On the Privacy Risks of Model Explanations

Reza Shokri,[1] Martin Strobel, [1] Yair Zick [2]
[1] National University of Singapore
[2] University of Massachusetts, Amherst
reza@comp.nus.edu.sg, mstrobel@comp.nus.edu.sg, yzick@umass.edu

Abstract

Privacy and transparency are two key foundations of trustworthy machine learning. Model explanations offer insights into a model's decisions on input data, whereas privacy is primarily concerned with protecting information about the training data. We analyze connections between model explanations and the leakage of sensitive information about the model's training set. We investigate the privacy risks of feature-based model explanations using membership inference attacks: quantifying how much model predictions plus their explanations leak information about the presence of a datapoint in the training set of a model. We extensively evaluate membership inference attacks based on feature-based model explanations, over a variety of datasets. We show that backpropagation-based explanations can leak a significant amount of information about individual training datapoints. This is because they reveal statistical information about the decision boundaries of the model about an input, which can reveal its membership. We also empirically investigate the trade-off between privacy and explanation quality, by studying the perturbation-based model explanations.

1 Introduction

Black-box machine learning models are often used to make high-stakes decisions in sensitive domains. However, their inherent complexity makes it extremely difficult to understand the reasoning underlying their predictions. This development has resulted in increasing pressure from the general public and government agencies; several proposals advocate for deploying (automated) model explanations (Goodman and Flaxman 2017). In recent years, novel explanation frameworks have been put forward; Google, Microsoft, and IBM now offer model explanation toolkits as part of their ML suites.[1]

Model explanations offer users additional information about how the model made a decision with respect to their data records. Releasing additional information is, however, a risky prospect from a privacy perspective. The explanations, as functions of the model trained on a private dataset, might inadvertently leak information about the training set, beyond what is necessary to provide useful explanations. Despite

Copyright© 2021, Association for the Advancement of Artificial

2 Background and Preliminaries

We are given a labeled dataset $X \quad R^n$, with n features and k labels. The labeled dataset is used to train a model c, which maps each datapoint $\vec{x}$ in R^n to a distribution over k labels, indicating its belief that any given label $\vec{x}$ is Black-

is a practically feasible attack if the attacker has access to similar data sources.

3 Privacy Analysis of Backpropagation-Based Explanations

In this section we describe and evaluate our membership inference attack on gradient-based explanation methods. We use the Purchase and Texas datasets in (Nasr, Shokri, and Houmansadr 2018); we also test CIFAR-10 and CIFAR-100 (Sablayrolles et al. 2019), the Adult dataset (Dua and Graff 2017) as well as the Hospital dataset (Strack et al. 2014). The last two datasets are the only binary classi cation tasks considered. Where possible, we use the same training parameters and target architectures as the original papers (see Table 1 for an overview of the datasets). We study four types of information the attacker could use: loss, prediction variance, gradient variance and the SmoothGrad variance.

Table 1: Overview of the target datasets for membership inference

Name	Points	Features	Type	# Classes
Purchase	197,324	600	Binary	100
Texas	67,330	6,170	Binary	100
CIFAR-100	60,000	3,072	Image	100
CIFAR-10	60,000	3,072	Image	10
Hospital	101,766	127	Mixed	2
Adult	48,842	24	Mixed	2

Table 2: The average training and testing accuracies of the target models.

	Purchase	Texas	CIFAR -100	CIFAR -10	Hospital	Adult
Train	1.00	0.98	0.97	0.93	0.64	0.85
Test	0.75	0.52	0.29	0.53	0.61	0.85

3.1 General setup

For all datasets, we rst create one big dataset by merging the original training and test dataset, to have a large set of points for sampling. Then, we randomly sample four smaller datasets that are not overlapping. We use the smaller sets to train and test four target models and conduct four attacks. In each instance, the other three models can respectively be used as shadow models. We repeat this process 25 times, producing a total of 100 attacks for each original dataset. Each small dataset is split 50/50 into a training set and testing set. Given the small dataset, the attacker has an a priori belief that 50% of the points are members of the training set, which is the common setting for this type of attack (Shokri et al. 2017b).

3.2 Target datasets and architectures

The overview of the datasets is provided in Table 1 and an overview of the target models accuracies in Table 2.

