On the Protection of Private Information in Machine Learning Systems: Two Recent Approaches

(Invited Paper)

Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, H. Brendan McMahan,
Ilya Mironov, Nicolas Papernot*, Kunal Talwar, and Li Zhang
Google

Abstract—The recent, remarkable growth of machine learning has led to intense interest in the privacy of the data on which machine learning relies, and to new techniques for preserving privacy. However, older ideas about privacy may well remain valid and useful. This note reviews two recent works on privacy in the light of the wisdom of some of the early literature, in particular the principles distilled by Saltzer and Schroeder in the 1970s.

1. Introduction

In their classic tutorial paper, Saltzer and Schroeder described the mechanics of protecting information in computer systems, as it was understood in the mid 1970s [1]. They were interested, in particular, in mechanisms for achieving privacy, which they defined as follows:

> The term "privacy" denotes a socially defined ability of an individual (or organization) to determine whether, when, and to whom personal (or organizational) information is to be released.

Saltzer and Schroeder took "security" to refer to the body of techniques for controlling the use or modification of computers or information. In this sense, security is an essential element of guaranteeing privacy. Their definitions are roughly in line with our current ideas, perhaps because they helped shaped those ideas. (In contrast, other early definitions took "security" to refer to the handling of classified information, more narrowly [2].)

Although some of that early literature may appear obsolete (for example, when it give statistics on computer abuse [3, Table 1]), it makes insightful points that remain valid today. In particular, Ware noted that "one cannot exploit the good will of users as part of a privacy system's design" [2]. Further, in their discussion of techniques for ensuring privacy [3, p. 9], Turn and Ware mentioned randomized response—several decades before the RAPPOR system [4] made randomized response commonplace on the Web.

The goal of this note is to review two recent works on privacy [5], [6] in the light of the wisdom of some of the early literature. Those two works concern the difficult

* Nicolas Papernot is an intern at Google; he is a student at Pennsylvania State University.

problem of guaranteeing privacy properties for the training data of machine learning systems. They are particularly motivated by contemporary deep learning [7], [8]. They employ two very different techniques: noisy stochastic gradient descent (noisy SGD) and private aggregation of teacher ensembles (PATE). However, they both rely on a rigorous definition of privacy, namely differential privacy [9], [10]. The growing body of research on machine learning with privacy includes several techniques and results related to those we review (e.g., [11], [12], [13], [14], [15], [16], [17], [18]), and undoubtedly neither noisy SGD nor PATE constitutes the last word on the subject; surveying this research is beyond the scope of this note.

Saltzer and Schroeder observed that no complete method existed for avoiding flaws in general-purpose systems. Despite much progress, both in attacks and in defenses, their observation is still correct; in particular, neither the concept of differential privacy nor the sophisticated techniques for achieving differential privacy are a panacea in this respect. Indeed, they may even present opportunities for new flaws [19].

As a mitigation, Saltzer and Schroeder identified several useful principles for the construction of secure systems. In this note, we discuss how those principles apply (or fail to apply) to modern systems that aim to protect private information. Specifically, we review those two recent works on protecting the privacy of training data for machine learning systems, and comment on their suitability by the standards of those principles. Accordingly, our title echoes Saltzer and Schroeder's, "The Protection of Information in Computer Systems".

In the next section, we describe the problem of interest in more detail. In Sections 3 and 4 we review the works on noisy SGD and PATE, respectively. In Section 5 we discuss these works, referring to Saltzer and Schroeder's principles. We conclude in Section 6.

2. The Problem

Next, in Section 2.1, we describe the problem on which we focus, at a high level. In Section 2.2, we describe related problems that are not addressed by the techniques that we discuss in the remainder of the note.

2.1. Framing

We are broadly interested in supervised learning of classification tasks. A classification task is simply a function f from examples to classes, for instance from images of digits to the corresponding integers. (In a more refined definition, f may assign a probability to each class for each example; we omit these probabilities below.) The learning of such a task means finding another function g, called a model, that approximates f well by some metric. Once the model g is picked, applying it to inputs is called inference. The learning is supervised when it is based on a collection of known input-output pairs (possibly with some errors); this collection is the training data.

