Private Retrieval, Computing, and Learning: Recent Progress and Future Challenges

Sennur Ulukus, *Fellow, IEEE*, Salman Avestimehr, *Fellow, IEEE*, Michael Gastpar, *Fellow, IEEE*, Syed A. Jafar, *Fellow, IEEE*, Ravi Tandon, *Senior Member, IEEE*, and Chao Tian, *Senior Member, IEEE*

***Abstract*— Most of our lives are conducted in the cyberspace. The human notion of privacy translates into a cyber notion of privacy on many functions that take place in the cyberspace. This article focuses on three such functions: how to privately retrieve information from cyberspace (privacy in information retrieval), how to privately leverage large-scale distributed/parallel processing (privacy in distributed computing), and how to learn/train machine learning models from private data spread across multiple users (privacy in distributed (federated) learning). The article motivates each privacy setting, describes the problem formulation, summarizes breakthrough results in the history of each problem, and gives recent results and discusses some of the major ideas that emerged in each field. In addition, the cross-cutting techniques and interconnections between the three topics are discussed along with a set of open problems and challenges.**

***Index Terms*— Private information retrieval, private distributed computing, private distributed learning, federated learning.**

I. Introduction

PRIVACY is an important part of human life. This article considers privacy in the context of three distinct but related engineering applications, namely, privacy in retrieving information, privacy in computing functions, and privacy in learning. In the first sub-topic of private information retrieval, a user wishes to download a content from publicly accessible databases in such a way that the databases do not learn which particular content the user has downloaded. Towards that goal, the user creates ambiguity by its actions during the download. This strategy prevents databases from guessing which content the user has downloaded. This in turn preserves the user's privacy because what is downloaded leaks information about interest and intent on the part of the user. In the second sub-topic of private computing, a user wishes to compute a function but does not have resources to perform the computation on its own. Thus, the user outsources the computation to many distributed servers. This necessitates the user send its data, which is private, to the distributed servers. The goal of the user is to utilize the servers for computation while preserving the privacy of its own data. To achieve that goal, the user introduces randomness in its data so that the servers cannot decipher the data while they are able to perform the computation successfully. In the third sub-topic of privacy in learning, a centralized unit (parameter server) wishes to train a learning model by utilizing distributed users (clients). The parameter server needs labeled training data to train the model. The data resides at the users, and the users prefer to keep their data at their site, i.e., not send it to the parameter server, to preserve the privacy of their data. Thus, such distributed (federated) learning has built-in privacy advantages. However, even then, the computations (e.g., gradients calculated on the data) may leak some information about the raw data. To prevent that, the users may want to add randomness to the calculation they send to the parameter server in order to further preserve their privacy.

The underlying threat model common to all three settings is the undesired leak of information that is considered private by the respective entities. In the case of private information retrieval, the leak is about the identity (index) of the content being downloaded/accessed. In the case of private computation, the leak is about the user data on which computation needs to be performed by distributed servers. In the case of private learning, the leak is user (client) data that is used to train the learning model. A common aspect of the solution approach to these problems is to randomize the information/actions in such a way to hide the private information. In private information retrieval, this corresponds to randomizing the downloads such that a certain download may happen equally likely for all possible user content requirements. In private computation, randomization is achieved by adding appropriate noise to the data whose effect can be nullified during the computation. In private learning, privacy of clients is achieved by keeping the data at the client side, and also my randomizing the transmitted calculations so that leaks are prevented.

Manuscript received June 10, 2021; revised November 1, 2021; accepted December 21, 2021. Date of publication January 12, 2022; date of current version February 17, 2022. The work of Sennur Ulukus was supported in part by the Army Research Office (ARO) under Grant W911NF2010142, and in part by the National Science Foundation (NSF) under Grant CCF 17-13977 and Grant ECCS 18-07348. The work of Ravi Tandon was supported by NSF under Grant CNS 17-15947, Grant CAREER 16-51492, and Grant CCF 21-00013. The work of Chao Tian was supported in part by the National Science Foundation under Grant CCF-2007067. *(Corresponding author: Sennur Ulukus.)*

Sennur Ulukus is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA (e-mail: ulukus@umd.edu).

Salman Avestimehr is with the Electrical and Computer Engineering Department, University of Southern California, Los Angeles, CA 90007 USA (e-mail: avestimehr@ee.usc.edu).

Michael Gastpar is with the School of Computer and Communication Sciences, EPFL, 1015 Lausanne, Switzerland (e-mail: michael.gastpar@epfl.ch).

Syed A. Jafar is with the Department of Electrical Engineering and Computer Science, University of California at Irvine, Irvine, CA 92697 USA (e-mail: syed@ece.uci.edu).

Ravi Tandon is with the Department of Electrical and Computer Engineering, The University of Arizona, Tucson, AZ 85721 USA (e-mail: tandonr@email.arizona.edu).

Chao Tian is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77842 USA (e-mail:chao.tian@tamu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2022.3142358.

Digital Object Identifier 10.1109/JSAC.2022.3142358

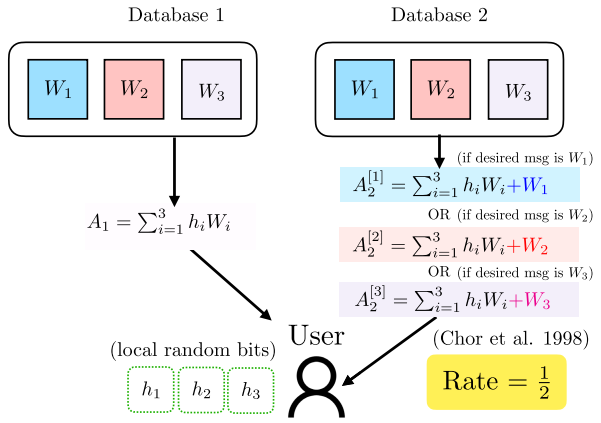

Fig. 1. The PIR scheme of Chor *et al.* [1] for $N = 2$ databases which achieves a rate of $1/2$ for any number of messages K (for the case shown in the figure, i.e., for $K = 3$ messages, the capacity is $4/7$ [4]). The main idea is to use the fact that since the databases cannot collude, one can send correlated queries across databases, allowing the user to leverage information retrieved across databases to increase the rate (download efficiency).

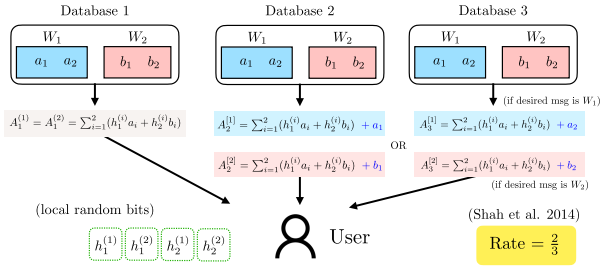

Fig. 2. The PIR scheme of Shah *et al.* [9] for $N = 3$ databases and $K = 2$ messages which achieves a rate of $2/3$ (the capacity for this setting is $3/4$ [4]). More generally, the scheme achieves a rate of $1 - 1/N$, irrespective of K. The new ingredient beyond Chor *et al.* [1] involved message sub-packetization.

We present private information retrieval in Section II, private distributed computation in Section III and private distributed machine learning in Section IV. We conclude this article in Section Section V by listing a few challenges and open problems.

II. Private Information Retrieval

The private information retrieval (PIR) problem was introduced by Chor *et al.* [1] as a privacy-preserving primitive for retrieving information in a private manner. In the canonical PIR setting, a user wishes to retrieve one of K available messages, from N non-communicating servers, each of which has a copy of these K messages. User request privacy needs to be preserved during the retrieval process, i.e., the identity of the desired message remains unknown to any single server. A generic protocol to retrieve, e.g., message k, is as follows in this setting:

1) The user generates N queries with a private random key and the message index k, one query per server, which are sent to the respective servers;
2) Each server, based on the query it receives and the content it stores, sends back an answer to the user;
3) The user collects the answers from N databases, and reconstructs the message based on the answers, the private key, and the requested message index k.

The privacy requirement can be either information-theoretic (e.g., [2]) or computational (e.g., [3]). The former requires that each server cannot infer any information on the identity of the requested message, even if assuming the server has infinite computation power; in contrast, the latter assumes that each server has only limited computation power, and under such computational constraint, it is required that the server cannot learn anything on the identity of the requested message. In this article, we only consider information-theoretic privacy.

Since the introduction of the PIR problem by Chor *et al.*, tremendous advances have been made using the computer science theoretic approach, including on the canonical form and many variations; see the survey article [5] and references therein. Within the context of this approach, the effort usually focuses on the scaling behavior of the communication costs, including both the query communication (upload) cost and the answer communication (download) cost, with respect to N and K. Moreover, the messages are usually assumed to be very short, the most common of which is in fact a single bit per message. There have been many variations on the canonical setting, and it has been recognized that the PIR problem has deep connections to other coding or security primitives, such as locally decodable codes and oblivious transfer [6]–[8].

It was shown in Chor *et al.* that for a single database, perfect information-theoretic privacy can only be achieved by downloading the entire database (of K messages), i.e., the optimal rate defined as the ratio of the amount of desired information (one message) and the total downloaded amount (K messages) in this case is $1/K$. Thus, a natural question that was first explored by Chor et. al is the following: can one achieve a better rate than $1/K$ by exploiting $N > 1$ databases? This question was answered in the affirmative, and it was shown that even with $N = 2$ databases, one can achieve a rate of $1/2$ for any number of messages. We explain the main idea through a simple example as shown in Fig. 1 for $K = 3$ messages. The main idea behind the scheme is as follows: assuming each message W_k is one-bit, the user generates K random bits $\{h_1, \ldots, h_K\}$ and requests the linear combination (denoted by $\sum_{k=1}^{K} h_k W_k$) of the K messages from database 1, whereas requests $\sum_{k=1}^{K} h_k W_k + W_\theta$ from the other database whenever the user wants to retrieve W_θ. Since $\{h_k\}$'s are uniformly generated bits, the distribution of $\sum_{k=1}^{K} h_k W_k + W_\theta$ is identical for every $\theta \in \{1, 2, \ldots, K\}$, ensuring perfect privacy.