Purchase dataset The dataset originated from the "Acquire Valued Shoppers Challenge" on Kaggle.[4] The goal of the challenge was to use customer shopping history to predict shopper responses to offers and discounts. For the original membership inference attack, Shokri et al. (2017b) create a simpli ed and processed dataset, which we use as well. Each of the 197,324 records corresponds to a customer. The dataset has 600 binary features representing customer shopping behavior. The prediction task is to assign customers to one of 100 given groups (the labels). This learning task is rather challenging, as it is a multi-class learning problem with a large number of labels; moreover, due to the relatively high dimension of the label space, allowing an attacker access to the prediction vector — as is the case in (Shokri et al. 2017b) — represents signi cant access to information. We sub-sampled smaller datasets of 20,000 points i.e. 10,000 training and testing points for each model. We use the same architecture as (Nasr, Shokri, and Houmansadr 2018), namely a four-layer fully connected neural network with tanh activations. The layer sizes are [1024, 512, 256, 100]. We trained the model of 50 epochs using the Adagrad optimizer with a learning rate of 0.01 and a learning rate decay of 1e-7.

Texas hospital stays The Texas Department of State Health Services released hospital discharge data public use les spanning from 2006 to 2009.[5] The data is about inpatient status at various health facilities. There are four different groups of attributes in each record: general information (e.g., hospital id, length of stay, gender, age, race), the diagnosis, the procedures the patient underwent, and the external causes of injury. The goal of the classi cation model is to predict the patient's primary procedures based on the remaining attributes (excluding the secondary procedures). The dataset is ltered to include only the 100 most common procedures. The features are transformed to be binary resulting in 6,170 features and 67,330 records. We sub-sampled smaller datasets of 20,000 points i.e. 10,000 training and testing points for each model. As the dataset has only 67,330 points we allowed resampling of points. We use the same architecture as (Nasr, Shokri, and Houmansadr 2018), namely a ve-layer fully connected neural network with tanh activations. The layer sizes are [2048, 1024, 512, 256, 100]. We trained the model of 50 epochs using the Adagrad optimizer with a learning rate of 0.01 and a learning rate decay of 1e-7.

CIFAR-10 and CIFAR-100 CIFAR-10 and CIFAR-100 are well-known benchmark datasets for image classi cation (Krizhevsky and Hinton 2009). They consists of 10 (100) classes of $32 \times 32 \times 3$ color images, with 6,000 (600) images per class. The datasets are usually split in 50,000 training and 10,000 test images. For CIFAR-10, we use a small convolutional network with the same architecture as in (Shokri et al. 2017b; Sablayrolles et al. 2019), it has two convolutional layers with max-pooling, and two dense layers, all with Tanh activations. We train the model for 50 epochs with a learn-

[4] https://www.kaggle.com/c/acquire-valued-shoppers-challenge/data

[5] https://www.dshs.texas.gov/THCIC/Hospitals/Download.shtm

ing rate of 0.001 and the Adam optimizer. Each dataset has 30,000 points (i.e. 15,000 for training). Hence, we only have enough points to train one shadow model per target model. For CIFAR-100, we use a version of Alexnet (Krizhevsky, Sutskever, and Hinton 2012), it has ve convolutional layers with max-pooling, and to dense layers, all with ReLu activations. We train the model for 100 epochs with a learning rate of 0.0001 and the Adam optimizer. Each dataset has 60,000 points (i.e. 30,000 for training). Hence, we don't have enough

Figure 1: Results for the threshold-based attacks using different attack information sources. The ESTIMATE attack uses the optimal threshold; the SHADOW trains a shadow model on data from the same distribution, and uses an optimal threshold for the shadow model. Using three such models results in nearly optimal attack accuracy.

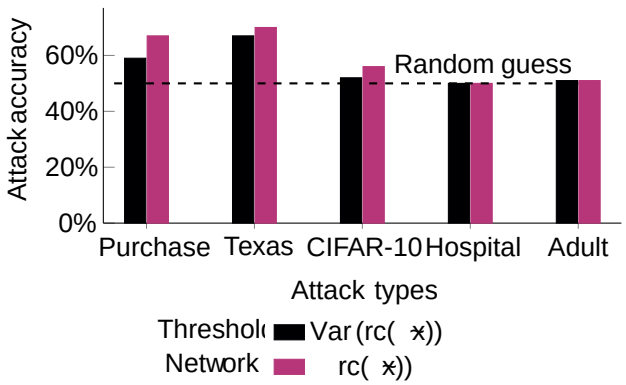


Figure 2: A comparison between attacks using only the variance of the gradient and attacks using the entire gradient explanation as input.