Since this training data may be sensitive, its protection is an obvious concern, but the corresponding threat model is somewhat less obvious. Attacks may have at least two distinct goals, illustrated by the work of Fredrikson et al. [20] and that of Shokri et al. [21], respectively:

- the extraction of training data (total or partial) from a model g, or
- testing whether an input-output pair, or simply an input or an output, is part of the training data.

We wish to prevent both. The definition of differential privacy gives bounds on the probability that two datasets can be distinguished, thus rigorously addressing membership tests. The extraction of training data seems a little difficult to characterize formally; intuitively, however, it appears harder than membership tests. (Shokri et al. also discuss a weak form of model inversion that appears incomparable with membership tests.) Therefore, we focus on membership tests, and rely on the definition of differential privacy.

Moreover, at least two kinds of threats are worth considering; we call them "black-box" and "white-box", respectively:

- "black-box": attackers can apply the model g to new inputs of their choice, possibly up to some number of times or under other restrictions, or
- "white-box": attackers can inspect the internals of the model g.

"White-box" threats subsume "black-box" threats (in other words, they are more severe), since attackers with access to the internals of a model can trivially apply the model to inputs of their choice. "Black-box" threats may be the most common, but full-blown "white-box" threats are not always unrealistic, in particular if models are deployed on devices that attackers control. Finally, the definition of "white-box" threats is simpler than that of "black-box" threats, since it does not refer to restrictions. For these reasons, we focus on "white-box" threats.

Attackers may also act during the learning process, for example tampering with some of the training data, or reading intermediate states of the learning system. Noisy SGD and PATE are rather resilient to those attacker capabilities, which we do not consider in detail for simplicity.

The term "protection" sometimes refers, specifically, to protection from programs [22]. When we discuss the protection of training data, we may use the term with a broader, informal meaning. This distinction is unimportant for our purposes.

2.2. Other Problems

Related problems pertain to other learning tasks, to other data, and to other aspects of machine learning systems.

2.2.1. Privacy in Other Learning Tasks. There is more to machine learning than the supervised learning of classification tasks. In particular, generative models, which aim to generate samples from a distribution, are another important province of machine learning.

Thinking about privacy has not progressed at an even pace across all areas of machine learning. However, we may hope that some core ideas and techniques may be broadly applicable. For example, many learning techniques employ SGD-like iterative techniques, and noisy variants of those may be able to guarantee privacy properties.

2.2.2. Privacy for Inference Inputs. This note focuses on training data, rather than on the inputs that machine learning systems receive after they are trained and deployed, at inference time. Cryptographic techniques, such as such as CryptoNets [23], can protect those inputs.

Technically, from a privacy perspective, training time and inference time are different in several ways. In particular, the ability of machine learning systems to memorize data [24] is a concern at training time but not at inference time. Similarly, it may be desirable for a machine learning system to accumulate training data and to use the resulting models for some time, within the bounds of data-retention policies, while the system may not need to store the inputs that it receives for inference.

Finally, users's concerns about privacy may be rather different with regards to training and inference. While in many applications users see a direct, immediate benefit at inference time, the connection between their providing data and a benefit to them or to society (such as improved service) is less evident at training time. Accordingly, some users may well be comfortable submitting their data for inference but not contributing it for training purposes.

Nevertheless, some common concerns about privacy focus on inference rather than training. The question "what can Big Brother infer about me?" pertains to inference. It is outside the scope of the techniques that we discuss below.

2.2.3. A Systems Perspectives. Beyond the core of machine learning algorithms, privacy may depend on other aspects of the handling of training data in machine learning systems, throughout the data's life cycle:

- measures for sanitizing the data, such as anonymization, pseudonymization, aggregation, generalization, and the stripping of outliers, when the data is collected;

- traditional access controls, for the raw data after its collection and perhaps also for derived data and the resulting machine learning models; and
- finally, policies for data retention and mechanisms for data deletion.

In practice, a holistic view of the handling of private information is essential for providing meaningful end-to-end protection.