The PIR problem was recently reintroduced to the information theory and coding community [9]–[11], with initial effort focused on using advanced coding technique to improve the storage, upload, and download efficiency. Specifically, Shah *et al.* [9] generalized the Chor *et al.* scheme to any arbitrary number $(N > 1)$ of databases. The key new idea herein was to subpacketize each message into $(N - 1)$ parts, and then follow an approach similar to Chor *et al.* In Fig. 2, we highlight this through an example when $N = 3$ and for

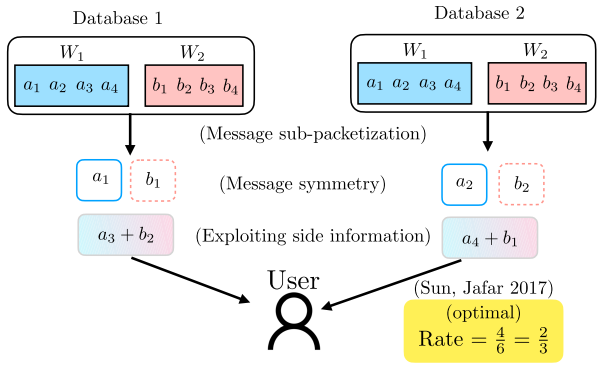

Fig. 3. The capacity-achieving PIR scheme of Sun and Jafar [4] for $N = 2$ databases and $K = 2$ messages which achieves a rate of $2/3$. The optimal scheme requires a combination of the following ideas: a) message sub-packetization, b) maintaining message symmetry for privacy and c) fully exploiting side information (from other databases) at each database.

$K = 2$ messages. Each message is partitioned into $(N-1) = 2$ parts, following which the user then requests a random linear combination (A_1 from database 1), followed by requesting A_1 XORed with the $(N-1)$ subpackets individually from the remaining $(N-1)$ databases. This leads to a rate of $\frac{N-1}{N} = 1 - \frac{1}{N}$, which is independent of K.

A significant milestone of this renewed effort on the PIR problem is the result obtained in [4], where the optimal download cost of the PIR capacity of the canonical setting was fully characterized. A key new ingredient that leads to this breakthrough is the information-theoretic reformulation of the problem. In contrast to typical computer science theoretical formulation, here the number of bits in each message is allowed to approach infinity, and the capacity is defined as the supremum of the number of useful message bits that can be retrieved per total downloaded bits.

A code construction was provided which relies on a few key code design principles. The scheme by Sun and Jafar for $N = 2$ databases and $K = 2$ messages is illustrated in Fig. 3. The key design principles behind the scheme include the following: a) sub-packetization of each message into $L = N^K$ symbols, followed by b) downloading parts of each message from every database (i.e., maintaining *message symmetry*) for maintaining privacy of the desired message index, and c) fully exploiting side information at every database (from remaining $(N-1)$ databases). For the example in Fig. 3, this amounts to breaking the two messages into $L = 4$ symbols. If the user wants to download message $W_1 = (a_1, a_2, a_3, a_4)$, it downloads one symbol from each message from both databases (namely, (a_1, b_1) from database 1 and (a_2, b_2) from database 2). Subsequently it downloads $a_3 + b_2$ from database 1 and $a_4 + b_1$ from database 2 (i.e., the remaining desired symbols together with the undesired symbols downloaded from the other database). A matching converse is proved using the conventional information theoretic approach.

The surprising result inspired many subsequent works using such a capacity formulation and led to many new discoveries which will be surveyed in this part of the article.

A. The Canonical PIR System

For the canonical PIR setting, shown in Fig. 4, a rigorous computer science theoretic problem definition of the problem was given in the seminal paper [1], and a more explicit information theoretic translation was given in [12]. The breakthrough work of Sun and Jafar [4] instead directly represented the coding function relations using information measure relations, which we explain next.

The random query $Q_n^{[k]}$ intended for server-n when requesting message k is determined by the private random key F, i.e., $H(Q_n^{[k]}|\mathsf{F}) = 0$, for $n = 1, 2, \ldots, N$, $k = 1, 2, \ldots, K$. The answers $A_n^{[k]}$ from server-n, in response to the query $Q_n^{[k]}$, is determined by the stored messages and the query, i.e., $H(A_n^{[k]}|W_{1:K}, Q_n^{[k]}) = 0$, for $n = 1, 2, \ldots, N$, $k = 1, 2, \ldots, K$. Given the above, there are two key constraints/requirements from a PIR scheme:

1) *Decodability Constraint:* The reconstructed message $\hat{W}_k$, by the user for the requested message k, is determined from the answers $A_{1:N}^{[k]}$ and the random key F, i.e.,

$$H(\hat{W}_k|A_{1:N}^{[k]}, \mathsf{F}) = 0, \text{ for } k \in \{1, 2, \ldots, K\}. \quad (1)$$

2) *Privacy Constraint:* The privacy requirement is that the queries for any message pairs k and k' have an identical distribution

$$\mathbf{Pr}(Q_n^{[k]} = q) = \mathbf{Pr}(Q_n^{[k']} = q), \quad (2)$$

which can be represented as $I(\theta; Q_n^{[\theta]}, A_n^{[\theta]}, W_{1:K}) = 0$, for $n \in \{1, 2, \ldots, N\}$, where θ is the random variable representing the index of the requested message.

In the information-theoretic setting, the download cost D dominates the upload cost. The definition of the download cost D requires some elaboration. Two obvious information-theoretic measures directly related to the download cost are $\sum_{i=1}^{N} H(A_n^{[k]})$ and $\sum_{i=1}^{N} H(A_n^{[k]}|\mathsf{F})$. The latter is a lower bound of the expected number of total download bits, which is how we usually measure the download cost. The former is an upper bound of the latter, and can be viewed as a surrogate, particularly in asymptotic (large number of information bits in each message) settings.

The efficiency of the download is then measured by the number of requested message bits obtained per downloaded bit, which leads to the following capacity notion, when error is not allowed. More precisely, a rate R is said to be achievable for zero-error PIR, if there exists a PIR code of download cost D such that $R = \frac{L}{D}$ with no decoding errors. The supremum of achievable rates for zero-error PIR is called the (zero-error) PIR capacity C_0.

For zero-error PIR code, there is no need to explicitly specify the probability distribution for each message, and also no need to specify the message retrieval probability. However, when a more general capacity notion, the ϵ-error capacity, is adopted, this is no longer the case, since the error probability is not strictly zero. In this case, the convention is to assume that each message is distributed in its range uniformly at random, and the message is also being requested

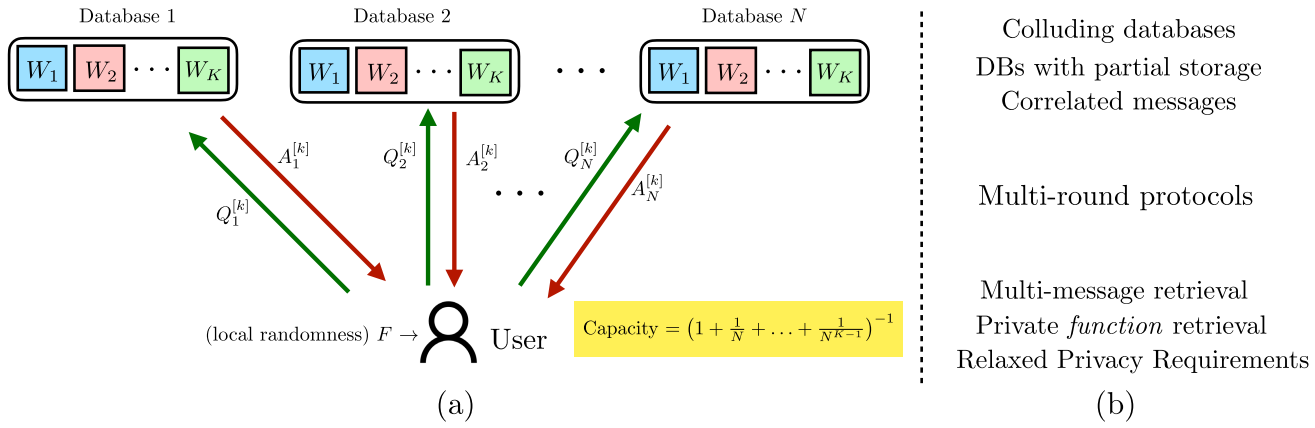

Fig. 4. (a) The canonical PIR system; (b) extensions of the canonical system.

uniformly at random [4]. Correspondingly, a rate R is said to be ϵ-error achievable if there exists a sequence of PIR codes, each of rate greater than or equal to R, for which $\frac{1}{K}\sum_{k=1}^{K} \mathbf{Pr}(\hat{W}_k \neq W_k) \rightarrow 0$ as $L \rightarrow \infty$. The supremum of ϵ-error achievable rates is called the ϵ-error capacity C_ϵ.

The download cost used above is the expected number of downloaded bits, and the capacities are defined accordingly, which is usually the notion adopted in subsequent works. However another slightly different notion of the download cost is the worst-case download cost, which was used in [13] (and later adopted in [14] and [15]). The worst-case download cost is the largest number of downloaded bits over all query combinations that are used with non-zero probability. Using this notion, we can similarly define the zero-error worst-case PIR capacity $\bar{C}_0$, and ϵ-error worst-case PIR capacity $\bar{C}_\epsilon$. The breakthrough work [4] essentially established that

$$C_0 = \bar{C}_0 = C_\epsilon = \bar{C}_\epsilon = \left(1 + \frac{1}{N} + \cdots + \frac{1}{N^{K-1}}\right)^{-1}. \quad (3)$$

The capacity definition C_0 is the most straightforward, and often adopted in the literature for generalized PIR settings. When the problem setting deviates from the canonical setting, it is known that C_ϵ can be different from C_0 in certain cases, but it is not well understood when this is the case. It is not known whether $\bar{C}_0$ and $\bar{C}_\epsilon$ can in fact be different for any generalized PIR systems.

The general code construction for the canonical PIR system turns out to be rather elegant, and plays an important role for subsequent works. The code construction to obtain the capacity result in [4] relies on several important design principles, which are illustrated using the example of $N = K = 2$ case in Fig. 3. This original construction required sub-packetization of each message into N^K parts (alternatively, a message length of $L = N^K$ symbols), followed by invoking the design principles of server/message symmetry and exploiting side information.