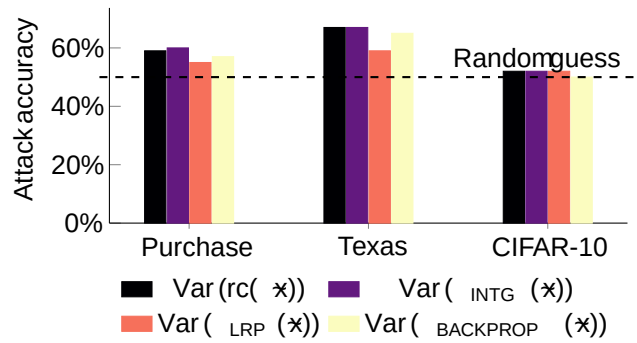


Figure 3: Results for the threshold-based attacks using-different backpropagation-based explanations as sources of information for the attacker.

Shokri et al. (2017b) to ensure a valid comparison between the methods. We trained the attack model for 15 epochs using the Adagrad optimizer with a learning rate 0.01 of and a learning rate decay of 1e-7. As data for the attacker, we used 20,000 explanations generated by the target 10,000 each for members and non-members. The training testing split for the attacker was 0.7 to 0.3. We repeated the experiment 10 times. We omitted CIFAR-100 for computational reasons.

As can be seen in Figure 2, attacks based on the entire explanation perform slightly better than attacks based only on the variance. However, they are qualitatively the same and still perform very poorly for CIFAR-10, Adult, and Hospital.

3.4 Combining different information sources

The learning attacks described in the previous paragraph allow for a combination of different information sources. For example, an attacker can train an attack network using both the prediction and the explanation as input. Experiments on combining the three information sources (explanation, prediction, and loss) lead to outcomes identical to the strongest used information source. Especially if the loss is available to an attacker, we could not nd evidence that either the prediction vector or an explanation reveals additional information.

3.5 Results for other backpropagation-based explanations

Besides the gradient, several other explanation methods based on backpropagation have been proposed. We conducted the attack described in Section 2.2 replacing the gradient with some other popular of these explanation methods. The techniques are all implemented in the INNVESTIGATE library[7] (Alber et al. 2018). An in-depth discussion of some of these measures, and the relations between them, can also be found in (Ancona et al. 2018). As can be seen in Figure 3 on the Purchase, Texas, and CIFAR-10 datasets, the results for other backpropagation based methods are relatively similar to the attack based on the gradient. Integrated gradients performing most similar to the gradient. For Adult, Hospital and CIFAR-100 small-scale experiments indicated that this type of attack would not be successful for these explanations as well, we omitted the datasets from further analysis.

[7] https://github.com/albermax/innvestigate

4 Analysis of factors of information leakage

In this section, we provide further going analysis to validate our hypothesis and broaden understanding.

4.1 The In uence of the Input Dimension

The experiments in Section 3 indicate that $Var(rc(x^*))$, and $\|rc(x^*)\|_1$ predict training set membership. In other words, high absolute gradient values at a point x^* signal that x^* is not part of the training data: the classi er is uncertain about the label of x^*, paving the way towards a potential attack. Let us next study this phenomenon on synthetic datasets, and the extent to which an adversary can exploit model gradient information in order to conduct membership inference attacks. The use of arti cially generated datasets offers us control over the problem complexity, and helps identify important facets of information leaks.

To generate datasets, we use the Sklearn python library.[8] For n features, the function creates an n-dimensional hyper cube, picks a vertex from the hypercube as center of each class, and samples points normally distributed around the centers. In our experiments, the number of classes is either 2 or 100 while the number of features is between 1 to 10,000 in the following steps,

$$n \in \{1;2;5;10;14;20;50;100;127;200;500;600;\\ 1000;2000;3072;5000;6000;10000\}.$$

For each experiment, we sample 20,000 points and split them evenly into training and test set. We train a fully connected neural network with two hidden layers with fty nodes each, the $\tanh$ activation function between the layers, and softmax as the nal activation. The network is trained using Adagrad with learning rate of 0.01 and learning rate decay of 10^{-7} for 100 epochs.

