SEQUENTIAL GAME NETWORK (SEGANE) WITH APPLICATION TO ONLINE DATA SANITIZATION

Zahir Alsulaimawi, Jinsub Kim, Thinh Nguyen

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331
alsulaiz@oregonstate.edu, kimjinsu@oregonstate.edu, thinhq@eecs.oregonstate.edu

ABSTRACT

This paper proposes SEquential GAme NEtwork (SEGANE), a novel deep neural network (DNN) architecture for optimizing the performance of machine learning applications with multiple competing objectives. Specifically, SEGANE is evaluated in the context of data sanitization which aims to remove any pre-specified private information from the data in real time while keeping the relevant information used to improve the inference accuracy about the non-private information. In some settings, preserving private information and improving inference performance about non-private information are competing objectives. In such cases, SEGANE provides a sequential game framework and algorithmic tools to implement data sanitization schemes with flexible trade-off between these two objectives. We use two datasets: MNIST (hand-written digits) and IMDB (gender and age) to evaluate SEGANE. For MNIST, even numbers are considered private while numbers larger than 10 are considered non-private. For IMDB, in one setting, gender is considered private while age is non-private, and vice versa in another setting. Our experimental results on these datasets show that SEGANE is highly effective in removing private information from the dataset while allowing non-private data to be mined effectively.

***Index Terms*—** Privacy-preserving machine learning, deep learning, sequential game

1. INTRODUCTION

Recent years have witnessed the proliferation of Artificial Intelligence (AI) technologies that fundamentally change the ways we live, play, and work. From drug design research to consumer products such as smart phones and self-driving cars, all are driven by AI/machine learning algorithms that help better our lives. Central to these ever more intelligent algorithms is the vast amount of data being collected and shared by billions of interconnected wireless devices/sensors. While more data leads to better machine learning algorithms, data collection and sharing in an insecure and open environment such as the Internet where most of Internet of Things (IoT) devices operate, will undermine the user's security and privacy issues. For example, Amazon Alexa allows one to conveniently access information on the Internet or to order pizza by voice command, but it also inadvertently records the customer's private conversations in the background.

Cryptography has long been used to address the privacy and security issues. However, cryptographic techniques are computationally expensive for the massive amount of data being collected constantly. Furthermore, it is difficult to design a secure, efficient, and reliable mechanism for key exchange/distribution required for cryptographic techniques in an open environment. All cryptographic measures are also based on the assumption that it is computationally infeasible to decipher an encoded message without the knowledge of the secret keys. This assumption has not been mathematically proven, and many implementations of cryptographic schemes have been shown to be vulnerable [1]. Importantly, encrypted data prevents a benevolent machine learning algorithm to use the raw data to mine non-private information. To that end, privacy-preserving algorithms have been proposed to transform the raw data into the "sanitized" data in such a way that private information cannot be mined from the sanitized data while the sanitized data can be used to mine some other non-private information.

Related Work. Within the private preserving framework, there exist privacy-preserving data mining (PPDM) techniques in the database community [2] [3] [4] whose goal is to prevent association of any instance in a database to a person. In addition to PPDM, many privacy-preserving machine learning (PPML) techniques [5],[6], [7], [8], [9], [10] have been proposed to deal with data beyond those in the traditional databases. Most existing PPML literature focus on ensuring that the private information cannot be mined and make no assumption about the non-private information. On the other hand, our work assumes pre-specified sets of private and non-private information. Such formulation not only makes the proposed data sanitization more effective, but also provides a flexible trade-off between privacy and the ability to mine non-private information from the sanitized data.

