Privacy Leakage in GAN Enabled Load Profile Synthesis

Jiaqi Huang, and Chenye Wu†
School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong 518172 China,
and the Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, Guangdong 518129 China
jiaqihuang@link.cuhk.edu.cn, chenyewu@yeah.net

***Abstract*—Load profile synthesis is a commonly used technique for preserving smart meter data privacy. Recent efforts have successfully integrated advanced generative models, such as the Generative Adversarial Networks (GAN), to synthesize high-quality load profiles. Such methods are becoming increasingly popular for conducting privacy-preserving load data analytics. It is commonly believed that performing analyses on synthetic data can ensure certain privacy.**

In this paper, we examine this common belief. Specifically, we reveal the privacy leakage issue in load profile synthesis enabled by GAN. We first point out that the synthesis process cannot provide any provable privacy guarantee, highlighting that directly conducting load data analytics based on such data is extremely dangerous. The sample re-appearance risk is then presented under different volumes of training data, which indicates that the original load data could be directly leaked by GAN without any intentional effort from adversaries. Furthermore, we discuss potential approaches that might address this privacy leakage issue.

***Index Terms*—Privacy; Data Synthesis; GAN; Differential Privacy; Load Profiling**

I. Introduction

Smart meters collect user load profiles, which enable a variety of data-driven applications such as demand side management, customer behavior analysis, and load forecasting. Unfortunately, the sensitive information in the load data raises public concerns over privacy and security. For example, using some prior knowledge of appliances' power signature, adversaries may apply None-Intrusive Load Monitoring (NILM) [1] techniques to disambiguate the load data, knowing each appliance's status at every moment. Moreover, it has been shown that even without any prior knowledge, it is possible to extract complex usage patterns from load profiles using off-the-shelf statistical methods [2], revealing detailed appliance information of users. At the same time, from the load data, an adversary can immediately know the household occupancy and may use it to infer habits or life patterns of households or even plan for burglaries. Such privacy risks harm users' willingness to share their load profiles, impeding the digitization of the power grid.

†C. Wu is the corresponding author. This work was supported by the National Natural Science Foundation of China (Grant No. 72271213), the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

A. Opportunities and Challenges

To address the privacy issue, a common technique is to generate synthetic load profiles, thereby avoiding the direct use of the original data. Generative models can be classified into white-box models and black-box models. Building a white-box model requires explicit rules to generate load data from sketch, requiring tremendous efforts in making assumptions about the data attributes. Even so, data generated in this manner may be insufficient to reflect the real data's characteristics accurately. In addition, privacy disclosure may occur during the preliminary modeling stage. Black-box models, on the other hand, generate synthetic data without the need for extensive preliminary analyses. These models can generalize well and produce high-quality synthetic data after training. Because black-box models are highly complex and their internal workings are unknown, it is commonly believed that carrying out analyses on the synthetic dataset generated by black-box models can protect certain privacy. However, no previous research has proved an explicit privacy guarantee of this scheme. It remains understudied whether these synthetic data will reveal the private information of users, which motivates our study in this paper.

B. Related Works

To address the privacy issue, many recent studies have investigated the use of Generative Adversarial Networks (GAN) to generate synthetic load data. The early effort from Fekri *et al.* [3] integrates a recurrent GAN (R-GAN) to generate load data. They demonstrate the data quality by the high performance of machine learning models trained on this data. Gu *et al.* further incorporate the auxiliary classifier GAN (ACGAN) to generate load profiles under typical load patterns [4]. Experimental results show that the generated load profiles have good diversity and similarity. Zhang *et al.* [5] use a conditional GAN to learn distribution from the real dataset and generate samples accordingly. The synthetic datasets they generate have a low maximum mean discrepancy compared to real datasets. Despite the success in providing high-quality load data, none of these studies have proved a privacy guarantee for the generated samples. To the best of our knowledge, only Wang *et al.* mention this issue in their work [6] that GANs serve the function of anonymizing smart meter data, thereby protecting certain privacy. On the contrary, in this study we examine

the boundaries of utilizing GAN in terms of smart meter data privacy protection.