An alternative code construction was provided in [16] and [12], which is illustrated in Fig. 5(a). In this construction, the server symmetry and message symmetry are not used on the per retrieval basis, but across all the retrieval patterns. The random key $F \in \{0, 1\}$ is invoked with probability $1/2$ each, and thus the expected retrieval download cost is still the same $3/2$. The advantage of this alternative code construction is that it can be shown to have the minimum message length ($L = 1$ in this example and $L = N - 1$ in general), comparing to the exponential growth of message length for the code given in [4]. A similar code construction was discovered by [17] for special case of $N = 2$, and was later extended to the case of more general number of databases [18] as shown in Fig. 5(b). The difference from that in [16] and [12] is additional layer of symmetrization enforced across the databases. It is clear that both code constructions have the same download cost, however, the upload cost of the former is lower ($\log 2$ vs. $\log 4$). The symmetry structure in the canonical PIR setting is quite sophisticated and plays an important role in constructing efficient code design. The overall symmetry is induced by the database symmetry, the message symmetry, and the (retrieval) variety symmetry; see [12] for a detailed discussion.

B. Relation to Computer Science Theoretic PIR

For the canonical PIR system, the computer science theoretic description of the coding operations is exactly equivalent to the information-theoretic version we just provided. The main difference between them is in terms of the performance measure, i.e., regarding the definition of C_0 and C_ϵ. Since in the computer science theoretic setting the messages are short, usually only one bit each, the upload cost, i.e., $\sum_{i=1}^{N} \log |\mathcal{Q}_n|$, plays an important role in the overall communication cost, and thus the total communication cost must consist of both components. In contrast, in the information-theoretic setting, since the message size is allowed to be very large, the download cost dominates and the upload cost can be essentially ignored, and only the normalized cost is meaningful, whose inverse is the PIR rate.

Since in the computer science theoretic setting, the message length is fixed and not allowed to grow to infinity, it is not meaningful to consider the ratio between the message length and the communication cost. In contrast, in the information-theoretic setting, this ratio between the message length and the download cost is the key metric to consider.

	Requesting a		Requesting b	
	DB 1	DB 2	DB 1	DB 2
$F=0$	0	a	0	b
$F=1$	$a+b$	b	$a+b$	a

(a)

	Requesting a		Requesting b	
	DB 1	DB 2	DB 1	DB 2
$F=0$	0	a	0	b
$F=1$	$a+b$	b	$a+b$	a
$F=2$	a	0	b	0
$F=3$	b	$a+b$	a	$a+b$

(b)

Fig. 5. Low-subpacketization schemes for PIR for $(N, K) = (2, 2)$. The scheme in (a) was given in [16] and [12], and the scheme in (b) in [17] and [18]. One achieves the minimum possible subpacketization for each message while achieving an expected download cost of $3/2$.

It is in fact rather difficult to fully characterize the sum of the upload cost and the download cost (for any fixed message length), and thus the optimal scaling laws are usually sought after in the computer science theoretic setting. In contrast, in the information-theoretic setting, the ratio between the message length and the download cost leads to the concept of capacity, and the problem in fact becomes much more tractable. Instead of the scaling law, the capacity of the PIR system can be fully identified.

C. Extended PIR Systems

The canonical PIR system given in the previous sub-section can be viewed as consisting of four key components: a single-round query-answer protocol, a set of independent messages of the same length, an absolute privacy requirement, and also inherently a star-shape communication network; see Fig. 4. The last item regarding the communication network may require some elaboration, since it is usually not explicitly introduced: the user communicates to each server through a dedicated link, and each server answers through a dedicated link, and as a consequence, the communication costs are measured in a straightforward manner on each link.

Any of these components can be generalized:

1) The query-answer protocol structure. The user sends a single round of queries, and the servers answer in a single round; this protocol can be generalized to allow multiple rounds.
2) The message structure. In more general systems, the messages can be dependent; in a less obvious variation, the user in fact requests a function of the messages.
3) The privacy (or security) requirement. The user may wish to enforce the privacy requirements that even certain subsets of the servers collude, they still will not be able to infer any knowledge on the request. The server may place security constraint that the authors cannot learn about other messages than the one being requested.
4) The communication network structure. This communication models can be generalized in various way, for example, to allow a more complex communication network, or using additional communication module, such as caches, to facilitate the communication. In such general settings, the communication costs are measured in rather different manners.

In the next subsections, we survey various generalizations of the canonical PIR problem.

1) Multi-Round and Multi-Message PIR Systems: Sun and Jafar [19] considered the extended PIR system where the user and the servers are allowed multiple rounds of queries and answers. It was shown that the capacity of multiround PIR is in fact the same as single round PIR, when there is no constraint placed on the storage cost. This equality continues to hold even when T-colluding is allowed. However, when the storage is more constrained, this equality would indeed break. Yao *et al.* [20] considered using multiround communication in the settings with Byzantine databases, and showed that multiround communication is also beneficial in this setting. In multi-message PIR, the user wishes to download multiple messages privately. The question that arises is whether downloading multiple messages one-by-one sequentially is optimum. Reference [21] shows that downloading multiple messages jointly is more efficient and beats the sequential use of single-message PIR. Reference [21] determines the PIR capacity when the number of desired messages is at least half of the total number of messages, while the multi-message PIR capacity in other cases remains open.

2) Cache or Side Information Aided PIR: Cache aided private information retrieval (PIR) (e.g., [22]–[24]) and side information aided PIR (e.g., [25]–[34]) are both interesting extensions of the original information-theoretic PIR problem [4] because they both lead to reduction in download costs due to the fact that, under both settings, the user possesses cache or side information, respectively. The PIR capacity and the corresponding PIR schemes with cache/side-information vary, depending on a) if the databases are aware or unaware of the side-information at the user; b) if the user wishes to only keep the message index private or both the message index and side-information private from the databases; and c) the type of side-information available at the user (e.g., subset

of messages or fraction of some/all messages). Recently, the role of side information is investigated in the context of symmetric PIR (SPIR) where the side information is a subset of shared database common randomness; this work showed that with appropriate amount of user-side side information the capacity of SPIR can be increased to the capacity PIR, and single-database SPIR can be made possible [35].

3) PIR From Databases With Limited Storage: The assumption of fully replicated databases (all N databases storing all K messages) can be unrealistic in practice. However, the amount of redundancy across the databases has an impact on the capacity of PIR. Specifically, on one extreme for replicated databases, the capacity is the highest, whereas on the other extreme, if there is no redundant storage across databases, then the only feasible strategy is to download all K messages. There have been several recent works which have explored the trade-off between the capacity and storage for PIR. The case when each message is encoded by a maximum distance separable (MDS) code and stored across the databases, referred to as the MDS-PIR code, was studied in [36] and [37] and the capacity was settled by Banawan and Ulukus [36]; also see recent results on MDS-PIR with minimum message size [38]. The problem of MDS-PIR with colluding databases turns out to be more challenging and the capacity remains unknown for general parameters; see [39], [40]. Several other variants have been studied including PIR from databases storing data using an arbitrary linear code [41], [42], impact on capacity versus storage when using arbitrary (possibly, non-linear codes) [43]–[47], when databases only store fraction of uncoded messages [48]–[50], and when data is not perfectly replicated across the databases, but rather partially replicated according to graph based structures [51]–[54].

4) PIR Under Additional Abilities and Constraints for the Databases: The original setting in [4] considers privacy against individual databases. In practice, a subset of databases may have the ability to collude; this may happen, for instance, if the databases belong to the same entity. Reference [55] considers the case where up to T out of N databases may collude, and finds the PIR capacity as a function of T. Further, [56] considers the case where in addition to the T colluding databases, up to B databases may exhibit Byzantine behavior, meaning that they can return arbitrarily random or incorrect answers to the queries, and finds the PIR capacity as a function of T and B. In addition, databases may require *database privacy*. This means that the user does not learn anything further than the message it wished to download. The resulting setting is coined as symmetric PIR (SPIR) to emphasize the symmetry of privacy requirements of the user and the databases. The capacity of SPIR is found in [57]. The SPIR capacity is smaller than the PIR capacity, as SPIR is a more constrained problem than PIR. SPIR achievable scheme is similar to the schemes in [1] and [9] but it requires a shared common randomness among the databases. Recent paper [35] explores making some of that common randomness available (randomly) to the user to increase SPIR capacity. Further, SPIR proves to be an important privacy primitive that is a building block in many problems that involve symmetric privacy requirements among participating parties, such as in private set intersection [58], [59].

Databases may be subject to a set of practical limitations due to the way that the databases are accessed or the way they need to return their answers. For instance, if the databases need to return their answers via noisy and/or multiple-access wireless channels, then the PIR schemes should be designed together with channel coding techniques to deal with the uncertainty in the channels as in [60]. In another example setting, if the rate at which the user can download information from the databases is different for each database, then the user access to the databases and the PIR schemes across the databases may need to be asymmetric. This may happen, for instance, if the databases have different distances to the user (with a more distant database having a smaller bit-rate) or if they have different channel qualities (some channels from the databases being in deep fades). In this case, asymmetric access conditions need to be taken into consideration [61]. An interesting observation in [61] is that if the asymmetry is mild, the full unconstrained PIR capacity may still be maintained. Another set of practical constraints arise if the database-to-user channels are being eavesdropped by an external entity. This gives rise to a problem formulation at the intersection of information-theoretic privacy and information-theoretic security [62]–[65]. Yet another practical constraint is that the messages stored at the databases do not have to be of equal length, and their a priori probabilities of retrieval (popularities) do not have to be the same. These give rise to message semantics that need to be taken into account during a PIR code design [66]. An interesting observation in [66] is that if longer messages have higher popularities then the semantic PIR capacity may be larger than classical PIR capacity.

5) Relaxed Privacy Notions: Perfect information-theoretic privacy requirements (either for the user as in PIR or for both the user and the databases as in SPIR) usually come at the expense of high download cost and do not allow tuning the PIR efficiency and privacy according to the application requirements. In applications which may require frequently retrieving messages, trading user or database privacy for communication efficiency could be desirable. Ideally, one would select a desired leakage level and then design a leakage-constrained retrieval scheme that guarantees such privacy while maximizing the download efficiency. Asonov and Freytag introduced the concept of repudiative information retrieval [67]. The repudiation property is achieved if the probability that the desired message index was i given the query is non-zero for every index i, i.e., there is always some remaining uncertainty at the database about the desired message index. Recently, Toledo *et al.* [68] adopted a game-based differential privacy definition to increase the PIR capacity at the expense of bounded privacy loss. With the goal of allowing bounded leakage for the information-theoretic PIR/SPIR formulations (as initiated in [69]), there have been a series of recent works. In [17], the perfect privacy constraint was relaxed by requiring that the log likelihood of the posterior distribution for any two message indices given the query is bounded by ϵ. When $\epsilon = 0$, this recovers perfect privacy, and allows leakage for $\epsilon > 0$.

