scales of manipulations.

accuracy of our models obtained on  $D_{\text{top}}$ ,  $D_{\text{bottom}}$  and  $D_{\text{rand}}$  degrades with different scales. In particular, the accuracy obtained on  $D_{\text{top}}$  and  $D_{\text{bottom}}$  shows the largest and smallest scales of degradation, respectively. The result for  $D_{\text{top}}$  is surprising in that even on these simple synthetic graphs, the robustness of a neural network model is far from satisfactory. Specifically, similar to the results shown by Dai et al. (2018), by masking top- $l$  influential nodes in the testing graphs, the accuracy of a neural network is brought down to 40%  $\sim$  50%.

In the case when we only focus on the vertex with the highest importance score, the experiment can be taken as a sanity check. However, as shown in Figure 5, when we evaluate the influence from manipulating vertices that have been identified to have different levels of influence on the performance of classifying graphs, our framework also provides a quantitative evaluation of the importance of different vertices, which goes beyond a sanity check.

## 6.2 Grammar Recognition

In this set of experiments, we used three Tomita grammars introduced in Table 1 due to their simplicity and wide adoption in various grammar learning research (Weiss et al., 2017; Q. Wang, Zhang, Ororbia II, et al., 2018; Q. Wang, Zhang, Liu, & Giles, 2018).

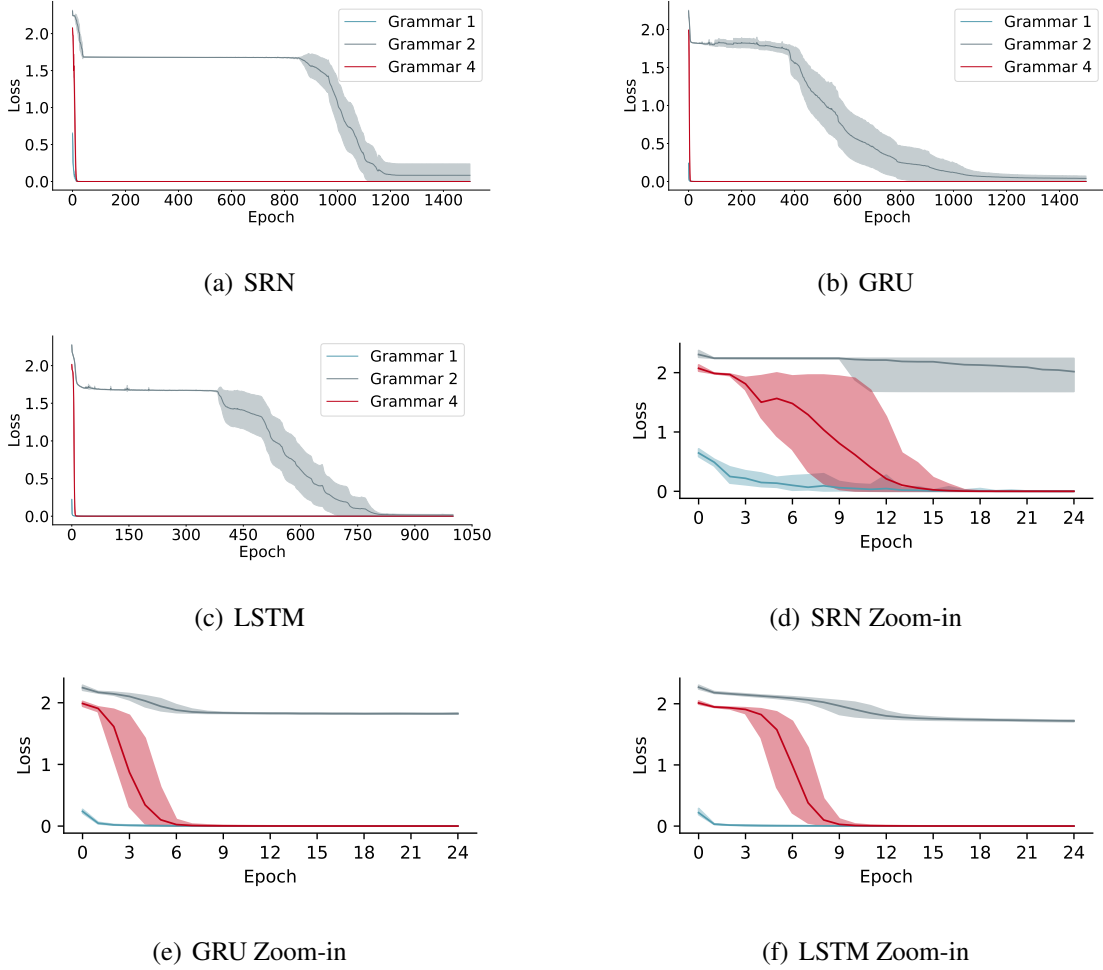


Figure 6: Training Performance for Different Recurrent Networks on Grammars 1, 2 and 4

In Table 3, we show for each grammar its entropy value with different radii calculated by our algorithm. We set the radius as discrete integers since we apply the edit distance to regular grammars. Note that it can be computationally intractable if we apply the Shapley homology framework to a large scale data set, since our algorithm under our framework had exponential time complexity. Therefore, we apply an approximation, i.e. we compute the spectrum of  $K$  subcomplexes instead of all subcomplexes, where  $K$  is a fixed number. More specifically, when computing the specific sample influence, we randomly selected  $K$  subcomplexes that contain this sample and use the same formula to obtain its Shapley value.