C. Our Contributions

In this paper, we study the privacy risk of the synthetic load profiles generated by GAN. Our main contributions can be summarized as follows:

- *Analysis on GAN's risk:* GAN is a complex model with a large parameter space, but no provable privacy guarantee. We analyze why this makes GAN a threat to data privacy.
- *Revealing Sample Re-appearance Risk:* We present the sample re-appearance risk of GAN, showing that the output samples may directly recover the training data.
- *Discussion on Potential Solutions:* We discuss some potential solutions to the privacy issue of load profile synthesis and present some preliminary findings.

The remaining of this paper is structured as follows. In Section II we review some necessary concepts on GAN. Section III provides a detailed analysis of the privacy problems of GAN enabled and load profile synthesis. The sample re-appearance risk is presented in Section IV, and the potential solutions are discussed in Section V. We conclude this study in Section VI.

II. Preliminaries

A. Generative Adversarial Networks (GAN)

GAN is a generative model paradigm that has gained popularity for producing highly realistic images that do not exist in the real world [7]. In the past few years, GAN algorithms have been consistently showing their capabilities of generating high-quality synthetic data, resulting in varieties of applications (see [8] for a comprehensive survey).

GAN is composed of two deep neural networks, the Discriminator Network (DN) and the Generative Network (GN). These two networks compete against each other during training. The goal of GN is to generate realistic samples from random noise in order to mimic the training data, whereas DN is a classifier that distinguishes the generated samples from the real ones. The training procedure of GAN can be viewed as a zero-sum game played by GN and DN with value function $V(G, D)$:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{\text{data}}\,(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))], \quad (1)$$

where G and D are functions representing GN and DN, respectively. During training, DN and GN are trained alternately using gradient descent. They both use $V(G, D)$ as their loss function, but GN aims to minimize this loss while DN aims to maximize it. At each DN's step, DN is shown both real samples and fake samples (generated by current GN) with corresponding real or fake labels and trained to distinguish the two classes of samples. At each GN's step, all parameters of DN are fixed. GN is then trained to maximize $log(D(G(z)))$, where the back-propagation of GN's gradients goes through DN's parameters. Although GN does not see any real images during the training process, it can generate realistic samples only from random vectors after training.

III. Privacy Risk Analysis

In this section, we analyze the privacy leakage risks associated with models which have a lot of parameters but cannot satisfy any kind of privacy guarantee.

A. Lacking Theoretical Privacy Guarantees

In general, data privacy is a concept that states that users' personal information cannot be leaked or cause harm to the user in any way. However, it is hard to explicitly define what is information leakage and what is not, so it has always been challenging to give a rigorous definition of data privacy. Existing privacy definitions and standards are often derived from known potential risks. For example, differential privacy [9], a golden privacy standard, asks that it should be impossible for adversaries to infer whether or not any specific person is involved in the dataset. In other words, when an algorithm doesn't satisfy differential privacy, then it is very likely to leak the participation information of users.

None of the existing GAN-based load profile synthesis approaches can provide a provable privacy guarantee. Most works only achieve a preliminary level of anonymization by removing the labels of training data. However, in privacy research, data anonymization typically refers to a strict definition that the dataset cannot contain any information that is personally identifiable, and there exist explicit standards of data anonymization such as k-anonymity. If the data is only literally anonymized (i.e., having no label), it can be easily de-anonymized when the adversaries have some prior knowledge, shown in real-world cases [10]. In smart grids, for example, adversaries may use some unique features or signatures of appliances to identify specific users from load data. They can therefore infer other sensitive information such as the users' habits or lifestyles, causing privacy disclosure.

B. Memorization Leads to Privacy Disclosure

GAN is a very complex black-box model with a huge parameter space, which seems unlikely for adversaries to recover any training sample, providing some sense of security. However, the large parameter space may unintentionally record sensitive information from training samples. After all, the most secure way to protect privacy is not to use sensitive data at all. Once the data is used, especially when some information derived from the data is recorded or presented, the data owner suffers from the risk of being harmed. The more information recorded, the greater the risk might be.