Full capacity characterization			
PIR [4] Multiround [19] Computation [76], [77]	$C = \Psi(N, K)$	Multiround: allows sequential queries Computation: retrieves arbitrary linear combinations of messages	✓
TPIR [55]	$C = \Psi((N-U)/T, K)$	U unresponsive servers	✓
PIR [78] with arbitrary collusion pattern	$C = \Psi(S^*, K)$	Arbitrary collusion pattern P, $S^* = \max_y 1_N^T y$, s.t. $B_P^T y \leq 1_M, y \geq 0_N$ B_P: incidence matrix of P	✓
Cache-aided PIR [79]	$C = \Psi(N, K)/(1 - S/K)$	PIR aided by local **c**ache at user of size $S\times$ message size	✓
PIR-SI [80], [81]	$C = \Psi(N, \lceil \frac{K}{M+1} \rceil)$	**S**ide **I**nformation of M messages at User, Privacy of SI not required	✓
PIR-PSI [28], [80], [82]	$C = \Psi(N/T, K-M)$	**S**ide **I**nformation of M messages at User, θ and SI jointly T-private	✓
SPIR [57], [83]	$C = (1 - \frac{2B+\max(T,E)}{N}) \cdot \mathbb{I}_{(\rho \geq \frac{2B+\max(T,E)}{N-2B-\max(T,E)})}$	**S**ymmetric security, B-**B**yzantine servers, E-**E**avesdroppers, Common randomness ρ shared among servers	✓
Q-PIR, Q-STPIR [84]	$C = \min\left\{1, \frac{2(N-T)}{N}\right\}$	Quantum PIR, allows symmetric security, Servers share an entangled state	✓
B-TPIR [56]	$C = \left(1 - \frac{2B}{N}\right) \Psi(\frac{T}{N-2B}, K)$, $N > 2B+T$ $C = \frac{1}{(2B+1)K} \mathbb{I}_{(N>2B)}$, $N \leq 2B+T$	B-**B**yzantine servers	✓
MDS-PIR [36]	$C = \Psi(N/K_c, K)$	(N, K_c) MDS Coded Storage	
PIR with limited storage [48]–[50]	$C = \Psi(\mu N, K)$, $\mu = \frac{t}{N}, t \in [N]$	Each server stores no more than μ fraction of database; heterogeneous sizes μ_i; decentralized.	
Partial capacity characterization			
Multimessage PIR [21]	$C = \Psi(N, \frac{K}{M})$ if $K/M \in \mathbb{N}$ $C = \frac{MN}{MN+K-M}$ if $M \geq \frac{K}{2}$	**M**ulti-**m**essage Retrieval Retrieves M out of K messages	✓
MDS-TPIR [39], [40]	$C_l = 1 - (K_c + T - 1)/N$ $C = \frac{N^2-N}{2N^2-3N+T}$ if $K_c = N-1$	(N, K_c) MDS Coded Storage	
XS-TPIR [85]	$C_l = \left(1 - \frac{X+T}{N}\right)^+ = C_\infty$ $C^u = \left(1 - \frac{X}{N}\right) \Psi(\frac{N-X}{T}, K)$ $= C$ if $N \leq X+T$ or $(N, X, T) = (3, 1, 1)$	X-secure storage	
U-B-XS-MDS-TPIR [86], [87]	$C_l = \left(1 - \frac{K_c+X+T+2B-1}{N-U}\right)$ $= C_\infty$ if $K_c = 1, X = 0$	X-secure, $(N, K_c + X)$ MDS coded storage, U-**u**nresponsive, B-**B**yzantine servers	

Fig. 6. A sampling of capacity results for various forms of PIR, where N is the number of servers, K is the number of messages, and T is the privacy parameter with default value being 1. For the rows with a check mark, the storage on server is simple message replication. C represents capacity, C_∞ is the asymptotic capacity for large number of messages ($K \to \infty$), C^u and C_l are upper and lower bounds on C, respectively. $\Psi(A, B) \triangleq (1 + 1/A + 1/A^2 + \cdots + 1/A^{(B-1)})^{-1}$.

Lin *et al.* [70], [71] relaxed user privacy by allowing bounded mutual information between the queries and the corresponding requested message index. Unlike [70], [71], which deal with the average leakage measured by mutual information, the model studied in [17] provides stronger privacy guarantees. Zhou *et al.* [72] measured the leakage using the maximal leakage metric and argued this leakage measure is more applicable. Guo *et al.* [73] considered the problem of SPIR with perfect user privacy and relaxed database privacy. Database privacy was relaxed by allowing a bounded mutual information (no more than δ) between the undesired messages, the queries, and the answers received by the user. Similar to the original work on SPIR in [57], SPIR with relaxed database privacy in [73] requires sharing common randomness among databases and comes at the expense of a loss in the PIR capacity. Asymmetric leaky PIR was explored in [18] where bounded leakage is allowed in both directions. Recently, the model of latent-variable PIR was introduced and studied, where instead

of requiring privacy for the message index, one may require privacy of data correlated with the message [74].

D. Connections to Other Security Primitives

PIR holds particular significance as a point of convergence of complementary perspectives. It is well known that PIR shares intimate connections to prominent problems in theoretical computer science and cryptography, communication and information theory, and coding and signal processing. PIR protocols are often used as essential ingredients of oblivious transfer [86], instance hiding [87]–[89], multiparty computation [90], secret sharing schemes [91], [92] and locally decodable codes [7]. Through the topics of locally decodable, recoverable, repairable and correctable codes [93], PIR connects to distributed data storage repair [94], index coding [95] and the entire umbrella of network coding [96] in general. PIR schemes are essentially interference alignment schemes [97] as the downloads comprise a mix of desired messages with undesired messages (interference). Efficient retrieval requires the alignment of interference dimensions across the downloads from different servers while keeping desired signals resolvable. It is not surprising then that interference alignment has been used implicitly in PIR and index coding long before its applications in wireless networks [97]. Various equivalence results have been established between PIR and blind interference alignment (BIA) [4], [98]; BIA and topological interference management (TIM) [99]; TIM and index coding [99]; index coding and locally repairable codes [100], [101]; locally repairable and locally decodable codes [93]; and between locally decodable codes and PIR [7]. Add to this the equivalence between index coding and network coding [102], [103], storage capacity and index coding [104], index coding and hat guessing [105], or the application of asymptotic interference alignment schemes originally developed for wireless interference networks [106] to distributed storage exact repair [107], and it becomes evident that discoveries in PIR have the potential for a ripple effect in their impact on a number of related problems.

The remainder of this article explores in greater depth the topics of secure distributed computing and private federated learning.

III. Private Distributed Computing

A. Formulation

We consider a general distributed computing framework, where the goal is to compute a function g using N distributed workers, while keeping the input dataset $\boldsymbol{X}$ secure (illustrated in Fig. 7). The N workers are assigned encrypted versions of the input using N encoding functions $\boldsymbol{c} \triangleq (c_1, \ldots, c_N)$, then each worker computes a function f over the assigned share, which can be viewed as building blocks of computing g.

This framework captures many commonly used operations. One example is *block matrix multiplication*, where the goal is to compute the product $A^\intercal B$ given two large matrices $A \in \mathbb{F}^{s\times t}$ and $B \in \mathbb{F}^{s\times r}$. Here the input dataset is $X = (A, B)$, and the computation task is $g(A, B) = A^\intercal B$. Given some partitioning parameters p, m, and n, the input matrices are partitioned block-wise into p-by-m and p-by-n sub-blocks of equal sizes, respectively. Then each worker is assigned a pair of sub-blocks and computes their product, i.e., the function f is the multiplication of two matrices of sizes $\mathbb{F}^{\frac{t}{m}\times\frac{s}{p}}$ and $\mathbb{F}^{\frac{s}{p}\times\frac{r}{n}}$. If there are no security requirements, the final result can be recovered using $N = pmn$ workers, by having each worker compute a product of certain uncoded submatrices.

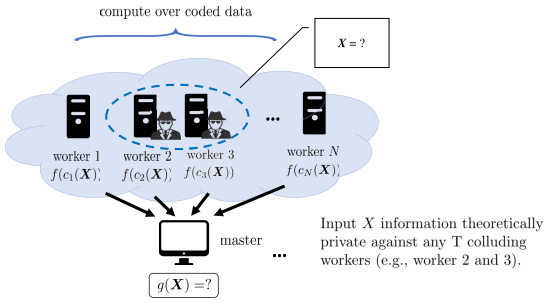

Fig. 7. An illustration of private computation.

Another example is to compute *multivariate polynomials* on a dataset X. Particularly, given a general polynomial f, the input dataset is partitioned into K subsets $X_1, \ldots, X_K$, and the goal is to compute $g(\boldsymbol{X}) = (f(X_1), \ldots, f(X_K))$. If each worker can compute a single evaluation of f, then a computing design using $N = K$ workers can be obtained by assigning each worker a disjoint uncoded subset of the input.

However, in secure computing, we aim to carry out the computation with an additional requirement that the entire input dataset is information-theoretically secure from the workers, even if up to a certain number of them can collude. In particular, a set of encoding functions $\boldsymbol{c} \triangleq (c_1, \ldots, c_N)$ is T-secure, if $I(\{c_i(X)\}_{i\in\mathcal{T}}; X) = 0$ for any subset $\mathcal{T}$ with a size of at most T, where X is generated uniformly at random.