Contributions. In this paper, we propose a novel deep neural network (DNN) architecture called SEquential GAme NEtwork (SEGANE) for optimizing the performance of machine learning applications with multiple competing objectives. Specifically, SEGANE is evaluated in the context of data sanitization which aims to remove any pre-specified private information from the data while keeping the relevant information used to improve the inference accuracy about some pre-specified non-private information. In some settings, preserving private information and improving inference performance about non-private information are competing objectives. In such cases, SEGANE provides a sequential game framework and algorithmic tools to implement data sanitization schemes with flexible trade-off between these two objectives. We use two datasets: MNIST (hand-written digits) and IMDB (gender and age) to evaluate SEGANE. For MNIST dataset, even numbers are considered private while numbers larger than 10 are considered non-private. For IMDB dataset, in one setting, gender is considered private while age is non-private, and vice versa in another setting. Our experimental results on these datasets show that SEGANE is highly effective in removing private information from the dataset while allowing non-private data to be mined effectively.

2. ONLINE DATA SANITIZATION

We consider the problem of online data sanitization in which we aim to transform a raw feature vector such that occurrence of certain *private* event cannot be inferred from the transformed feature vector, which we refer to as sanitized features, while occurrence of certain *public* event can be inferred as efficiently as in the case of using the raw features. Specifically, we assume that there is a training dataset $\mathcal{D} = \{(x^i, p^i, b^i), i = 1, \ldots, N\}$ where x^i is the i-th feature vector and p^i and b^i are binary labels associated with x^i, representing occurrence (label 1) or non-occurrence (label 0) of the public event and the private event respectively. We assume that (x^i, p^i, b^i)'s are independent and identically distributed samples from an unknown joint distribution $F(x, p, b)$. Based on the training dataset $\mathcal{D}$, our objective is to find an optimal sanitizer transformation S within $\mathcal{S}$, which denotes the set of feasible sanitizer transformations. In particular, we aim to solve the following minimization problem to find an optimal sanitizer design:

$$\min_{S\in\mathcal{S}} \left\{ \min_{P} \mathbb{E}[\mathcal{L}(P(S(x)), p)] - \lambda \min_{B} \mathbb{E}[\mathcal{L}(B(S(x)), b)] \right\} \quad (1)$$

In the above minimization, $\mathcal{L}$ denotes the loss function (e.g., cross-entropy loss, Hamming loss), $\min_{P} \mathbb{E}[\mathcal{L}(P(S(x)), p)]$ represents the minimum expected loss one can achieve for public event detection using the sanitized features $S(x)$, and $\min_{B} \mathbb{E}[\mathcal{L}(B(S(x)), b)]$ denotes the minimum expected loss one can achieve for private event detection using the sanitized features. The hyperparameter λ is set to properly balance the two objectives: (i) make private event detection based on the sanitized features infeasible and (ii) ensure that public event detection can be effectively performed based on the sanitized features.

In practice, we envision that the online sanitizer can be implemented at the data acquisition stage of sensor hardwares as illustrated in Fig. 1. The sanitizer will perform in situ data sanitization at a sensor such that only the sanitized sensor measurements will be transmitted over a possibly insecure network environment. Therefore, even when a data breach or a cyber attack occurs in the network or the server, only the sanitized data will be revealed to adversaries; adversaries will not be able to mine private event information from the sanitized data they acquire, even with the knowledge of the sanitizer design. The set $\mathcal{S}$ of feasible sanitizers is determined according to the types of sanitizers that can be implemented under the practical constraints of the system, considering computation, latency, and memory requirements. For instance, for an indoor person localization system based on sensors to be deployed in a public space, we can implement online sanitizers at sensors such that identity of people monitored by the sensors (private event information) cannot be inferred from the sanitized data while the number and locations of people in the space (public event detection) can be effectively inferred based on the sanitized data.

Sequential Game Formulation. The online data sanitization problem can be seen as a sequential game with three players: the sanitizer designer, the public event detector designer, and the private event detector designer. The players do not know the joint distribution of (x, p, b), but they have access to the training dataset $\mathcal{D}$. The game is played in the following sequence [11]:

- Stage 1: The sanitizer designer chooses a sanitizer transformation S from $\mathcal{S}$.
- Stage 2: The detector designers gain the knowledge of the sanitizer S. Given the knowledge of the sanitizer, the public event detector designer chooses the detector P, and the private event detector designer chooses the detector B.
- Payoffs: The payoff to the public event detector designer is $-\mathbb{E}[\mathcal{L}(P(S(x)), p)]$. The payoff to the private event detector designer is $-\mathbb{E}[\mathcal{L}(B(S(x)), b)]$. The payoff to the sanitizer designer is

$$-\mathbb{E}[\mathcal{L}(P(S(x)), p)] + \lambda \mathbb{E}[\mathcal{L}(B(S(x)), b)].$$

In choosing their strategies, each player aims to maximize its payoff. In other words, the detectors aim to minimize their loss terms while the sanitizer aims to solve the optimization problem (1). Note that the objective in online data sanitization problem corresponds to finding a good strategy for the sanitizer designer.

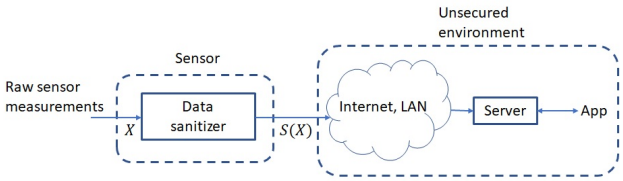

Fig. 1. Implementation of online data sanitization.

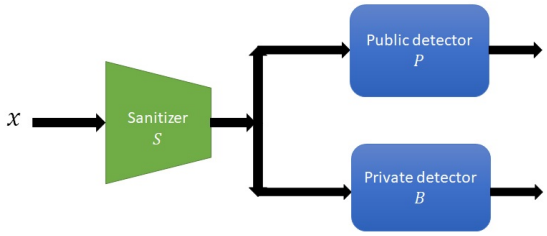

Fig. 2. Architecture of SEGANE for data sanitization.

SEquential GAme NEtwork. We present a novel deep neural network architecture, referred to as SEquential GAme NEtwork (SEGANE), which we can use to find a local optimum design of the sanitizer. In solving the sanitization problem (1), as the distribution information is not available, the expected loss terms cannot be computed. Therefore, we replace the expected loss terms with the empirical loss terms. Specifically, the optimization formulation with the empirical loss terms can be written as

$$\min_{S\in\mathcal{S}} \left\{ \min_{P\in\mathcal{P}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(P(S(x^i)), p^i) - \lambda \min_{B\in\mathcal{B}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(B(S(x^i)), b^i) \right\} \quad (2)$$

where $\mathcal{P}$ and $\mathcal{B}$ denotes the classes of classifiers represented by certain machine learning models. Similarly, in analyzing the sequential game, the expected loss terms in the payoffs are replaced by the above empirical loss terms.

Fig. 2 provides an overview of SEGANE. In SEGANE, the sanitizer S, the public event detector P, and the private event detector B are restricted to belong to certain classes of mappings represented by some machine learning models (e.g., neural networks, logistics regression), denoted by $\mathcal{S}$, $\mathcal{P}$, and $\mathcal{B}$ respectively. In this paper, we

focus on the case where $\mathcal{S}$ is the set of functions that can be represented as a deep neural network (DNN) encoder associated with certain DNN architecture, and $\mathcal{P}$ and $\mathcal{B}$ are the sets of classifiers that can be represented as DNN classifiers associated with certain DNN architectures [12]. We use SEGANE to emulate the sequential game of data sanitization for different santizer designs chosen by the sanitizer designer in Stage 1; based on the emulated game results, we incrementally improve the sanitizer design to make it converge to a local optimum. Specifically, we initialize the sanitizer design to $S^{(0)}$ and the iteration index k to 1 and then perform the following iterative procedure to train the sanitizer:

1. Update public event and private event detectors: assuming that the sanitizer design S is set to $S^{(k-1)}$, the public event detector designer and the private event detector designer use the sanitized training dataset $\mathcal{D}_{S^{(k-1)}} = \{(S^{(k-1)}(x^i), p^i, b^i), i = 1, \ldots, N\}$ to train their respective detectors, $P^{(k)}$ and $B^{(k)}$. In other words, they train their respective DNNs to compute $P^{(k)}$ and $B^{(k)}$ to minimize the empirical losses. This step emulates the player actions at Stage 2 of the sequential sanitize-learn game when the sanitizer designer chose $S^{(k-1)}$ as the sanitizer in Stage 1.