3. Noisy SGD

Many machine learning techniques rely on parametric functions as models. Such a parametric function g takes as input a parameter θ and an example x and outputs a class $g(\theta, x)$. For instance, θ may be the collection of weights and biases of a deep neural network [7], [8]. With each g and θ one associates a loss $L(g, \theta)$, a value that quantifies the cost of any discrepancies between the model's prediction $g(\theta, x)$ and the true value $f(x)$, over all examples x. The loss over the true distribution of examples x is approximated by the loss over the examples in the training data and, for those, one takes $f(x)$ to be as given by the training data, even though this training data may occasionally be incorrect. Training the model g is the process of searching for a value of θ with the smallest loss $L(g, \theta)$, or with a tolerably small loss—global minima are seldom guaranteed. After training, θ is fixed, and new examples can be submitted. Inference consists in applying g for a fixed value of θ.

Often, both the model g and the loss L are differentiable functions of θ. Therefore, training often relies on gradient descent. With SGD, one repeatedly picks an example x (or a mini-batch of such examples), calculates $g(\theta, x)$ and the corresponding loss for the current value of θ, and adjusts θ in order to reduce the loss by going in the opposite direction of the gradient. The magnitude of the adjustment depends on the chosen learning rate.

The addition of noise is a common technique for achieving privacy (e.g., [3]), and also a common technique in deep learning (e.g., [25]), but for privacy purposes the noise should be carefully calibrated [9]. The sensitivity of the final value of θ to the elements of the training data is generally hard to analyze. On the other hand, since the training data affects θ only via the gradient computations, we may achieve privacy by bounding gradients (by clipping) and by adding noise to those computations. This idea has been developed in several algorithms and systems (e.g., [14], [15], [16]). Noisy SGD, as defined in [5], is a recent embodiment of this idea, with several modifications and extensions, in particular in the accounting of the privacy loss.

4. PATE

The use of ensembles of models is common in machine learning [26]. If an ensemble comprises a large enough number of models, and each of the models is trained with a disjoint subset of the training data, we may reason, informally, that any predictions made by most of the models should not be based on any particular piece of the training data. In this sense, the aggregation of the models should protect privacy with respect to "black-box" threats.

Still, since the internals of each of the models in an ensemble is derived from the training data, their exposure could compromise privacy with respect to "white-box" threats. In order to overcome this difficulty, we may treat the ensemble as a set of "teachers" for a new "student" model. The "student" relies on the "teachers" only via their prediction capabilities, without access to their internals.

Training the "student" relies on querying the "teachers" about unlabelled examples. These should be disjoint from the training data whose privacy we wish to protect. It is therefore required that such unlabelled examples be available, or that they can be easily constructed. The queries should make efficient use of the "teachers", in order to minimize the privacy cost of these queries. Once the "student" is fully trained, however, the "teachers" (and any secrets they keep) can be discarded.

PATE and its variant PATE-G are based on this strategy. They belong in a line of work on knowledge aggregation and transfer for privacy [27], [28], [17]. Within this line of work, there is some diversity in goals and in specific techniques for knowledge aggregation and transfer. PATE aims to be flexible on the kinds of models it supports, and is applicable, in particular, to deep neural networks. It relies on noisy plurality for aggregation; the noise makes it possible to derive differential-privacy guarantees. The variant PATE-G relies on generative, semi-supervised methods for the knowledge transfer; currently, techniques based on generative adversarial networks (GANs) [29], [30] are yielding the best results. They lead, in particular, to state-of-the-art privacy/utility trade-offs on MNIST and SVHN benchmarks.

5. Principles, Revisited

In this section, we consider how the principles distilled by Saltzer and Schroeder apply to noisy SGD and to PATE. Those principles have occasionally been analyzed, revised, and extended over the years (e.g., [31], [32], [33]). We mainly refer to the principles in their original form.

Most of the protection mechanisms that Saltzer and Schroeder discussed are rather different from ours in that they do not involve data transformations. However, they also described protection by encryption (a "current research direction" when they wrote their paper); the training of a model from data is loosely analogous to applying a cryptographic transformation to the data. This analogy supports the view that noisy SGD and PATE should be within the scope of their principles.

Economy of mechanism. This principle says that the design of protection mechanisms should be kept as simple and small as possible.