1) Tradeoffs in Secure Distributed Computing: The goal is to design the encoding functions to achieve a tradeoff between the resources/constraints while ensuring the reliable recovery of the desired computation. Specifically, the resources could correspond to the number of available workers, storage and computation performed per worker. In addition to T-security, one may also be interested in communication efficient designs to account for bandwidth constraints (between master node and the workers). In the past few years, there have been significant progress in using ideas from coding/information theory to devise new schemes, and ultimately towards understanding these fundamental tradeoffs. As an example, a large body of work [108]–[112] have focused on the problem of distributed computing in the presence of stragglers. Here, since the overall latency of computation can be limited by the slowest workers, the goal is to design schemes which minimize the number of workers required to carry out the computation, while satisfying the security requirement. More rigorously, let $\mathcal{C}_T$ denotes the set of allowable[1] encoding function designs that are T-secure. Then, in a secure coded computing problem, given fixed parameters f, g, and $\mathcal{C}_T$, one aim could be to find computing

[1]The set $\mathcal{C}_T$ also captures practical constraints such as encoding complexities.

schemes $\boldsymbol{c} \in \mathcal{C}_T$ that use as small number of workers N as possible. Alternatively, when the total number of workers N is fixed, one may be interested in the design of T-secure schemes with minimum download communication overhead. The capacity of secure distributed computation, analogous to that of PIR, can then be defined as the supremum of the ratio of the number of bits of desired information (the desired function, $g(\mathbf{X})$), to the total number of bits downloaded from the N servers.

B. Schemes for Private Distributed Computing

Information-theoretically secure distributed computing has its origins in the celebrated work of Ben-Or Goldwasser Micali (BGW protocol) on tasks involving linear/bilinear computations. Specifically, the master node creates N coded shares with T-secure guarantees (using Shamir's secret sharing scheme) which are subsequently sent to the workers. The workers subsequently compute the function on the coded shares. In a recent work [113], *staircase codes*, presented originally for a PIR problem [114], were combined with the idea of secret sharing to minimize the overall latency for secure distributed matrix multiplication.

Lagrange coded computing (LCC) [115] has been proposed to provide a unified solution for computing general multivariate polynomials. In comparison to the classical BGW (or similar Shamir's secret sharing based protocols), LCC reduces the amount of storage, communication and randomness overhead. Given any fixed parameter T, LCC encodes the input variables using the following *Lagrange interpolation polynomial*

$$c(x) \triangleq \sum_{j\in[K]} X_j \cdot \prod_{k\in[K+T]\setminus\{j\}} \frac{x-x_k}{x_j-x_k} + \sum_{j=K+1}^{K+T} Z_j \cdot \prod_{k\in[K+T]\setminus\{j\}} \frac{x-x_k}{x_j-x_k},$$

where $x_1,\ldots,x_{K+T}$ are some arbitrary distinct elements from the base field $\mathbb{F}$, and Z_i's are some random cryptographic keys generated uniformly[2] at random on the domain of X_i's. Each worker i selects a distinct variable y_i from the base field that is not from $\{x_1,\ldots,x_K\}$, and obtains $\tilde{X}_i \triangleq c(y_i)$ as the coded variable. LCC is T-secure, because the coded variables sent to any subset of T workers are padded by an invertible linear transformation of T random keys, which are jointly uniformly random.

After each worker i applies function f over the coded inputs, they essentially evaluate the composed polynomial $f(c)$ at point y_i. On the other hand, the evaluations of the same polynomial at $x_1,\ldots,x_K$ are exactly the K needed final results. Hence, by polynomial interpolation, the decoder can recover all final results by recovering $f(c)$, by receiving results from any subset of workers with a size greater than the degree of $f(c)$. More precisely, let $\deg f$ denote the total degree of polynomial f, the degree of the composed polynomial equals $(K-1)\deg f$. Thus, LCC computes any multivariate polynomial with at most $N=(K-1)\deg f+1$ workers.

[2]We assume that $\mathbb{F}$ is finite so that the uniform distribution is well defined.

Secure coded computation has also been studied for different computation tasks and settings. A majority of works are on matrix multiplication [116]–[131], and it has been shown in [132] and [131] that for block-partition-based designs, the optimum number of workers to enable secure computation can be within a constant factor of a fundamental quantity called the bilinear complexity [133]. Private gradient computation was studied in [134] and it was shown that the optimal coding design is encoding the input variables using harmonic sequences. References [135] and [136] considered a setting where the workers send compressed versions of their computing results to tradeoff the download communication cost and the required number of workers.

LCC has also been widely leveraged to enable privacy-preserving machine learning [137], [138]. In particular, authors in [137] have considered a scenario in which a data-owner (e.g., a hospital) wishes to train a logistic regression model by offloading the large volume of data (e.g., healthcare records) and computationally-intensive training tasks (e.g., gradient computations) to N machines over a cloud platform, while ensuring that any collusions between T out of N workers do not leak information about the dataset. In this setting, CodedPrivateML [137] has been proposed, which leverages LCC, to provide three salient features:

1) it provides strong information-theoretic privacy guarantees for both the training dataset and model parameters.
2) it enables fast training by distributing the training computation load effectively across several workers.
3) it secret shares the dataset and model parameters using coding and information theory principles, which significantly reduces the training time.

LCC has also been leveraged to break a fundamental "quadratic barrier" for secure model aggregation in federated learning [139]. We defer the discussion on this topic to the next section.

Within the scope of this article, the connection between PIR, secure distributed computing, and private federated learning is exemplified by the idea of cross-subspace alignment (CSA) which extends to all three domains. CSA codes originated in [83] as a solution to XS-TPIR, i.e., the problem of T-private information retrieval from N servers that store K messages in an X-secure fashion. CSA codes then found applications in private secure coded computation [85], [140], [141], and in particular secure distributed matrix multiplication (SDMM) [142]. CSA codes were first applied to SDMM by Kakar *et al.* in [123], and subsequently applied to secure distributed *batch* matrix multiplication (SDBMM) by Jia and Jafar in [143]. These works produced sharp capacity[3] characterizations for various cases. For example, in [143] the capacity for X-secure distributed computation by N servers of a batch of outer products of two vectors is shown to be $(1-X/N)^+$, the capacity for computing the inner product of two length-K vectors is $\frac{1}{K}(1-\frac{X}{N})^+$ when $N\leq 2X$, and for long vectors $(K\to\infty)$ the capacity of computing inner

[3]Analogous to PIR, the capacity of SDBMM is defined as the supremum of the ratio of the number of bits of desired information (the desired matrix products), to the total number of bits downloaded from the N servers.

products is shown to be $(1-2X/N)^+$. While LCC [115] codes and CSA codes originated in seemingly unrelated contexts of distributed secure computing and secure PIR, there are interesting connections between them. For example, in a special case of secure multiparty/distributed batch matrix multiplications, CSA codes yield LCC codes as a special case [135], [144]. The generalization inherent in CSA codes is beneficial primarily in download-limited settings, where CSA codes are able to strictly outperform LCC codes. However, it is worth mentioning that to achieve order-optimal performances, entangled polynomial codes [131], [132] should be applied, which enables coding over bilinear-complexity-based algebraic structures, and achieving order-wise improvements. Finally, in the domain of federated learning, CSA codes were applied in [145] to find a solution to the problem of X-secure T-private federated submodel learning. Fundamentally, this is a problem of privately reading from and privately writing to a database comprising K files (messages/submodels) that are stored across N distributed servers in an X-secure fashion. The CSA read-write scheme of [145] is able to fully update the storage at all N servers after each write operation even if some of the servers (up to a specified threshold value) are inaccessible, and achieves a synergistic gain from the joint design of private-read and private-write operations. Intuitively, the connection between these problems arises because the operation required at each server for many (but not all) PIR (and private write) schemes can be interpreted as a *matrix multiplication* between a threshold-T secret-shared query vector/matrix (polynomial encoded for T-privacy) and a threshold-X secret-shared data vector/matrix (polynomial encoded for X-security), which produces various desired and undesired products. CSA codes are characterized by a Cauchy-Vandermonde structure that facilitates interference alignment of undesired products along the Vandermonde terms, while the desired products remain separable along the Cauchy terms. This alignment structure allows efficient downloads by reducing interference dimensions. Therefore, to the extent that a multiplication of polynomial encoded matrices is involved, and download efficiency is of concern, the same Cauchy-Vandermonde alignment structure facilitated by CSA codes turns out to be useful across these problems. It is also noteworthy that applications of CSA codes generalize naturally beyond matrix products, to tensor products, as seen in Double Blind Private Information Retrieval (M-way blind PIR in general) [146].

IV. Private Federated Learning

A. Threat Models for Private Federated Learning

A typical federated learning (FL) system [147], [148] comprises users/workers, a server/curator, and an analyst, where users are connected to the server, and the server is subsequently connected to the analyst. Users wish to jointly train a machine learning model using their local datasets with the help of the server. The training is typically done using iterative algorithms such as gradient descent and its variants, where users receive the global learning model that needs to be trained from the server and compute gradients using their local datasets, and subsequently send the gradients or the updated local models back to the server for aggregation. The analyst may request for the model at any given time. Depending on who the malicious party is, its capability and intent, we can have several different threat models. For example, the analyst can be assumed to be honest but curious who does not actively attack the trustworthy server or users but tries to learn as much information about users as possible through the output released to it. Similarly, the server can also be assumed to be honest but curious. However, different from the analyst, the server can also be an active attacker, who alters the training process and/or baits users into revealing their information. Another possible threat model is when a subset of users is malicious, who try to tamper with the training process by sending altered gradients or model updates. We refer the reader to a recent excellent comprehensive survey on the subject of FL [148], which gives an in-depth account of recent progress on various FL modalities, as well as challenges in achieving efficiency, privacy, fairness, and system level implementation.

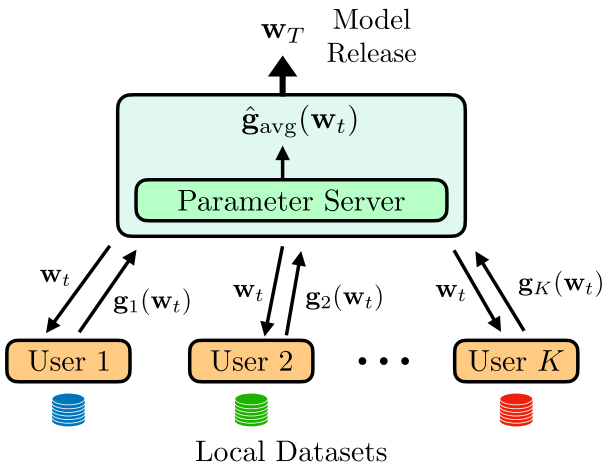

Fig. 8. Conventional federated learning system, where $\mathbf{w}_t$ denotes the model parameters at iteration t, and $\mathbf{g}_k(\mathbf{w}_t)$ denotes the gradient computed using $\mathbf{w}_t$ by user k. After training, the model is released publicly.