2. Update sanitizer design: while fixing P and B to $P^{(k)}$ and $B^{(k)}$, we compute a stochastic gradient direction using a small batch of training data points and use it to make incremental improvement of the sanitizer design and set $S^{(k)}$ to the new sanitizer.

3. Terminate and return $S^{(k)}$ if $k = K$. If $k < K$, increase k by 1 and go to Step 1.

Note that in updating P and B in each iteration, we aim to look for globally optimal detectors for the given sanitizer design. But, in updating S in each iteration, we are making an incremental improvement based on a stochastic gradient direction. This update structure is intended to ensure that the sanitizer design will converge to a local optimum[1]. The details of the sanitizer training procedure using SEGANE with the stochastic gradient descent method are presented in Algorithm 1. In Algorithm 1, θ_S, θ_P, and θ_B denote the DNN parameters for S, P, and B respectively.

Algorithm 1 The hyperparameter d ($d \gg 1$) is the number of stochastic gradient descent steps to be used for updating P and B in each iteration.

for K training iterations **do**
  for d steps **do**
    • Sample $\{x^1, ..., x^m\}$, a batch, from dataset.
    • Update P (θ_P) by descending its stochastic gradient:
$$\nabla_{\theta_P} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(P(S(x^i)), p^i)$$
    • Update B (θ_B) by descending its stochastic gradient:
$$\nabla_{\theta_B} \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(B(S(x^i)), b^i)$$
  end for
  • Sample $\{x^1, ..., x^m\}$ a batch from dataset.
  • Update S (θ_S) by descending its stochastic gradient:
$$\nabla_{\theta_S} \frac{1}{m} \sum_{i=1}^{m} \Big\{ \mathcal{L}(P(S(x^i)), p^i) - \lambda \mathcal{L}(B(S(x^i)), b^i) \Big\}$$
end for

[1]The intuition is that for any sanitizer within a small neighborhood of $S^{(k)}$, the corresponding optimal public event and private event detectors can be well approximated by $(P^{(k)}, B^{(k)})$. Therefore, if we make an incremental improvement as in the second step of iteration, this incremental update will approximate a stochastic gradient descent step for (2).

3. EXPERIMENTS

To evaluate SEGANE, we use two image datasets: MNIST (handwritten digits) dataset [13] and IMDB (gender and age) dataset [14]. Algorithm 1 is implemented using the PyTorch deep learning platform. We set $\lambda = 1$ and $m = 64$ for both datasets. We use $d = 5$ and $K = 15$ for MNIST while $d = 8$ and $K = 25$ for IMDB. In general, d should be sufficiently large to allow Algorithm 1 to converge to a local optimum. However, for these two datasets, we find that good empirical results can be obtained for small values of d. When using small values of d, Algorithm 1 can be viewed as a type of coordinate descent algorithms that have been known to produce good results in many settings.

To evaluate a trained sanitizer $S(\cdot)$, we implement public and private detectors as DNNs that are trained separately using the sanitized training instances $\{(S(x^i), p^i, b^i), i = 1, \ldots, N\}$. Presumably, these detectors act as ideal detectors for detecting public and private events. Therefore, the performance of the sanitizer can be characterized by the area under the ROC curves (AUC) resulted from these public and private detectors. A better sanitizer would have a larger AUC for the public detector and a smaller AUC for the private detector.