Because deep networks have a huge amount of parameters, such a large capacity (i.e., parameter space) allows the models to memorize training samples during training [11]. Even when the training data is completely made up of random noises with random labels, which means there is no feature to learn at all, a neural network is still capable of achieving zero loss, indicating that the network has memorized all of them [12]. This remarkable memorizing ability enables adversaries to

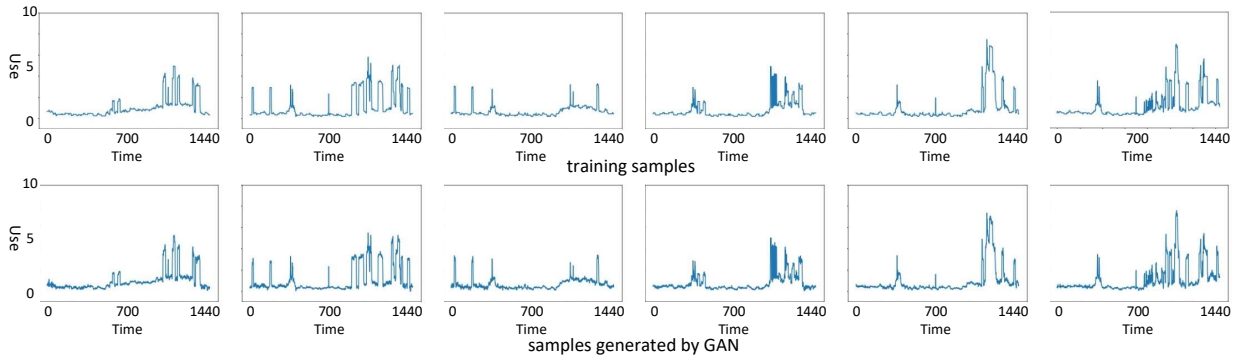

Fig. 1: The sample re-appearance phenomenon of GAN, when the training dataset has 6 samples. (The generated samples displayed above is manually ordered to align the nearest training sample (in l_2-distance) for each.)

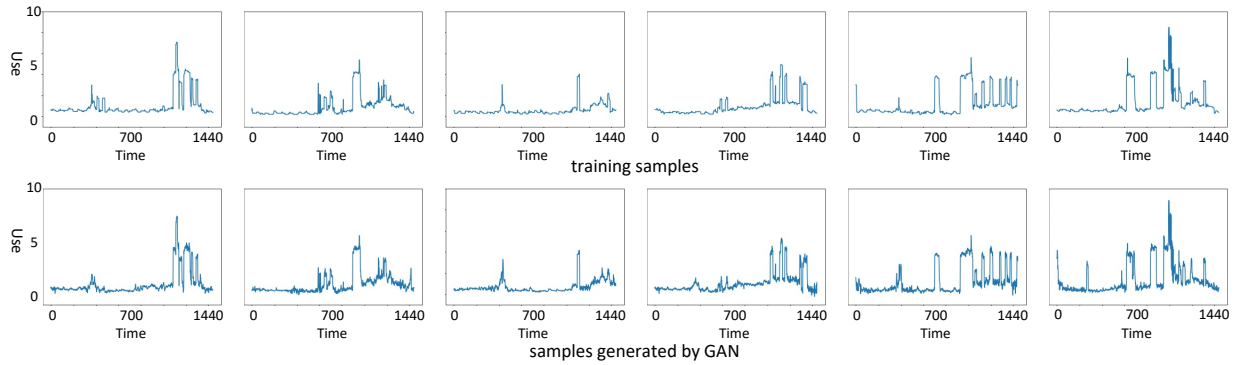

Fig. 2: The sample re-appearance phenomenon of GAN, when the training dataset has 60 samples. (The generated samples displayed above is manually ordered to align the nearest training sample (in l_2-distance) for each.)