Increasing the number of features does not increase the complexity of the learning problem as long as the number of classes is xed. However, the dimensionality of the hyper plane increases, making its description more complex. Fur thermore, for a xed sample size, the dataset becomes increasingly sparse, potentially increasing the number of points

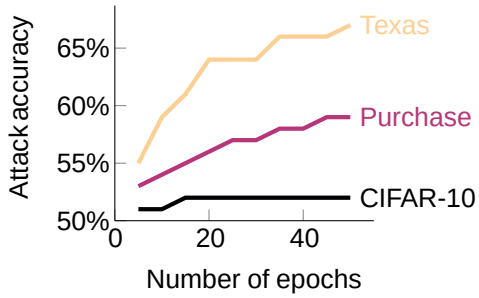


Figure 5: The attack accuracy of the attacker increases with increasing number of epochs.

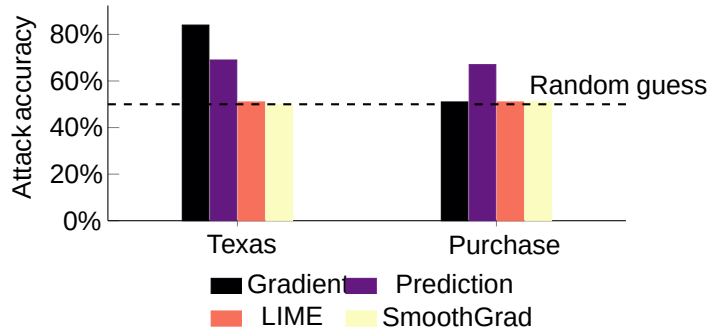


Figure 6: Attacks using LIME or SmoothGrad do not outperform random guessing in any of our experiments.

5.1 Attacks using LIME explanation

As a second perturbation-based method, we looked at the popular explanation method LIME (Singh, Ribeiro, and Guestrin 2016). The type of attack is the same as described in Section 2.2. We use an optimal threshold based on the variance of the explanation. However, the calculation of LIME explanations takes considerably longer than the computation of other methods we considered. Every single instance computes for a few seconds. Running experiments with 10,000 or more explanations would take weeks to months. To save time and energy, we restricted the analysis of the information-leakage of LIME to smaller-scale experiments where the models train on 1,000 points, and the attacks run on 2,000 points each (1,000 members and 1,000 non-members). We also repeated each experiment only 20 times instead of 100 as for the others. Furthermore, given that the experiments for the other explanations indicated that only for Purchase and Texas the attack was likely to be successful, we restricted our experiments to these two datasets. Figure 6 shows the results for these attacks. To ensure that it is not the different setting that determines the outcome, we also rerun the attacks for the gradient and SmoothGrad explanations, as well as the attack using the prediction variance in this new setting. Neither LIME nor SmoothGrad outperforms random guessing. For the Purchase dataset, however, the attack using the gradient variance fails as well. As a final interesting observation, which we are unable to explain at the moment: For the Texas dataset, the gradient-based attack performs better than on the larger dataset (shown in Figure 1) it even outperforms the attack based on the prediction in this specific setting. Something we want to explore further in future works.

5.2 Analysis

While it is entirely possible that perturbation-based methods are vulnerable to membership inference, we conjecture that this is not the case. This conjecture is due to an interesting connection between perturbation-based model explanations and the data-manifold hypothesis (Fefferman and Mitter 2016). The data-manifold hypothesis states that "data tend to lie near a low dimensional manifold" (Fefferman and Mitter 2016, p. 984). Many works support this hypothesis (Belkin and Niyogi 2003; Brand 2003; Narayanan and Mitter 2010), and use it to explain the pervasiveness of adversarial examples (Gilmer et al. 2018). To the best of our knowledge,

governance; both expressed a great deal of interest in the potential impact of our work on the ongoing debate over model explainability and its potential effects on user privacy.

Belkin, M.; and Niyogi, P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation 15(6): 1373–1396.

Brand, M. 2003. Charting a manifold. In Proceedings of the 17th Annual Conference on Neural Information Processing Systems (NIPS), 985–992.