In this survey, we focus on the models where (a) the server is trustworthy and the analyst is honest but curious; and (b) both the server and the analyst are honest but curious. One may think that no information can be learned by the curious party due to the fact that the local data never leave the users, therefore, the local data is private. However, it has been shown that even gradients or updated models can be used to recover the data used during training for feed-forward neural networks [149]–[151] and convolutional neural network [152], [153]. This type of attack is known as gradient/model inversion attack.

B. Differential Private Federated Learning

In private distributed computing, where the entire data is available at a central location (user), as discussed in the previous section, it is indeed possible to achieve perfect privacy (in an information-theoretic sense) when performing computations over distributed cluster of nodes. The federated learning paradigm, however, has several key distinctions as we

briefly highlight next: since the data is already locally spread at the users (and is required to be kept private), perfect privacy against a single server can only be achieved by completely sacrificing utility (in terms of the model learned by perfectly private interactions with the user). Thus, in conventional single-server FL, one seeks to relax the privacy requirements from perfect privacy to allowing some leakage in a graceful manner. Indeed, as is shown in [145], perfect privacy can be feasible with multiple servers, and when one may be interested in training multiple sub-models at the servers, or when some collaboration between the users is allowed (also see the discussion in Section IV-C). For the remainder of this section, we will exclusively focus on the single-server FL setting when the users cannot collaborate.

Differential privacy (DP) [154] is one of the most widely used privacy notions and has been shown to be effective to mitigate not only inversion attacks, but also differential attacks. The goal is to protect the private data by perturbing the output before it is released to untrustworthy parties. Depending on who performs the perturbation or who we wish to protect against, differential privacy can be further categorized into local DP and central DP. For a FL system with K users, the local and central DP are formally defined as follows.

Definition 1: ($(\epsilon_\ell^{(k)}, \delta_\ell)$-LDP) Let $\mathcal{X}_k$ be a set of all possible data points at user k. For user k, a randomized mechanism $\mathcal{M}_k : \mathcal{X}_k \to \mathbb{R}^d$ is $(\epsilon_\ell^{(k)}, \delta_\ell)$-LDP if for any $x,\ x' \in \mathcal{X}_k$, and any measurable subset $\mathcal{O}_k \subseteq$ Range$(\mathcal{M}_k)$, we have

$$\Pr(\mathcal{M}_k(x) \in \mathcal{O}_k) \leq \exp(\epsilon_\ell^{(k)}) \Pr(\mathcal{M}_k(x') \in \mathcal{O}_k) + \delta_\ell. \quad (4)$$

The setting when $\delta_\ell = 0$ is referred as pure $\epsilon_\ell^{(k)}$-LDP.

Definition 2: ((ϵ_c, δ_c)-DP) Let $\mathcal{D} \triangleq \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_K$ be the collection of all possible datasets of all K users. A randomized mechanism $\mathcal{M} : \mathcal{D} \to \mathbb{R}^d$ is (ϵ_c, δ_c)-DP if for any two neighboring datasets D, D' and any measurable subset $\mathcal{O} \subseteq$ Range$(\mathcal{M})$, we have

$$\Pr(\mathcal{M}(D) \in \mathcal{O}) \leq \exp(\epsilon_c) \Pr(\mathcal{M}(D') \in \mathcal{O}) + \delta_c. \quad (5)$$

The setting when $\delta_c = 0$ is referred as pure ϵ_c-DP.

We refer ϵ_c ($\epsilon_\ell^{(k)}$) and δ_c (δ_ℓ) as privacy parameters. These parameters are closely associated with a quantity called sensitivity, which is defined as the largest difference of a function over all available inputs. It is known that central DP is a weaker guarantee than local DP. Therefore, local DP guarantee implies central DP guarantee. Both central and local DP are first computed on a per-iteration basis. Then, the total leakage is computed by summing up the leakages over all iterations. However, simply summing up the leakages over all iterations provides bound on the actual total leakage due to the fact that data is often reused during training. It is known that the more a data point is used, the more information it leaks. Therefore, to capture this phenomenon, various of composition theorem/leakage accountant methods are used to tighten the bound on the total leakage, such as advanced composition theorem, and moment accountant [155].

1) *Basic privacy preserving mechanisms*: Let us first look at the case where the server is trustworthy, and the goal is to satisfy a desired central DP level against the curious analyst. The outputs, e.g., learning model iterates or gradients, have been shown to leak information about the local datasets. Therefore, the goal is to perturb the outputs so that it becomes difficult for the analyst to learn information about the local datasets. Typically, for works that focus on central DP, the perturbation is done at the server. The outputs can be perturbed by using random response, adding noise, or using approximations. For example, in [155], Gaussian noise is added to the gradient before the model update. However, gradients and model parameters often are represented with finite precision, which make Gaussian noise injecting mechanism impractical. Thus, other types of noise injecting mechanisms are also considered, such as Laplace mechanism, and for discrete values, binomial mechanism can be used [156]. Noise with custom density can also be used [157]. However, as mentioned earlier, the privacy guarantee degrades when the same data is used for training repeatedly.
2) *Privacy amplification via sampling*: To remedy this issue, another line of work focuses on reducing the exposure of the data through user [158], [159] or data point sampling [155], [160], [167], [168]. In [158], users are sampled i.i.d. according to some probability, who will then compute and send the gradients to the server for perturbation and model updates. In [160], exponential mechanism is studied. In [167], various data sampling schemes, such as Poisson sampling, sampling with/without replacement, are studied and analyzed. As a result of sampling, the privacy level is *amplified* [167], i.e., less noise is needed to achieve the same privacy level that is achieved by schemes without sampling. Amplification can also be obtained through shuffling [161], where a trusted shuffler shuffles the outputs from users before sending it to the server. Works that consider shuffling as part of the pipeline include [162], [163]. Another way to control leakage is to ensure that the sensitivity is small by carefully choosing the clipping norm [164].
3) *Private FL over new communication models*: The above works rely on assumption that the server (and the shuffler) is trustworthy. This assumption may not be practical in certain scenarios. To remove this assumption, local DP was proposed and studied, where each user is responsible for protecting their own data. Similar to central DP, one can ask users to directly perturb the information they want to send, e.g., in [165]. In addition to adding noise, one can also perturb the information by using approximation. In [173], the approximation is obtained by flipping a random bit in the input string of each user. In [174], a random vector that is roughly in the same direction as the original gradient is sampled and used as an approximation by each user. However, it has been shown that techniques that let users perturb information directly to achieve LDP may suffer greatly in terms of utility, i.e., accuracy. In order to satisfy LDP and provide reasonable utility, we can again use the idea of privacy amplification using sampling and/or shuffling [161],

TABLE I
A QUICK REFERENCE OF SOME OF THE KEY PRIVACY PRESERVING TECHNIQUES FOR PROVIDING CENTRAL AND LOCAL DP GUARANTEES IN FEDERATED LEARNING

	Noise Injection	Sampling	Shuffling	Others
Central DP	[161], [162]	[160], [163]–[165]	[166]–[168]	[169]
Local DP	[170], [171]	[172], [173]	[166], [174]–[177]	[178], [179]

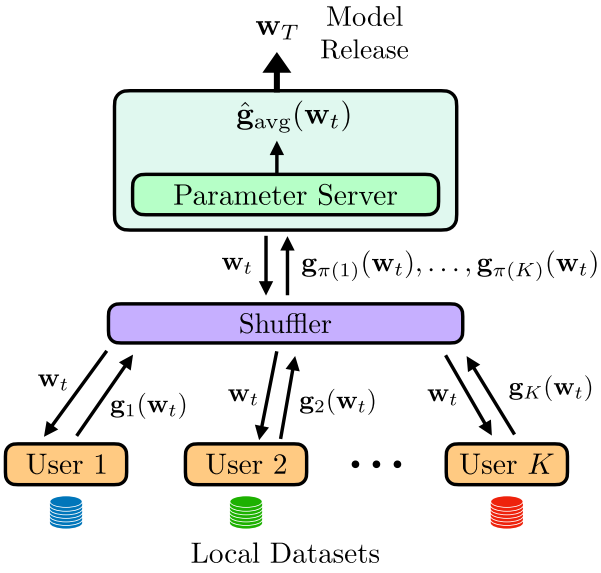

Fig. 9. Federated learning system with a shuffler, who shuffles the gradients before sending them to the PS. The trained model is subsequently released to the analyst.

[169]–[172]. Since information is perturbed at the user, honest but curious shuffler would not compromise the local DP. Intuitively, both sampling and shuffling are able to further confuse the curious party without injecting more noise. Therefore, one is able to inject less noise to maintain utility, and still achieve the desired privacy level via sampling and/or shuffling. Another line of work focuses on private FL over wireless channels [175]–[180]. With superposition property of wireless channel, the gradients can be naturally aggregated while being transmitted. It has been shown in [177] and [178] that, by carefully designing the power control factors, the channel noise can be used as perturbation and provides DP guarantee. Amplification results are shown in [175], where perturbation added by users that is aggregated over wireless channel enhances privacy guarantee, and in [180], the privacy is amplified by aggregated perturbation and the addition of user sampling in the FL pipeline. However, channel state information (CSI) is obtained with the help of the server and is crucial in these works. When the server is untrustworthy, CSI obtained from the server can be tampered to lurk users to leak information. Therefore, [179] and [180] study the case when CSI is not available.

4) *Other privacy notions and connections to DP*: There are also works that use different privacy notions, such as Concentrated DP [181], [182], Renyi DP [183], Bayesian DP [184], communication-constrained DP [172] and information-theoretic privacy [137]. Concentrated DP is a relaxed version of DP, where it ensures that leakage is centered around the expected privacy level ϵ_c, and is subgaussian. The probability that leakage exceeds ϵ_c by a small amount is bounded. Unlike the standard DP, where the expected leakage is not bounded and could potentially go to infinity with probability δ_c, leakage of concentrated DP does not go to infinity. Renyi DP is another relaxation of the standard DP that is based on the concept of Renyi divergence. Renyi DP can be translated to standard DP and it was shown to have better composition result in [183]. Bayesian DP is essentially standard DP, however, the data distribution is taken into account when quantifying privacy parameters. While one can show connection between DP and information-theoretic privacy, the approaches that are used to secure data are completely different. In [137], data is kept private by using error control codes and the idea of secret sharing. The data is considered private when the mutual information of the original data and encoded data is zero. Other works such as [185] studies how to allocate the amount of noise added to the data by each user in a decentralized setting (without the presence of a server) so that the collective noise does not reduce the utility.