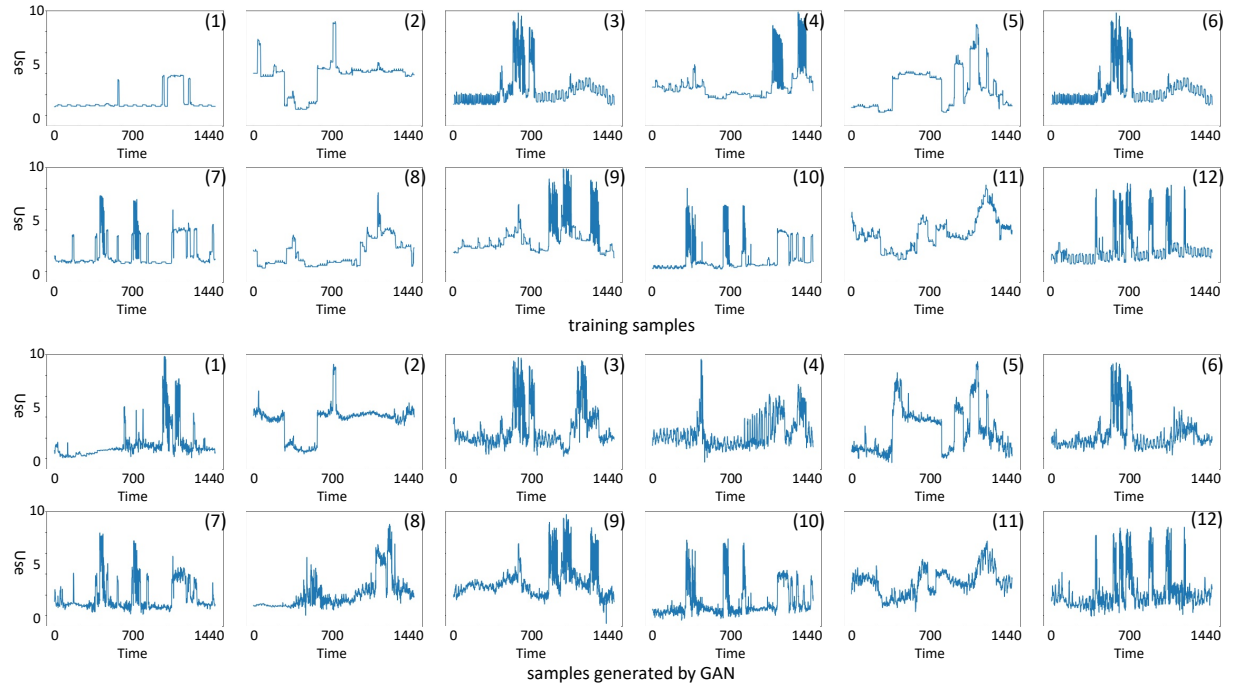

Fig. 3: The generated samples of GAN, when the training dataset has 300 samples.

recover information of the training samples, e.g., inferring the participation of specific user [13] (i.e., membership inference attacks).

Such realistic risks demonstrate that complex black-box models may not prevent the training samples from being recovered from the model parameters. The risks are not completely unexpected, as the attacks often have formulable objectives, so the attacks can be easily conducted by gradient computations when adversaries know the network parameters. Moreover, explicit risks are proven to exist even when adversaries know nothing about the model [14]. In the next section, we further show how users' load profiles are leaked automatically without any intentional design or computation.

IV. Empirical Evidence for the Leakage

In this section, we present empirical results showing that GAN may leak users' load profiles from the training dataset simply and directly.

A. Experimental Settings

We conduct our experiments on the Pecan Street Dataset [15]. In our experiments, each user's data are divided into samples by days, so all samples are of equal length 1440. Each 1440-length sample is treated as the data from a specific user. The model we use is a Wasserstein GAN (WGAN) [16] with 2D convolution layers, where each 1440-length sample is reshaped to 60×24 before entering the network. WGAN is used to improve the quality of generation, and we test LSTM layers, 1D convolution layers, and 2D convolution layers, finding that 2D convolution layers have the best performance.

B. Re-appearance of Samples

We train the WGAN on a tiny dataset of only six samples. After training, we obtain some output samples from GN and observe that each generated sample is nearly identical to a sample in the training dataset. We align the generated samples with the training samples, as shown in Figure 1. Such a phenomenon demonstrates that the model memorized all of the training samples and can leak these data directly without being intentionally attacked. We refer to this phenomenon as the re-appearance risk of GAN, where the harm of leaking the original data has been discussed in Section III.

To simulate more practical cases, we also train the WGAN on datasets with 60 or 300 samples, respectively. We observe that for most of the generated samples, we can still find very similar training samples. We pick the re-appearance samples when the dataset has 60 samples and show them in Figure 2.