Carlini, N.; Liu, C.; Kos, J.; Erlingsson, Ú.; and Song, D. 2018. The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets. arXiv preprint arXiv:1802.08232 .

Daligault, J.; and Thomassé, S. 2009. On Finding Directed

Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision 115(3): 211–252.

Sablayrolles, A.; Douze, M.; Ollivier, Y.; Schmid, C.; and Jégou, H. 2019. White-box vs Black-box: Bayes Optimal Strategies for Membership Inference. In Proceedings of the 36th International Conference on Machine Learning (ICML), 5558–5567.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017a.

A The In uence of the Input Dimension

In Figure 7 we report experiments on the in uence of the input dimension with networks of smaller/larger capacity, which have qualitatively similar behavior to our baseline model. However, the interval of in which correlation exists and the amount of correlation varies.

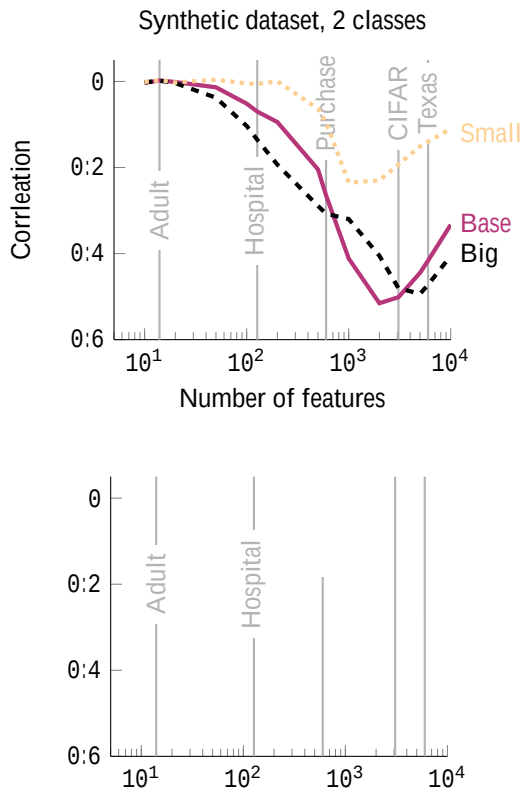


Figure 7: The correlation between $\|r_c(x)\|_1$ and training membership for synthetic datasets for increasing number of features and different number of classes $k \in \{2, 100\}$ for three different networks. The "Small" has one hidden layer with 5 nodes, "Base" has two layers with 50 nodes each, "Big" has 3 layers with 100 nodes each.

B Example-Based Explanations

To illustrate the privacy risk of example-based model explanations, we focus on the approach proposed by Koh and Liang (2017). It aims at identifying in uential datapoints; that is, given a point of interest y, nd a subset of points from the training data $(x, y) \in X$ that explains the label $c(y)$, where $\hat{}$ is a parameterization induced by a training algorithm A. It selects a training point x_{train} by measuring the importance of x_{train} for determining the prediction for y.
To estimate the effect of x_{train} on y, the explanation measures the difference in the loss function over y when the model is trained with and without x_{train}. Let $_{x_{train}} = A(X \setminus \{x_{train}\})$, in words, $_{x_{train}}$ is induced by training algorithm A given the dataset excluding x_{train}. The in uence of x_{train} on y is then

$$I_y(x_{train}) = L(y; _{x_{train}}) - L(y; \hat{}). \tag{2}$$

The Koh and Liang explanation releases k points with the highest absolute in uence value according to Equation (2). Additionally, it might release the in uence of these k points (the values of $I_y(z)$ as per Equation (2)), which allows users to gauge their relative importance.