C. Secure Model Aggregation in Federated Learning

While data is kept at the user-side in FL, a user's model still carries a significant amount of information about the local dataset of this user. Specifically, as shown recently, the private training data can be reconstructed from the local models through inference or inversion attacks (see e.g., [186]–[189]). To prevent such information leakage, *secure aggregation* protocols are proposed (e.g., [139], [190]–[195]) to protect the privacy of individual local models, both from the server and other users, while still allowing the server to learn the aggregate model of the users. More specifically, secure aggregation protocols ensure that, at any given round, the server can only learn the aggregate model of the users, and beyond that no further information is revealed about the individual local model of a particular user. The key idea of the secure aggregation protocols is that the users mask their models before sending them to the server. These masks then cancel out when the server aggregates the masked models, which allows the server to learn the aggregate of the local models without revealing the individual models.

In the secure aggregation protocol of [190], known as SecAgg, pairwise secret keys are generated between each pair of users. For handling user dropouts, the pairwise keys in [190] are secret shared among all users, and can be reconstructed by the server in case of dropouts. This protocol tolerates any D dropped users and ensures privacy guarantee against up to T colluding users, provided that $T + D < N$, where N is the number of users. The communication cost of constructing these masks, however, scales as $O(N^2)$, which limits the scalability of this approach. Several works have considered designing communication-efficient secure aggregation protocol [139], [191]–[193], [196]. SecAgg+ [193] improves upon SecAgg [190] by limiting the secret sharing according to a sparse random graph instead of the complete graph considered in SecAgg [190]. TurboAgg [139] overcomes the quadratic aggregation overhead of [190], achieving a secure aggregation overhead of $O(N \log N)$, while simultaneously tolerating up to a user dropout rate of 50% and providing privacy against up to $N/2$ colluding users with high probability. The key idea of Turbo-Aggregate that enables communication-efficient aggregation is that it employs a multi-group circular strategy in which the users are partitioned into groups. The dropout and the privacy guarantees of TurboAgg, however, are not worst-case guarantees and it requires $\log N$ rounds. FastSecAgg [191] is a 3-round secure aggregation interactive communication-efficient protocol that is based on the Fast Fourier Transform multi-secret sharing, but it provides lower dropout and privacy guarantees compared to SecAgg [190]. While all of aforementioned works in secure aggregation provide cryptographic security, the secure aggregation protocol [192] provides information-theoretic security. In addition, unlike all previous protocols that depend on the pairwise random-seed reconstruction of the dropped users, this protocol departs from the previous protocols by employing instead one-shot aggregate-mask reconstruction of the surviving users. This feature can reduce the aggregation complexity significantly. However, this protocol relies on a trusted-third party to distribute the masks over the users. While all of the aforementioned works do not consider the bandwidth heterogeneity among the different users in secure aggregation, an adaptive secure aggregation protocol has been proposed in [196] which quantizes the model of each user according to the available bandwidth to improve the training accuracy.

While secure aggregation seeks to resolve the issue of preserving user data privacy by masking the individual model updates, the learning protocol can be adversarially affected by Byzantine users that may aim to break or perturb the learning to their benefit [197]–[201]. As the local models are protected by random masks, the server cannot observe the individual user updates in the clear, which prevents the server from utilizing outlier detection protocols to protect the model against Byzantine manipulations. This problem has been recently addressed in [198], for the I.I.D. setting, where the first single-server Byzantine-resilient secure aggregation protocol for secure federated learning known as BREA has been developed. BREA is based on distance based adversarial detection and leverages quantization and verifiable secret sharing to provide robustness against malicious users, while preserving the privacy of the individual user models.

All works on secure aggregation only guarantee the privacy of the individual users over a single aggregation round [139], [190]–[193]. While the privacy of the users is protected in each single round, the server can reconstruct an individual model from the aggregated models over multiple rounds of aggregation. Specifically, as a result of the client sampling strategy and the users dropouts, the server may be able to recover an individual model by exploiting the history of the aggregate models [202], [203]. This problem was studied for the first time in [203] which developed a client selection strategy known as Multi-RoundSecAgg that ensures the privacy of the individual users over all aggregation rounds while taking into account other important factors such as the aggregation fairness among the users and average number of users participating at each round (average aggregation cardinality) which control the convergence rate.

V. Discussion: Challenges and Open Problems

In this article, we have surveyed the privacy issues in information retrieval, distributed computation, and distributed (federated) learning. We conclude this article with the following incomplete list of remaining challenges and open problems in these areas.

Challenges and open problems in PIR:

- Coded colluding databases: The PIR problem is completely solved when the database content is coded for the case when the databases do not collude, and is also completely solved when the databases collude for the case when the database content is replicated (uncoded). However, the problem is open when database content is coded, potentially secured and the databases may collude. Remarkably, even the asymptotic capacity (for large number of messages) remains open. While the lower bound for U-B-XS-MDS-TPIR in Table 6 is conjectured to be aymptotically optimal, the asymptotic capacity C_∞ remains unknown in almost all cases.
- Non-replicated databases: The basic form of PIR assumes that the databases contain exactly the same set of files. In reality, the databases will have some overlap in content and will also have distinct items. When the databases have arbitrary contents, the PIR capacity problem is open, with a few notable exceptions [51]–[54]. The challenge here is to be able exploit the replication to reduce the download cost, while at the same time deal with non-replication as efficiently as possible.
- Upload cost and message size: In the capacity formulation, the upload cost is largely ignored. However, when the message sizes are not very large, the consideration on the upload cost becomes important. The problem then becomes how to construct codes for small message sizes to achieve a smaller upload cost. In general, PIR schemes should be designed to minimize a combined measure of upload and download costs.
- Weakly private (leaky) information retrieval: In some practical applications of PIR, it may not be absolutely

necessary to require perfect privacy, and a small leakage may be tolerable. In this setting, two questions stand out: What are good metric(s) to measure the leakage, and how to characterize the capacity as a function of these metrics.
- More complex message structure: The messages are usually required to be independent (and of the same length). What is the optimal coding strategy when the messages are dependent, either as overlapping parts, or are dependent following a general probability law?
- Privacy, stragglers and timeliness of retrieval: While PIR focuses mainly only on the privacy of downloaded information, it assumes that the servers are ideal, which respond to queries immediately and with no delays. Robust PIR problem considers the case of servers being completely unresponsive [55], [114]. However, most servers respond eventually, albeit slowly in many cases. Therefore, there is a need to design systems where information may be downloaded privately but also in a timely manner. An initial consideration of this issue is presented in a recent paper in [204], but the problem remains largely open.

Challenges and open problems in private distributed computing:

- Optimal coding for block matrix multiplication: In a standard coded computing setup, we aim to find coding designs to minimize the recovery threshold [205] (or the number of workers when no straggler is present) given fixed constraints on computation, security, and privacy. While the optimal recovery thresholds have been characterized within a factor of 2 for block matrix multiplication [131], [132], its exact characterization is not known except for some boundary cases. In particular, a remaining interesting open problem is to show whether the factor-of-2 penalty in the state-of-the-art upper bound is necessary when all partition parameters are large.
- Analog coded computing: Most works in coded computing such as [115], [131], [132], [205], [206] rely on quantizing the data into finite fields and then leveraging coding-theoretic techniques to mitigate stragglers and Byzantine workers, and to provide data privacy. This approach, however, degrades the accuracy of the computations [207], [208]. In fact, this is a limitation of many other problems such as verifiable computing and machine learning [209]–[211]. Recently, several works have extended LCC to the analog domain to address these challenges [207], [208], [212], [213], but they either focus only on straggler-mitigation [213], Byzantine-robustness [212], both issues [214] or privacy [207], [208]. An interesting open problem is to design a framework that jointly tackles these three challenges in the analog domain.

Challenges and open problems in private federated learning:

- Secure and Byzantine-robust aggregation in federated learning: While BREA [198] has considered secure aggregation and mitigating Byzantine users jointly, it has only focused on the i.i.d. setting. Extending BREA to the non-i.i.d. setting is an interesting future direction. The main challenge in this direction is to determine whether the updates that may seem deviating are due to the users having non-i.i.d. data or because of Byzantine users sending erroneous updates.
- Secure aggregation and multi-round secure aggregation: There are many open problems related to the multi-round secure aggregation problem introduced in [203]. While the secure aggregation protocols are believed to protect the privacy the of the individual users, it is not clear whether such protocols ensure privacy in the information-theoretic sense. Specifically, the secure aggregation protocols ensure that the server only learns the aggregate model of the users. However, the aggregate model of the users may still reveal information about the individual users and characterizing such a leakage is an important problem. While Multi-RoundSecAgg provides a trade-off between the multi-round privacy, the average aggregation cardinality, and the aggregation fairness, investigating the optimality of Multi-RoundSecAgg remains an open problem.

REFERENCES

[1] B. Chor, O. Goldreich, E. Kushilevitz, and M. Sudan, "Private information retrieval," in *Proc. 39th Annu. Symp. Found. Comput. Sci.*, Oct. 1995, pp. 41–50.

[2] A. Beimel and Y. Ishai, "Information-theoretic private information retrieval: A unified construction," in *Automata, Languages and Programming*. Springer, 2001, pp. 912–926.

[3] C. Cachin, S. Micali, and M. Stadler, "Computationally private information retrieval with polylogarithmic communication," in *Proc. Int. Conf. Theory Appl. Cryptogr. Techn.* Springer, 1999, pp. 402–414.

[4] H. Sun and S. A. Jafar, "The capacity of private information retrieval," *IEEE Trans. Inf. Theory*, vol. 63, no. 7, pp. 4075–4088, Jul. 2017.

[5] W. Gasarch, "A survey on private information retrieval," *Bull. EATCS*, vol. 82, pp. 72–107, Feb. 2004.

[6] O. Goldreich, H. Karloff, L. J. Schulman, and L. Trevisan, "Lower bounds for linear locally decodable codes and private information retrieval," in *Proc. 17th IEEE Annu. Conf. Comput. Complex.*, May 2002, pp. 175–183.