Neither noisy SGD nor PATE (and its variants) seem to excel in this respect, though for somewhat different reasons:

- Although noisy SGD relies on simple algorithmic ideas that can be implemented concisely, these ideas

directly affect SGD, which is so central to many learning algorithms. Therefore, applying noisy SGD is akin to performing open-heart surgery. The protection mechanism is not a stand-alone system component. Moreover, optimizations and extensions of learning algorithms (for example, the introduction of techniques such as batch normalization [34]) may require new methods or new analysis.

- PATE involves somewhat more design details than noisy SGD. In particular, PATE-G incorporates sophisticated techniques based on GANs. On the other hand, those design details are entirely separate from the training of the "teacher" models, and independent of the internal structure of the "student" model.

It remains to be seen whether radically simpler and smaller mechanisms can be found for the same purpose.

Fail-safe defaults. This principle means that lack of access is the default. In particular, mistakes should result in refusing permission.

This principle, which is easy to interpret for traditional reference monitors, appears difficult to apply to noisy SGD, PATE, and other techniques with similar goals, which generally grant the same level of access to anyone making a request under any circumstances.

We note, however, that many of these techniques achieve a guarantee known as (ϵ, δ)-differential-privacy [35]; while the parameter ϵ from the original definition of differential privacy can be viewed as a privacy cost [9], the additional parameter δ can be viewed as a probability of failure. Prima facie, such a failure results in a loss of privacy, rather than a loss of accuracy. In this sense, the guarantee does not imply fail-safe defaults. A more refined analysis paints a subtle and interesting picture with a trade-off between the parameters ϵ and δ [36].

Complete mediation. This principle implies that every access to sensitive data should go through the protection mechanism.

Resistance to "white-box" threats means that the internals of models are not sensitive, so concerns about complete mediation should not apply to them, but these concerns still apply to the raw training data at rest or in transit. Complete mediation requires a system-wide perspective (see Section 2.2.3).

Open design. This principle, which echoes one of Kerckhoffs's [37], states that the designs of protection mechanisms should not depend on secrecy, and should not be kept secret.

Both noisy SGD and PATE are satisfactory in this respect. This property may seem trivial until one notes that not all current work on privacy (and, in particular, on differential privacy) is equally open.

Separation of privilege. This principle calls for the use of multiple independent "keys" for unlocking access.

Like the principle of fail-safe defaults, it appears difficult to apply to noisy SGD and to PATE. It may perhaps apply to a separate, outer level of protection.

Least privilege. This principle reads "Every program and every user of the system should operate using the least set of privileges necessary to complete the job".

This principle seems more pertinent to the implementations of noisy SGD and PATE than to their high-level designs. For instance, in the case of PATE, it implies that each of the "teacher" models should be configured in such a way that it would not have access to the training data of the others, even if its software has flaws.

The principle has the virtue of limiting the damage that may be caused by an accident or error. We simply do not have enough experience with noisy SGD and PATE to characterize the nature and frequency of accidents and errors, but it seems prudent to admit that they are possible, and to act accordingly.

Least common mechanism. This principle addresses the difficulty of providing mechanisms shared by more than one user. Those mechanisms may introduce unintended communication channels. Moreover, it may be hard to satisfy all users with any one mechanism.

The principle can be regarded as an end-to-end argument [31], since it suggests that shared mechanisms should not attempt that which can be achieved by each user separately. While each user could ensure the differential privacy of their data by adding noise, as in RAPPOR [4], the required levels of noise can sometimes conflict with utility. Therefore, shared mechanisms that achieve differential privacy are attractive.

In techniques such as noisy SGD and PATE, the privacy parameters (in particular the parameters ϵ and δ discussed above) are the same for all pieces of the training data, and for all accesses to the learning machinery. The addition of weights [38] could perhaps accommodate the privacy requirements of different pieces of training data, and thus those of different users.

Psychological acceptability. This principle advocates ease of use, and was later called the principle of least astonishment [31]. Saltzer and Schroeder noted: "to the extent that the user's mental image of his protection goals matches the mechanisms he must use, mistakes will be minimized".