C Membership-inference via Example-Based Model Explanations

In this section we analyze how many training points an attacker can recover with access to example-based explanations. We focus on logistic regression models, for which example-based explanations were originally used (Koh and Liang 2017). The results can be generalized to neural networks by focusing only on the last layer, we demonstrate this by considering a binary image classi cation dataset used in (Koh and Liang 2017); we call this dataset Dog/Fish. We also focus only on binary classi cation tasks. While technically the approaches discussed in previous sections could be applied to this setting as well, example-based explanations allow for stronger attacks. Speci cally, they explicitly reveal training points, and so offer the attacker certainty about a points' training set membership (i.e. no false positives). Formally, we say a point y reveals point x if for all $z \in X$, $|I_y(x)| \geq |I_y(z)|$. In other words, x will be offered if one requests a example-based explanation for y. Similarly, y k-reveals point x if there is a subset $S \subset X$, $|S| = k - 1$ such that $\forall z \in X \setminus S$: $|I_y(x)| \geq |I_y(z)|$. Hence, x will be one of the points used to explain the prediction of y if one releases the top k most in uential points. A point $x \in X$ that (k-)reveals itself, is called (k-)self-revealing.

Revealing membership for example-based explanations
While for feature-based explanations the attacker needs to rely on indirect information leakage to infer membership, for example-based explanations, the attacker's task is relatively simple. Intuitively, a training point should be in uential for its own prediction, so an example-based explanation of a training set member is likely to contain the queried point, revealing membership. We test this hypothesis via experiments, i.e. for every point $x \in X$ in the training set we obtain the $k \in \{1, 5, 10\}$ most in uential points for the prediction $f(x)$ and see if x is one of them.

C.1 Experimental setup

While the theoretical framework of in uence functions described in Section 2 can be applied to an arbitrary classi cation task, it requires the training of as many classi ers as there are points in the training set in practice. Koh and Liang (2017) propose an approximation method, but currently we only have access of its implementation for binary logistic

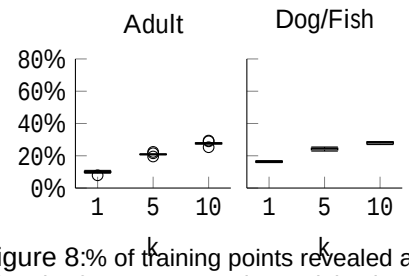


Figure 8: % of training points revealed as their own explanation, when $k \in \{1, 5, 10\}$ most influential points are revealed.

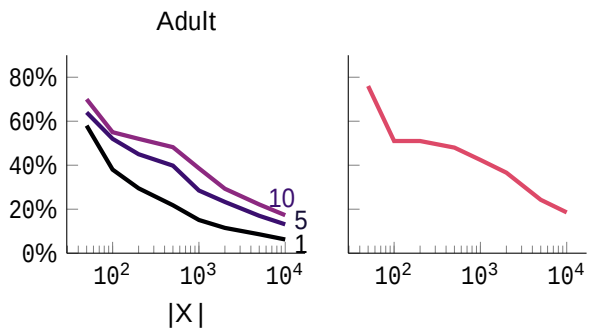


Figure 9: % of k-self-revealing points depending on the size of the dataset $|X|$ for $k \in \{1, 5, 10\}$. Increasing datasize the relative number of self-revealing points decreases. However, the decrease is relatively slow (the figure is in logarithmic scale) and the absolute number of revealed points is actually increasing.

Adult and Hospital datasets from our previous experiments for which we train binary logistic regression models. Furthermore, given the relatively long time it takes to compute all models, we reduce the size of the training set to 2,000 points, so we can run one experiment within a few hours (see Figure 9 for an exploration of the effect of dataset size). Koh and Liang (2017) use a specifically created dataset containing 2400 299 × 299-pixel dog and fish images, extracted from ImageNet (Russakovsky et al. 2015). The images were preprocessed by taking the output of the penultimate layer of a neural network trained on ImageNet.[10] These latent representations were then used to train a linear regression model. For more variety we include this dataset here as Dog/Fish dataset. Arguably, this dataset doesn't contain particularly private information, but it can be seen as representative of image data in general. It also allows us to attack a pre-trained deep neural network which last layer was fine tuned. This type of transfer learning for small dataset becomes increasingly popular. Given the small size of this dataset we randomly split it into 1800 training and 600 test points for each experiment. For each dataset we repeat the experiment 10 times.

C.2 Evaluation

Figure 8 shows the percentage of training points that would be revealed by explaining themselves. For the standard setting where the top 5 most influential points are revealed, a quarter of each dataset is revealed on average. For the Hospital dataset, two thirds of training points are revealed. Even when just the most influential point would be released for the Adult dataset (which exhibits the lowest success rates), 10% of the members are revealed through this simple test.