MNIST dataset: The original MNIST dataset is a handwritten digit dataset consisting of 60,000 training examples and 10,000 testing examples. Each example is a 28×28 grayscale image. We create a new synthetic dataset where each synthetic image is a two-digit image (ranging from 00 to 19) generated by concatenating two handwritten images into one with 28×56 pixels. We use 50,000 synthetic images for training and the remaining 5,000 were used for testing. A private event is defined as the event that the two-digit number in the synthetic image is greater than or equal to 10 and a public event is defined as the event that the two-digit number in the image is even. The sanitizer is designed to map a 28×56 input image to $S(x) \in R^{100}$. Fig. 3 depicts the ROC curves for the public and private detectors trained based on the sanitized training dataset. As seen, the AUC is close to 1 for public detector and near 0.5 for the private detector. This indicates that the sanitizer produces sanitized features that allow the public events to be mined effectively while the private events cannot be inferred, i.e., the private detector performs like a random guess. To compare the effectiveness of the proposed sanitizer to retain the data features that allow for accurate classification of a subsequent classifier, Fig. 4 shows the two ROC curves for public detectors operating on raw and sanitized data. As seen, the two curves are almost identical, demonstrating that the proposed sanitizer is very effective, i.e., the sanitized data is as useful as the raw data with regard to the accuracy of public event classification.

IMDB dataset: IMDB consists of 460,723 facial images with gender and age labels. We created a clean dataset consists of 55,000 64×64 images of single frontal faces. We use 50,000 images for training and 5,000 images for testing. We define the public event as the event that the person in an image is male, and a private event as the event that the age of the person is greater than or equal to 25. The sanitizer maps an 64×64 input image to $S(x) \in R^{256}$. As seen in Figs. 5, the ROC for the public detector is quite good while the ROC for the private detector is slightly better than that of random guess. We now let the gender to be the private event while age to be the public event. Fig. 6 shows that the AUC for the public detector is not as large as that in Fig. 5. This is because classifying ages is more difficult than gender. On the other hand, Fig. 6 shows that the performance of the private detector is similar to a random guess. Similar to Fig. 4, Fig. 7 shows the two ROCs of public detectors performing on the raw and santized data. As seen, they are almost

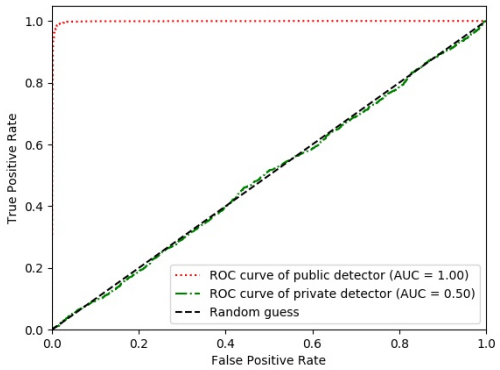

Fig. 3. MNIST experiment: ROC curves for public and private detectors trained with the sanitized dataset

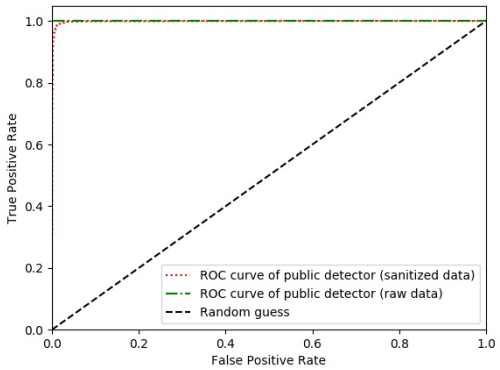

Fig. 4. MNIST experiment: ROC curves for public detectors trained with the raw dataset and the sanitized dataset

identical, indicating that the performance of good subsequent public event classifier on the sanitized data is as good as it is used on the original raw data.

4. CONCLUSION

In this paper, we have proposed SEGANE, a novel DNN architecture for optimizing the performance of machine learning applications with multiple competing objectives. Specifically, SEGANE is evaluated in the context of data sanitization which aims to remove any pre-specified private information from the data in real time while keeping the relevant information used to improve the inference accuracy about the non-private information. The experimental results on two datasets MNIST and IMDB show that SEGANE is highly effective.