This remark by Saltzer and Schroeder was in fact one of the starting points for the work on PATE. While noisy SGD provides mathematical guarantees, understanding them requires a fair amount of sophistication in differential privacy and in machine learning, which many users and operators of systems will lack. In contrast, the way in which PATE guarantees privacy should be intuitively clear, at least at a high level. No advanced background is required in order to accept that, if 100 independently trained machine learning models say that a picture is an image of a cat, then this prediction is probably independent of any particular picture in their disjoint sets of training data.

Work factor. This principle calls for comparing the resources of attackers with the cost of circumventing the protection mechanism.

The "white-box" threat model is helpful in simplifying work-factor considerations. The privacy guarantees for noisy SGD and PATE benefit from the fact that attackers need not be limited to any particular number of queries at inference time.

Compromise recording. The final principle suggests that it is advantageous to be able to detect and report any failures of protection.

As noted above, (ϵ, δ)-differential-privacy includes the possibility of failures. In our setting, it allows for the possibility that noisy SGD and PATE plainly reveal one or a few pieces of the training data that they should safeguard. No detection or reporting of such failures has been contemplated. The seriousness of this problem seems open to debate, as it may be a shortcoming of the theory, rather than of the algorithms.

6. Conclusion

The current, vibrant research on privacy is developing sophisticated concepts and techniques that apply, in particular, to cutting-edge machine learning systems. Noisy SGD and PATE, which this note reviews, aim to contribute to this line of work. Looking beyond the core algorithms, it is important to understand how these algorithms fit into systems and society. Economy of mechanism, psychological acceptability, and other fundamental principles should continue to inform the design and analysis of the machinery that aims to protect privacy.

Acknowledgments

References

[1] J. H. Saltzer and M. D. Schroeder, "The protection of information in computer systems," *Proceedings of the IEEE*, vol. 63, no. 9, pp. 1278–1308, 1975. [Online]. Available: http://dx.doi.org/10.1109/PROC.1975.9939

[2] W. H. Ware, "Security and privacy: Similarities and differences," in *Proceedings of the April 18-20, 1967, Spring Joint Computer Conference*, ser. AFIPS '67 (Spring). ACM, 1967, pp. 287–290. [Online]. Available: http://doi.acm.org/10.1145/1465482.1465525

[3] R. Turn and W. H. Ware, "Privacy and security in computer systems," Jan. 1975. [Online]. Available: https://www.rand.org/content/dam/rand/pubs/papers/2008/P5361.pdf

[4] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: randomized aggregatable privacy-preserving ordinal response," in *Proceedings of the 21st ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2014, pp. 1054–1067. [Online]. Available: http://doi.acm.org/10.1145/2660267.2660348

[5] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318. [Online]. Available: http://doi.acm.org/10.1145/2976749.2978318

[6] N. Papernot, M. Abadi, Ú. Erlingsson, I. J. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *CoRR*, vol. abs/1610.05755, 2016, presented at the 5th International Conference on Learning Representations, 2017. [Online]. Available: http://arxiv.org/abs/1610.05755

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[9] C. Dwork, F. McSherry, K. Nissim, and A. D. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006*, 2006, pp. 265–284. [Online]. Available: http://dx.doi.org/10.1007/11681878_14

[10] C. Dwork, "A firm foundation for private data analysis," *Commun. ACM*, vol. 54, no. 1, pp. 86–95, Jan. 2011. [Online]. Available: http://doi.acm.org/10.1145/1866739.1866758

[11] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. D. Smith, "What can we learn privately?" *SIAM J. Comput.*, vol. 40, no. 3, pp. 793–826, 2011. [Online]. Available: http://dx.doi.org/10.1137/090756090

[12] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate, "Differentially private empirical risk minimization," *J. Machine Learning Research*, vol. 12, pp. 1069–1109, 2011.

[13] D. Kifer, A. D. Smith, and A. Thakurta, "Private convex optimization for empirical risk minimization with applications to high-dimensional regression," in *Proceedings of the 25th Annual Conference on Learning Theory*, 2012, pp. 25.1–25.40. [Online]. Available: http://www.jmlr.org/proceedings/papers/v23/kifer12/kifer12.pdf

[14] S. Song, K. Chaudhuri, and A. Sarwate, "Stochastic gradient descent with differentially private updates," in *GlobalSIP Conference*, 2013.