As mentioned in Appendix B, the influence score of the most influential points might be released to a user as well. In our experiments, the influence scores are similarly distributed between training and test points (i.e. members and non-members); however, the distribution is significantly different once we ignore the revealed training points. Figure 10 illustrates this for one instance of the Dog/Fish dataset; similar results hold for the other datasets. An attacker can exploit these differences, using techniques similar to those discussed in Section 2.2; however, we focus on other attack models in this work.

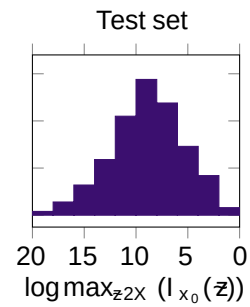


Figure 10: Histogram of the influence of the most influential points for every point in the training set (left) and test set (right) on a logarithmic scale (for one instance of Dog/Fish). The points in the training set for which the membership inference is successful (i.e. $x_0 = \operatorname{argmax}_{z \in X} (I_{x_0}(z))$) are highlighted in red (dark).

C.3 Minority and outlier vulnerability to inference attacks for example-based explanations

Visual inspection of datapoints for which membership attacks were successful indicates that outliers and minorities are more susceptible to being part of the explanation. Images of animals (a bear, a bird, a beaver) eating fish (and labeled as such) were consistently revealed (as well as a picture containing a fish as well as a (more prominent) dog that was labeled as fish). We label three "minorities" in the dataset to test the hypothesis that pictures of minorities are likelier to be revealed (Table 3a).

With the exception of $k = 1$ (for lion and clown fish), minorities are likelier to be revealed. While clown fish (which are fairly "standard" fish apart from their distinct coloration) exhibit minor differences from the general dataset, birds are more than three times as likely to be revealed. The Hospital dataset exhibits similar trends (Table 3b). Young children, which are a small minority in the dataset, are revealed to a greater degree; ethnic minorities also exhibit slightly higher rates than Caucasians. This is mirrored (Table 3c) for the Adult dataset, with the exception of the age feature and the

[10] Specifically, the authors used an Inceptionv3 architecture, for which pre-trained models are available for Keras: https://keras.io/applications/\#inceptionv3.

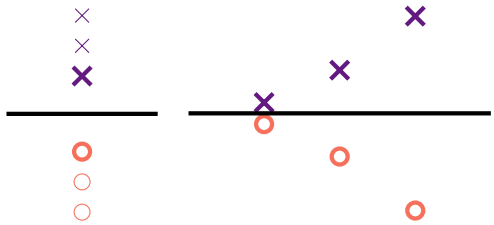


Figure 11: Illustrations of datasets for which only two (left) or all (right) points can be revealed under standard training procedures.

□

D.1 Bounds on number of revealable points

It is relatively easy to construct examples in which only two points in the dataset can be revealed (see Figure 11). In fact, there are speci c instances in which only a single datapoint can be revealed (see Lemma D.1), however these cases are neither particularly insightful nor do they re ect real-world settings. On the other side, there exists datasets where, independent of the number of features, the entire dataset can be recovered. The right side of Figure 11 illustrates such an example.

The following Lemma characterizes the situations in which only a single point of the dataset can be revealed X for R. The conditions for higher dimensions follow from this.

Lemma D.1. Given a training set X let f and F_X be induced by X with $|F_X| \geq 2$, then one of the following statements is true

1. $\forall x \in X : w = w_*$ and $(b \geq b_*; \forall x \in X) \vee (b \leq b_*; \forall x \in X)$ (i.e all functions in F_X are shifted in one direction of f),
2. $\exists y \in R^n : \forall x \in X : w_* \cdot y + b_* = w \cdot y + b$ (i.e all functions in F_X intersect with f in the same point) ,
3. $\forall x \in X : w = w_*$ and there exists a numbering of the points in X such that $b_1 \leq b_2 \leq \ldots \leq b_k \leq b \leq b_{k+1} \leq \ldots \leq b_m$ such that $b \geq \log(\frac{1}{2}(e^{b_1} + e^{b_m}))$.
4. at least two points can be revealed.