It becomes particularly interesting in the 300-sample case. To present in detail, we pick more and not only re-appearing samples and number them as in Figure 3. As before, we present the nearest training sample of each generated sample. We can see that the sample re-appearance phenomenon still happens at generated samples (2), (5), (6), (7), (9), (10), (11), and (12). Meanwhile, the patterns of generated samples (1), (4), and (8) are distinctly different from the training samples. In addition, it seems that recombination happens sometimes. For generated sample (3), its left-hand side is exactly the same as the training sample (3), but its right-hand side seems like a pattern obtained from another training sample.

V. Potential Solutions

In this section, we discuss some potential solutions that might address the privacy risk of load profile synthesis.

A. Insights from the Sample Re-appearance Risk

The experimental results presented in the previous section can give us some insights on mitigating the GAN's privacy risk. In our experiment, we find that if the volume of training data is too small (i.e., 6-sample case), then every output sample from the network will be identical to one of the training samples. When the training data volume gets larger, the generated samples become not always identical to the training samples. This indicates that a larger group of training samples may promote the model's generalization and reduce memorizing specific samples.

Some simple tricks may also be helpful. For example, we can prohibit the output of a sample if its l_2-distance is too close to a training sample. However, such tricks are still unable to provide any guarantee of privacy. To obtain a strong sense of security, we can integrate rigorous privacy standards into our approach, as discussed in the following.

B. Differentially Private Deep Learning

As a *de-facto* standard of privacy, differential privacy has been incorporated into deep learning to provide privacy guarantees solid for the training samples, known as private learning [17]. To conduct private learning, we can inject some specific amount of Gaussian noise into the gradients at each iteration of training. The injected noise is large enough to overwhelm each sample's gradients but at the same time being not that large to change the direction of the aggregated gradients. Therefore, it can protect each sample's privacy while minimizing its impact on the training process. The level of differential privacy protection is represented by a privacy parameter ε, which can be calculated through the noise amplitude.

To study how differential privacy can prevent GAN from memorizing the training samples, we train a GAN using the differentially private stochastic gradient descent (DPSGD) algorithm presented in [17] so that it satisfies differential privacy. We use the same setting as the sample re-appearance experiment with 6 training samples, and see if it can reproduce the re-appearance phenomenon.

The experimental results are presented in Figure 4, which shows even when $\varepsilon = 200,000$, it is still a long way from reproducing the re-appearance phenomenon. An acceptable reproduction is achieved when ε arrives 4,000,000. As a reference, the privacy parameter ε is usually required to be less than 1.0 in order to provide meaningful privacy guarantees, and the protection is often regarded as meaningless when the value of ε exceeds 10. From this point of view, we can confirm that private learning does well in mitigating the sample re-appearance phenomenon. After all, differential privacy is a