We used these grammars to generate three sets of binary strings with length ranging

		Grammar 1	Grammar 2	Grammar 4
Entropy ( $K = 100$ )	$r = 1$	2.38	5.67	5.07
	$r = 2$	2.47	5.65	5.08
	$r = 3$	2.46	5.65	5.07

Table 3: Entropy of different grammars with various radii.

from 2 to 13 and split the data set of each grammar into a training set and testing set with the ratio of 7 to 3. We then trained several different recurrent networks (Simple Recurrent Network (SRN) (Elman, 1990), gated-recurrent-unit networks (GRU) (Cho et al., 2014) and long-short-term-memory networks (LSTM) (Hochreiter & Schmidhuber, 1997)) for a binary classification task on the data set generated by each grammar. Essentially, the binary classification task is designed to train a recurrent network to correctly differentiate and accept strings that are accepted by a certain regular grammar, and reject strings that are rejected by this grammar. For each type of recurrent network, we use the same number of parameters for models used for all three grammars in order to avoid any bias that these models would have any different learning ability. We trained the model of each type of RNN on each grammar for 20 trials. In each trial, we randomly split the training and testing set and randomly initialize the model. The results are demonstrated in Figure 6, in which the results from all trials fit the shaded area associated with each plot.

In Figure 6(a), 6(b) and 6(c), we see that for the first and fourth grammars which have lower entropy values, the learning process converges much faster and more consistently than the process for the second grammar, which has the highest entropy value. This effect holds for all types of recurrent networks we evaluate. To better illustrate the difference of the difficulty of learning the first and fourth grammars, we provide in Figure 6(d), 6(e) and 6(f) the zoomed view for each plot at the top row of Figure 6. While the learning process of all models converges within ten epochs for both grammars, it is still clear that the learning process is slower for the fourth grammar. These results agree with both our analysis on the entropy of these grammars and the intuition. Specifically,

the second grammar defines two sets of strings with equal cardinality when the string length is even. In this case, by flipping any binary digit of a string to its opposite (e.g., flipping a 0 to 1 or vice versa), a valid or invalid string can be converted into a string with the opposite label. This implies that a model must pay equal attention to any string to learn the underlying grammar. This corresponds to our analysis result that in the data space defined by the second grammar, each string sample shares equal influence on affecting the topological feature of this space.

## **7 Conclusion**

We proposed a Shapley Homology framework to study the sample influence on topological features of a data space and its associated entropy. This provides an understanding of the intrinsic properties of both individual samples and the entire data set. We also designed an algorithm for decomposing the 0th-Betti number using a cooperative game theory and provide analytical results for several special representative topological spaces. Furthermore, we empirically verified our results with two carefully designed experiments. We show that data points identified by our algorithm that have a larger influence on the topological features of their underlying space also have more impact on the accuracy of neural networks for determining the connectivity of their underlying graphs. We also show that a regular grammar with higher entropy has more difficulty of being learned by neural networks.

## **8 Acknowledgments**

We gratefully acknowledge useful comments from the referees and Christopher Griffin.

## **References**

Anirudh, R., Thiagarajan, J. J., Sridhar, R., & Bremer, T. (2017). Influential sample



- selection: A graph signal processing approach. *arXiv preprint arXiv:1711.05407*.
- Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16(1), 77–102.
- Carlsson, G., & Gabrielsson, R. B. (2018). Topological approaches to deep learning. *CoRR*, *abs/1811.01122*.
- Carlsson, G., Ishkhanov, T., De Silva, V., & Zomorodian, A. (2008). On the local behavior of spaces of natural images. *International journal of computer vision*, 76(1), 1–12.
- Chazal, F., & Michel, B. (2017). An introduction to topological data analysis: fundamental and practical aspects for data scientists. *CoRR*, *abs/1710.04019*.
- Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). L-shapley and c-shapley: Efficient model interpretation for structured data. *CoRR*, *abs/1808.02610*.
- Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. *CoRR*, *abs/1712.05526*.
- Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of ssst@emnlp 2014, eighth workshop on syntax, semantics and structure in statistical translation, doha, qatar, 25 october 2014* (pp. 103–111).
- Cohen-Steiner, D., Kong, W., Sohler, C., & Valiant, G. (2018). Approximating the spectrum of a graph. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1263–1271).
- Conrad, K. (2008). *Group actions*.
- Dai, H., Dai, B., & Song, L. (2016). Discriminative embeddings of latent variable models for structured data. *arXiv preprint arXiv:1603.05629*.

- Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J., & Song, L. (2018). Adversarial attack on graph structured data. , 1123–1132.
- Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE symposium on security and privacy, SP 2016, san jose, ca, usa, may 22-26, 2016* (pp. 598–617).
- De la Higuera, C. (2010). *Grammatical inference: learning automata and grammars*. Cambridge University Press.
- Edelsbrunner, H., & Harer, J. (2008). Persistent homology-a survey. *Contemporary mathematics*, 453, 257–282.
- Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2), 179–211.
- Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*.
- Hanneke, S. (2016). The optimal sample complexity of pac learning. *The Journal of Machine Learning Research*, 17(1), 1319–1333.
- Hatcher, A. (2005). *Algebraic topology*. Cambridge University Press.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.
- Hofer, C., Kwitt, R., Niethammer, M., & Uhl, A. (2017). Deep learning with topological signatures. In *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, 4-9 december 2017, long beach, ca, USA* (pp. 1633–1643).
- Ishigami, Y., & Tani, S. (1997). Vc-dimensions of finite automata and commutative finite automata with k letters and n states. *Discrete Applied Mathematics*, 74(2), 123–134.

- Khoury, M., & Hadfield-Menell, D. (2018). On the geometry of adversarial examples. *CoRR*, *abs/1811.00525*.
- Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In *Proceedings of the 34th international conference on machine learning, ICML 2017, sydney, nsw, australia, 6-11 august 2017* (pp. 1885–1894).
- Lee, D., & Yoo, C. D. (2019). *Learning to augment influential data*. Retrieved from <https://openreview.net/forum?id=BygIV2CcKm>
- Li, C., Ovsjanikov, M., & Chazal, F. (2014). Persistence-based structural recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1995–2002).
- Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, 4-9 december 2017, long beach, ca, USA* (pp. 4768–4777).
- Marsden, A. (2013). Eigenvalues of the laplacian and their relationship to the connect- edness of a graph. *University of Chicago, REU*.
- Narahari, Y. (2012). . the shapley value.
- Newman, M. E. (2005). A measure of betweenness centrality based on random walks. *Social networks*, *27*(1), 39–54.
- Ren, M., Zeng, W., Yang, B., & Urtasun, R. (2018). Learning to reweight examples for robust deep learning. In *Proceedings of the 35th international conference on machine learning, ICML 2018, stockholmmsässan, stockholm, sweden, july 10-15, 2018* (pp. 4331–4340).
- Rezaei, S. S. C. (2013). Entropy and graphs. *arXiv preprint arXiv:1311.5632*.

- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, san francisco, ca, usa, august 13-17, 2016* (pp. 1135–1144).
- Rote, G., & Vegter, G. (2006). Computational topology: An introduction. In *Effective computational geometry for curves and surfaces* (pp. 277–312). Springer.
- Schapira, P. (2001). *Categories and homological algebra*. Société mathématique de France.
- Tomita, M. (1982). Dynamic construction of finite-state automata from examples using hill-climbing. In *Proceedings of the fourth annual conference of the cognitive science society* (pp. 105–108).
- Turner, K., Mukherjee, S., & Boyer, D. M. (2014). Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3(4), 310–344.
- Vapnik, V. N. (2000). *The nature of statistical learning theory, second edition*. Springer.
- Wang, Q., Zhang, K., Liu, X., & Giles, C. L. (2018). Verification of recurrent neural networks through rule extraction. *arXiv preprint arXiv:1811.06029*.
- Wang, Q., Zhang, K., Ororbia, I., Alexander, G., Xing, X., Liu, X., & Giles, C. L. (2018). A comparative study of rule extraction for recurrent neural networks. *arXiv preprint arXiv:1801.05420*.
- Wang, Q., Zhang, K., Ororbia II, A. G., Xing, X., Liu, X., & Giles, C. L. (2018). An empirical evaluation of rule extraction from recurrent neural networks. *Neural computation*, 30(9), 2568–2591.
- Wang, T., Zhu, J., Torralba, A., & Efros, A. A. (2018). Dataset distillation. *CoRR*, abs/1811.10959.