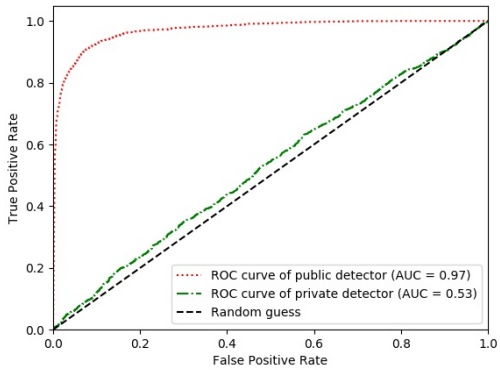

Fig. 5. IMDB experiment: ROC curves for public and private detectors trained with the sanitized dataset (public event is gender, and private event is age)

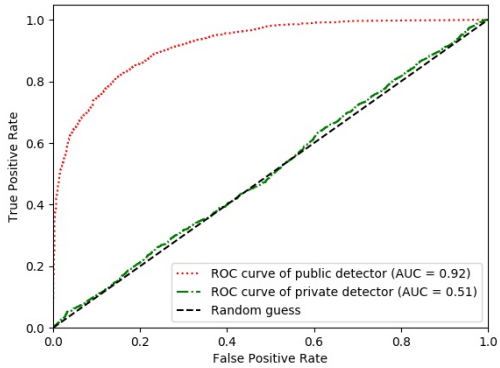

Fig. 6. IMDB experiment: ROC curves for public and private detectors trained with the sanitized dataset (public event is age, and private event is gender)

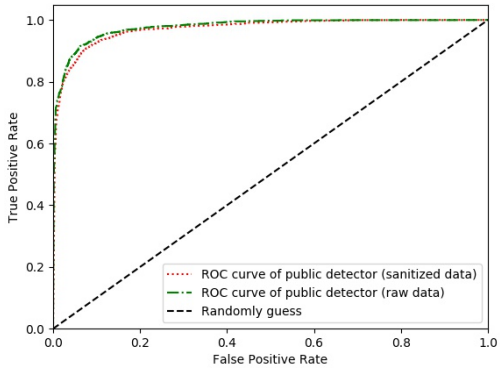

Fig. 7. IMDB experiment: ROC curves for public detectors trained with the raw dataset and the sanitized dataset (public event is gender, and private event is age)

5. REFERENCES

[1] Christof Paar and Jan Pelzl, "Introduction to cryptography and data security," in *Understanding Cryptography*, pp. 1–27. Springer, 2010.

[2] Stan Matwin, "Privacy-preserving data mining techniques: survey and challenges," in *Discrimination and Privacy in the Information Society*, pp. 209–221. Springer, 2013.

[3] Ricardo Mendes and João P Vilela, "Privacy-preserving data mining: Methods, metrics, and applications," *IEEE Access*, vol. 5, pp. 10562–10582, 2017.

[4] Aleksandra Korolova, "Privacy violations using microtargeted ads: A case study," in *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*. IEEE, 2010, pp. 474–482.

[5] Ismini Psychoula, Erinc Merdivan, Deepika Singh, Liming Chen, Feng Chen, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist, "A deep learning approach for privacy preservation in assisted living," *arXiv preprint arXiv:1802.09359*, 2018.

[6] Jianxin Zhao, Richard Mortier, Jon Crowcroft, and Liang Wang, "Privacy-preserving machine learning based data analytics on edge devices," 2018.

[7] Anand D Sarwate and Kamalika Chaudhuri, "Signal processing and machine learning with differential privacy: Algorithms and challenges for continuous data," *IEEE signal processing magazine*, vol. 30, no. 5, pp. 86–94, 2013.

[8] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft, "Learning in a large function space: Privacy-preserving mechanisms for svm learning," *arXiv preprint arXiv:0911.5708*, 2009.

[9] Kamalika Chaudhuri, Anand D Sarwate, and Kaushik Sinha, "A near-optimal algorithm for differentially-private principal components," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2905–2943, 2013.

[10] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 308–318.

[11] Martin J Osborne and Ariel Rubinstein, *A course in game theory*, MIT press, 1994.

[12] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio, *Deep learning*, vol. 1, MIT press Cambridge, 2016.

[13] Yann LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[14] Rasmus Rothe, Radu Timofte, and Luc Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 10–15.