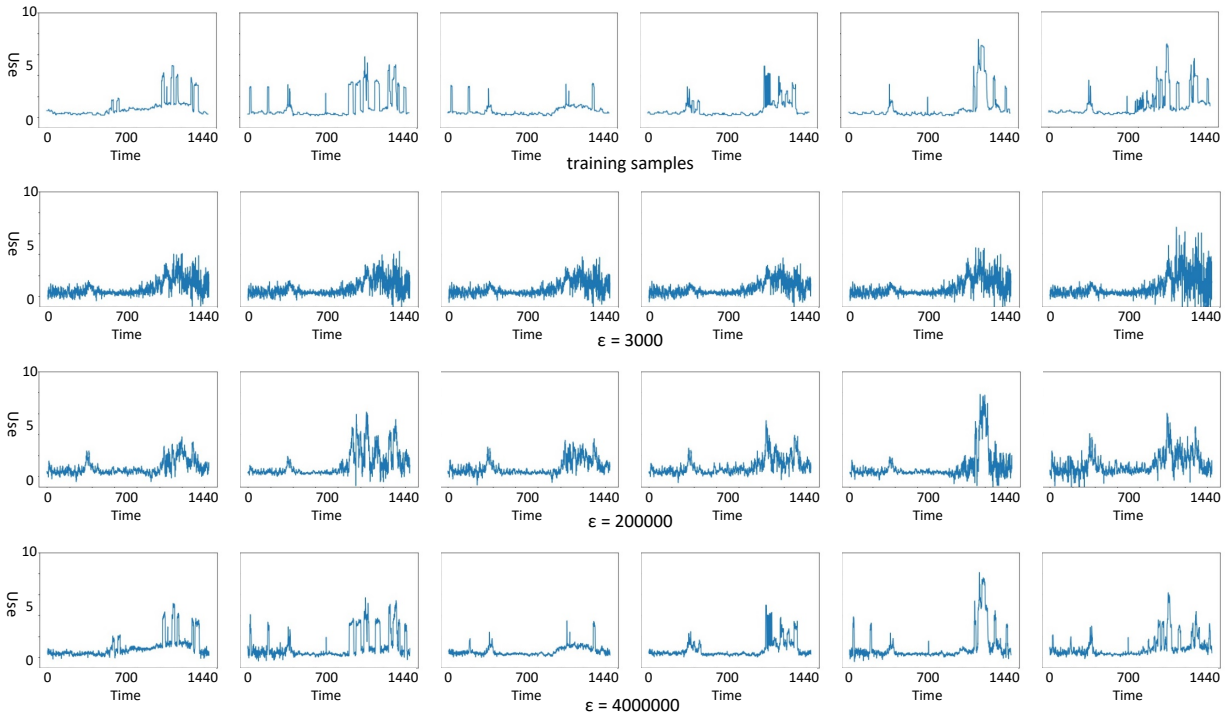

Fig. 4: The memorization ability of GAN is largely weakened when trained with differential privacy.

rigorous standard that prevents adversaries from even knowing the participation of data. The privacy risk of load profile synthesis will be largely reduced if we can develop a differentially private GAN with a meaningful privacy parameter ε.

VI. Conclusion

This paper takes GAN as a representative example to study the privacy leakage problem of load profile synthesis. We show the sample re-appearance risk of GAN when trained with different volumes of training samples, indicating that a complex black-box model may leak the original data in a simple and direct way. We hope this finding will encourage the community to focus on developing data generative methods that can provide rigorous privacy guarantees in future research.

References

[1] G. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, 1992.

[2] A. Molina-Markham, P. J. Shenoy, K. Fu, E. Cecchet, and D. E. Irwin, "Private memoirs of a smart meter," in *Prof. of ACM BuildSys '10*, 2010.

[3] M. N. Fekri, A. M. Ghosh, and K. Grolinger, "Generating energy data for machine learning with recurrent generative adversarial networks," *Energies*, 2019.

[4] Y. Gu, Q. Chen, K. Liu, L. Xie, and C. Kang, "Gan-based model for residential load generation considering typical consumption patterns," in *Proc. of IEEE PES ISGT*, 2019, pp. 1–5.

[5] C. Zhang, S. R. Kuppannagari, R. Kannan, and V. K. Prasanna, "Generative adversarial network for synthetic time series data generation in smart grids," in *Proc. of IEEE SmartGridComm*, 2018, pp. 1–6.

[6] Z. Wang and T. Hong, "Generating realistic building electrical load profiles through the generative adversarial network (gan)," *Energy and Buildings*, vol. 224, p. 110299, 2020.

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27. Curran Associates, Inc., 2014.

[8] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *ArXiv*, vol. abs/2001.06937, 2021.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography*, S. Halevi and T. Rabin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 265–284.

[10] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 2008, pp. 111–125.

[11] D. Arpit, S. Jastrzundefinedbski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," in *Proc. of ICML*. JMLR.org, 2017, p. 233–242.

[12] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Commun. ACM*, vol. 64, no. 3, p. 107–115, feb 2021.

[13] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18.

[14] A. Sablayrolles, M. Douze, Y. Ollivier, C. Schmid, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," 2019. [Online]. Available: https://arxiv.org/abs/1908.11229

[15] Pecan Street. (2012) Pecan street energy research. https://www.pecanstreet.org/. [Online]. Available: https://www.pecanstreet.org/

[16] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 70. PMLR, 06–11 Aug 2017, pp. 214–223.

[17] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2016.