- Wang, Y., & Chaudhuri, K. (2018). Data poisoning attacks against online learning. *CoRR*, *abs/1808.08994*.
- Weiss, G., Goldberg, Y., & Yahav, E. (2017). Extracting automata from recurrent neural networks using queries and counterexamples. *arXiv preprint arXiv:1711.09576*.
- Williams, S. G. (2004). Introduction to symbolic dynamics. In *Proceedings of symposia in applied mathematics* (Vol. 60, pp. 1–12).
- Yeh, C., Kim, J. S., Yen, I. E., & Ravikumar, P. (2018). Representer point selection for explaining deep neural networks. In *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, neurips 2018, 3-8 december 2018, montréal, canada*. (pp. 9311–9321).

## A Proof of the Results for the Six Special Sets of Graphs

Here we provide proofs for results shown in Table 2.<sup>3</sup> With these results one can easily derive corollaries using basic inequalities.

**The complete graph and cycles** For the complete graph and cycles, the results presented in Table 2 are trivial since these two graphs preserve invariance under permutation.

**The wheel graph** For wheel graph shown in Table 2, we have

$$\begin{aligned} \text{Periphery: } & \frac{1}{N!} \left( \sum_{k=0}^{N-4} \binom{N-4}{k} k!(N-k-1)! + \sum_{k=0}^{N-5} \binom{N-4}{k} (k+2)!(N-k-3)! \right), \\ \text{Center: } & \frac{1}{N} + \frac{1}{N!} \sum_{m=2}^{N-3} \sum_{k=2}^l T(N-1, k, m) (k-1)m!(N-m-1)!, \end{aligned}$$

where  $l = \min(m, N-m-1)$  and  $T(N, k, m) = \frac{N}{m} \binom{m}{k} \binom{N-m-1}{k-1}$ , and simplifying the above formulas gives the results in Table 2.

**The star graph** For the star graph, we have

$$\begin{aligned} \text{Periphery: } & \frac{1}{N!} \sum_{k=0}^{N-2} \binom{N-2}{k} k!(N-k-1)! = \frac{1}{2}, \\ \text{Center: } & \frac{1}{N} + \frac{1}{N!} \sum_{k=2}^{N-1} \binom{N-1}{k} (k-1)(k-1)!(N-k)! = \frac{N^2 - 3N + 4}{2N}. \end{aligned}$$

**The path graph** For path graph we have

$$\begin{aligned} \text{Ends: } & \frac{1}{N!} \sum_{k=0}^{N-2} \binom{N-2}{k} k!(N-k-1)! = \frac{1}{2}, \\ \text{Middle: } & \frac{2}{N!} \sum_{k=0}^{N-3} \binom{N-3}{k} k!(N-k-1)! = \frac{2}{3}. \end{aligned}$$

---

<sup>3</sup>Only the calculation of Shapley value is given and influence score can be easily obtained by normalization of Shapley value.

**The complete bipartite graph** For complete bipartite graph we have

$$\text{m side: } \frac{1}{(m+n)!} \left( \sum_{k=0}^{m-1} \binom{m-1}{k} k!(m+n-k-1)! + \sum_{k=2}^n \binom{n}{k} (k-1)k!(m+n-k-1)! \right),$$

$$\text{n side: } \frac{1}{(m+n)!} \left( \sum_{k=0}^{n-1} \binom{n-1}{k} k!(m+n-k-1)! + \sum_{k=2}^m \binom{m}{k} (k-1)k!(m+n-k-1)! \right),$$

which after simplifying the above gives the results in Table 2.

The above derivation has applied the following facts which can be easily verified by combinatorics:

$$\frac{1}{N!} \sum_{k=0}^{N-m} \binom{N-m}{k} k!(N-k-1)! = \frac{1}{m},$$

$$\frac{1}{(m+n)!} \sum_{k=2}^m \binom{m}{k} (k-1)k!(m+n-k-1)! = \frac{m(m-1)}{n(n+1)(m+n)},$$

$$\frac{1}{N!} \sum_{m=2}^{N-3} \sum_{k=2}^l T(N-1, k, m) (k-1)m!(N-m-1)! = \frac{(N-3)(N-4)}{6N},$$

where  $l = \min(m, N-m-1)$  and  $T(N, k, m) = \frac{N}{m} \binom{m}{k} \binom{N-m-1}{k-1}$ .

## B Experiment and Model Settings

Here we provide all the parameters set in the experiments.

Experiment	Graph		Grammar			
Model Setting	Component	Hidden	Model	Hidden (Total)		
	Struct2vec	32	SRN	200 (40800)		
			GRU	80 (19920)		
	Neural Network	16	LSTM	80 (19920)		
Optimizer			Rmsprop			
Dataset Setting	Class	3	Class	2		
	Total	6000	Total	G1 8677	G2 8188	G4 8188
	Training (per class)	5400 (1800)	Training	6073	5730	5730
	Testing (per class)	600 (200)	Testing	2604	2458	2458
	Prob.	0.02 - 0.21	Length	2- 13		
	Nodes	8-14				

Table 4: Parameter Settings For Neural Network Models and Datasets.