[15] R. Bassily, A. D. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Proceedings of the 55th IEEE Annual Symposium on Foundations of Computer Science*. IEEE, 2014, pp. 464–473. [Online]. Available: http://dx.doi.org/10.1109/FOCS.2014.56

[16] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1310–1321. [Online]. Available: http://doi.acm.org/10.1145/2810103.2813687

[17] J. Hamm, Y. Cao, and M. Belkin, "Learning privately from multiparty data," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, 2016, pp. 555–563. [Online]. Available: http://jmlr.org/proceedings/papers/v48/hamm16.html

[18] X. Wu, A. Kumar, K. Chaudhuri, S. Jha, and J. F. Naughton, "Differentially private stochastic gradient descent for in-RDBMS analytics," *CoRR*, vol. abs/1606.04722, 2016. [Online]. Available: http://arxiv.org/abs/1606.04722

[19] I. Mironov, "On significance of the least significant bits for differential privacy," in *Proceedings of the 19th ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2012, pp. 650–661. [Online]. Available: http://doi.acm.org/10.1145/2382196.2382264

[20] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2015, pp. 1322–1333. [Online]. Available: http://doi.acm.org/10.1145/2810103.2813677

[21] R. Shokri, M. Stronati, and V. Shmatikov, "Membership inference attacks against machine learning models," *CoRR*, vol. abs/1610.05820, 2016. [Online]. Available: http://arxiv.org/abs/1610.05820

[22] B. W. Lampson, "Protection," *Operating Systems Review*, vol. 8, no. 1, pp. 18–24, 1974. [Online]. Available: http://doi.acm.org/10.1145/775265.775268

[23] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. E. Lauter, M. Naehrig, and J. Wernsing, "CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016*, 2016, pp. 201–210. [Online]. Available: http://jmlr.org/proceedings/papers/v48/gilad-bachrach16.html

[24] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *CoRR*, vol. abs/1611.03530, 2016, presented at the 5th International Conference on Learning Representations, 2017. [Online]. Available: http://arxiv.org/abs/1611.03530

[25] A. Neelakantan, L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens, "Adding gradient noise improves learning for very deep networks," *CoRR*, vol. abs/1511.06807, 2015. [Online]. Available: http://arxiv.org/abs/1511.06807

[26] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[27] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*. ACM, 2007, pp. 75–84.

[28] M. Pathak, S. Rane, and B. Raj, "Multiparty differential privacy via aggregation of locally trained classifiers," in *Advances in Neural Information Processing Systems*, 2010, pp. 1876–1884.

[29] 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*, 2014, pp. 2672–2680.

[30] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," *arXiv preprint arXiv:1606.03498*, 2016.

[31] J. H. Saltzer and M. F. Kaashoek, *Principles of Computer System Design: An Introduction*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2009.

[32] S. L. Garfinkel, "Design principles and patterns for computer systems that are simultaneously secure and usable," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 2005.

[33] R. Smith, "A contemporary look at Saltzer and Schroeder's 1975 design principles," *IEEE Security and Privacy*, vol. 10, no. 6, pp. 20–25, Nov. 2012. [Online]. Available: http://dx.doi.org/10.1109/MSP.2012.85

[34] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, 2015, pp. 448–456. [Online]. Available: http://jmlr.org/proceedings/papers/v37/ioffe15.html

[35] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *EUROCRYPT*. Springer, 2006, pp. 486–503.

[36] I. Mironov, "Renyi differential privacy," *CoRR*, vol. abs/1702.07476, 2017. [Online]. Available: http://arxiv.org/abs/1702.07476

[37] A. Kerckhoffs, "La cryptographie militaire," *Journal des sciences militaires*, vol. IX, pp. 5–38, Jan. 1883. [Online]. Available: http://www.petitcolas.net/kerckhoffs/crypto_militaire_1.pdf

[38] D. Proserpio, S. Goldberg, and F. McSherry, "Calibrating data to sensitivity in private data analysis: A platform for differentially-private analysis of weighted datasets," *Proc. VLDB Endow.*, vol. 7, no. 8, pp. 637–648, Apr. 2014. [Online]. Available: http://dx.doi.org/10.14778/2732296.2732300