Proof. It is easy to see that in the rst two situations only one point can be revealed (the one corresponding to the largest shift or largest angle at intersection). In the third case all functions in F_X are shifts of f, but not all in the same direction. Only, the left most and right most shift are candidates for being revealed as they clearly dominate all other shifts. Also we assume $b_1 < b < b_m$ (as soon as one shift coincidences with the original model the statement is trivially true). Some calculus reveals the condition for which the two points would have the same in uence is

$$y = \ln \frac{2e^{b} + e^{b_1} + e^{b_m}}{e^{b+b_1} + e^{b+b_m} - 2e^{b_1+b_m}} \cdot w$$

, which is well de ned when the expression inside the logarithm is positive and $e^{b+b_1} + e^{b+b_m} - 2e^{b_1+b_m} \neq 0$ The for b

□

the model parameters $= (\mathbf{w}; b)$ are induced by

$\square$

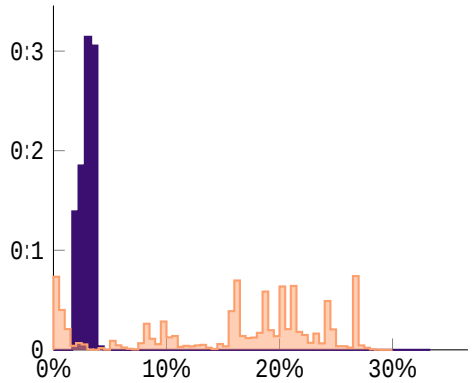


Figure 12: Distribution over the size of the revealed training data, starting from a random point in the in uence graph. This is obtained by averaging over 10 initiations of the training set. Each time roughly the size of the largest SCC is recovered. For the Hospital dataset the size of the SCC varies the most, hence the multimodal distribution.

Dataset	n	jX j	# points recovered	% of jX j
Dog/Fish	2048	1,800	1790	99.4
Adult	104	2,000	91.5	4.6
Hospital	127	2,000	81.1	4.0

Table 4: The number of points recovered using our attack based on subspace reduction. For small sized, high dimensional data, notably Dog/Fish, the attack can recover (almost) the entire dataset.

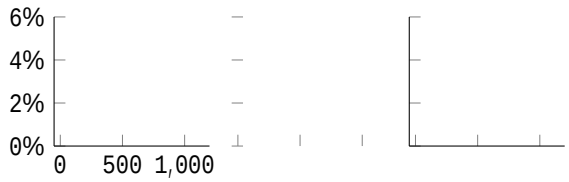
Number of strongly connected components (SCCs): a high number of SCCs implies that the training set is harder to recover: an adaptive algorithm can only extract one SCC at a time. It also implies that the underlying prediction task is fragmented: labels in one part of the dataset are independent from the rest.
Size of the SCCs: large SCCs help the attacker: they are more likely to be discovered, and recovering just some of them already results in recovery of signi cant portions of the training data.
Distribution of in-degrees: the greater a node's in-degree is, the likelier its recovery; for example, nodes with zero in-degree may be impossible for an attacker to recover. Generally speaking, a uniform distribution of in-degrees makes the graph easier to traverse.

	Dog/Fish	Adult	Hospital
#SCC	1709	1742	1448
#SCC of size 1	1679	1701	1333
Largest SCC size	50	167	228
Max in-degree	1069	391	287
#node in-degree=0	1364	1568	727

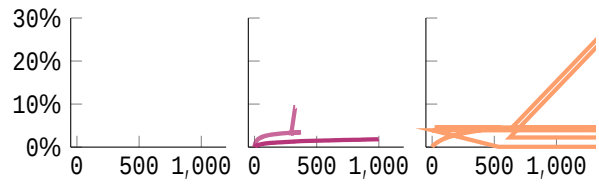
Table 5: Some key characteristics of the in uence graphs induced by the example-based explanations (for lik,i9e2d (1568 (8.

E Dataset Reconstruction for Example-Based Explanations

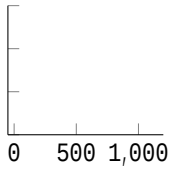
We evaluate our reconstruction algorithms on the same datasets as the membership inference for example-based explanations. In this section we describe the results for the two approaches described above. In Appendix E.3 we compare them to some general baselines.



(a) Uniform datapoint sampling.



(b) Marginal feature distribution sampling (for the Dog/Fish dataset points are sampled in the latent space).



(c) True point distribution sampling.

Figure 13:% of training data revealed by an attacker using dif-