[7] S. Yekhanin, "Locally decodable codes and private information retrieval schemes," Ph.D. dissertation, Dept. EECS, Massachusetts Inst. Technol., Cambridge, MA, USA, 2007.

[8] G. Di Crescenzo, T. Malkin, and R. Ostrovsky, "Single database private information retrieval implies oblivious transfer," in *Proc. Int. Conf. Theory Appl. Cryptogr. Techn.* Springer, 2000, pp. 122–138.

[9] N. B. Shah, K. V. Rashmi, and K. Ramchandran, "One extra bit of download ensures perfectly private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 856–860.

[10] T. H. Chan, S.-W. Ho, and H. Yamamoto, "Private information retrieval for coded storage," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Nov. 2015, pp. 2842–2846.

[11] A. Fazeli, A. Vardy, and E. Yaakobi, "Codes for distributed PIR with low storage overhead," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Nov. 2015, pp. 2852–2856.

[12] C. Tian, H. Sun, and J. Chen, "Capacity-achieving private information retrieval codes with optimal message size and upload cost," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7613–7627, Nov. 2019.

[13] H. Sun and S. A. Jafar, "Optimal download cost of private information retrieval for arbitrary message length," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 12, pp. 2920–2932, Dec. 2017.

[14] Z. Zhang and J. Xu, "The optimal sub-packetization of linear capacity-achieving PIR schemes with colluding servers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2723–2735, May 2019.

[15] J. Xu and Z. Zhang, "On sub-packetization and access number of capacity-achieving PIR schemes for MDS coded non-colluding servers," *Sci. China Inf. Sci.*, vol. 61, no. 10, Oct. 2018, Art. no. 100306.

[16] C. Tian, H. Sun, and J. Chen, "Capacity-achieving private information retrieval codes with optimal message size and upload cost," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[17] I. Samy, R. Tandon, and L. Lazos, "On the capacity of leaky private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1262–1266.

[18] I. Samy, M. Attia, R. Tandon, and L. Lazos, "Asymmetric leaky private information retrieval," *IEEE Trans. Inf. Theory*, vol. 67, no. 8, pp. 5352–5369, Aug. 2021.

[19] H. Sun and S. A. Jafar, "Multiround private information retrieval: Capacity and storage overhead," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5743–5754, Aug. 2018.

[20] X. Yao, N. Liu, and W. Kang, "The capacity of multi-round private information retrieval from byzantine databases," 2019, *arXiv:1901.06907*.

[21] K. Banawan and S. Ulukus, "Multi-message private information retrieval: Capacity results and near-optimal schemes," *IEEE Trans. Inf. Theory*, vol. 64, no. 10, pp. 6842–6862, Oct. 2018.

[22] R. Tandon, "The capacity of cache aided private information retrieval," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2017, pp. 1078–1082.

[23] Y.-P. Wei, K. Banawan, and S. Ulukus, "Fundamental limits of cache-aided private information retrieval with unknown and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3215–3232, May 2019.

[24] Y.-P. Wei, K. Banawan, and S. Ulukus, "Cache-aided private information retrieval with partially known uncoded prefetching: Fundamental limits," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1126–1139, Jun. 2018.

[25] A. Heidarzadeh, B. Garcia, S. Kadhe, S. E. Rouayheb, and A. Sprintson, "On the capacity of single-server multi-message private information retrieval with side information," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2018, pp. 180–187.

[26] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private information retrieval with side information," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2032–2043, Apr. 2020.

[27] Z. Chen, Z. Wang, and S. A. Jafar, "The capacity of T-private information retrieval with private side information," *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 4761–4773, Aug. 2020.

[28] Y.-P. Wei, K. Banawan, and S. Ulukus, "The capacity of private information retrieval with partially known private side information," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8222–8231, Dec. 2019.

[29] Y.-P. Wei and S. Ulukus, "The capacity of private information retrieval with private side information under storage constraints," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2023–2031, Apr. 2020.

[30] S. P. Shariatpanahi, M. J. Siavoshani, and M. A. Maddah-Ali, "Multi-message private information retrieval with private side information," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[31] A. Heidarzadeh, S. Kadhe, S. El Rouayheb, and A. Sprintson, "Single-server multi-message individually-private information retrieval with side information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1042–1046.

[32] S. Li and M. Gastpar, "Single-server multi-message private information retrieval with side information," in *Proc. 56th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Oct. 2018, pp. 173–179.

[33] S. Li and M. Gastpar, "Single-server multi-message private information retrieval with side information: The general cases," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1083–1088.

[34] A. Heidarzadeh, F. Kazemi, and A. Sprintson, "Capacity of single-server single-message private information retrieval with private coded side information," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1662–1666.

[35] Z. Wang and S. Ulukus, "Symmetric private information retrieval with user-side common randomness," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 1–6.

[36] K. Banawan and S. Ulukus, "The capacity of private information retrieval from coded databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1945–1956, Mar. 2018.

[37] R. Tajeddine, O. W. Gnilke, and S. El Rouayheb, "Private information retrieval from MDS coded data in distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7081–7093, Nov. 2018.

[38] R. Zhou, C. Tian, H. Sun, and T. Liu, "Capacity-achieving private information retrieval codes from MDS-coded databases with minimum message size," *IEEE Trans. Inf. Theory*, vol. 66, no. 8, pp. 4904–4916, Aug. 2020.

[39] R. Freij-Hollanti, O. W. Gnilke, C. Hollanti, and D. A. Karpuk, "Private information retrieval from coded databases with colluding servers," *SIAM J. Appl. Algebra Geometry*, vol. 1, no. 1, pp. 647–664, Nov. 2017.

[40] H. Sun and S. A. Jafar, "Private information retrieval from MDS coded data with colluding servers: Settling a conjecture by Freij-Hollanti," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 1000–1022, Feb. 2018.

[41] S. Kumar, H.-Y. Lin, E. Rosnes, and A. G. i Amat, "Achieving maximum distance separable private information retrieval capacity with linear codes," *IEEE Trans. Inf. Theory*, vol. 65, no. 7, pp. 4243–4273, Jul. 2019.

[42] H.-Y. Lin, S. Kumar, E. Rosnes, and A. G. i Amat, "An MDS-PIR capacity-achieving protocol for distributed storage using non-MDS linear codes," 2018, *arXiv:1801.04923*.

[43] T. Guo, R. Zhou, and C. Tian, "New results on the storage-retrieval tradeoff in private information retrieval systems," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 403–414, Mar. 2021.

[44] K. Banawan, B. Arasli, and S. Ulukus, "Improved storage for efficient private information retrieval," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Aug. 2019, pp. 1–5.

[45] C. Tian, "On the storage cost of private information retrieval," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7539–7549, Dec. 2020.

[46] C. Tian, H. Sun, and J. Chen, "A Shannon-theoretic approach to the storage-retrieval tradeoff in pir systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 1904–1908.

[47] H. Sun and C. Tian, "Breaking the MDS-PIR capacity barrier via joint storage coding," *Information*, vol. 10, no. 9, p. 265, Aug. 2019.

[48] M. A. Attia, D. Kumar, and R. Tandon, "The capacity of private information retrieval from uncoded storage constrained databases," 2018, *arXiv:1805.04104*.

[49] K. Banawan, B. Arasli, Y.-P. Wei, and S. Ulukus, "The capacity of private information retrieval from heterogeneous uncoded caching databases," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3407–3416, Jun. 2020.

[50] Y.-P. Wei, B. Arasli, K. Banawan, and S. Ulukus, "The capacity of private information retrieval from decentralized uncoded caching databases," *Information*, vol. 10, no. 12, p. 372, Nov. 2019.

[51] N. Raviv, I. Tamo, and E. Yaakobi, "Private information retrieval in graph-based replication systems," *IEEE Trans. Inf. Theory*, vol. 66, no. 6, pp. 3590–3602, Jun. 2020.

[52] K. Banawan and S. Ulukus, "Private information retrieval from non-replicated databases," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1272–1276.

[53] Z. Jia and S. A. Jafar, "On the asymptotic capacity of X-secure T-private information retrieval with graph-based replicated storage," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6280–6296, Oct. 2020.

[54] B. Sadeh, Y. Gu, and I. Tamo, "Bounds on the capacity of PIR over graphs," 2021, *arXiv:2105.07704*.

[55] H. Sun and S. A. Jafar, "The capacity of robust private information retrieval with colluding databases," *IEEE Trans. Inf. Theory*, vol. 64, no. 4, pp. 2361–2370, Apr. 2018.

[56] K. Banawan and S. Ulukus, "The capacity of private information retrieval from Byzantine and colluding databases," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 1206–1219, Feb. 2019.

[57] H. Sun and S. A. Jafar, "The capacity of symmetric private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 1, pp. 322–329, Jan. 2019.

[58] Z. Wang, K. Banawan, and S. Ulukus, "Private set intersection: A multi-message symmetric private information retrieval perspective," 2020, *arXiv:1912.13501*.

[59] Z. Wang, K. Banawan, and S. Ulukus, "Multi-party private set intersection: An information-theoretic approach," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 366–379, Mar. 2021.

[60] K. Banawan and S. Ulukus, "Noisy private information retrieval: On separability of channel coding and information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8232–8249, Dec. 2019.

[61] K. Banawan and S. Ulukus, "Asymmetry hurts: Private information retrieval under asymmetric traffic constraints," *IEEE Trans. Inf. Theory*, vol. 65, no. 11, pp. 7628–7645, Nov. 2019.

[62] Q. Wang and M. Skoglund, "On PIR and symmetric PIR from colluding databases with adversaries and eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3183–3197, May 2019.

[63] Q. Wang, H. Sun, and M. Skoglund, "The capacity of private information retrieval with eavesdroppers," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 3198–3214, May 2019.

[64] K. Banawan and S. Ulukus, "Private information retrieval through wiretap channel II: Privacy meets security," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4129–4149, Jul. 2020.

[65] H. Yang, W. Shin, and J. Lee, "Private information retrieval for secure distributed storage systems," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 12, pp. 2953–2964, Dec. 2018.

[66] S. Vithana, K. Banawan, and S. Ulukus, "Semantic private information retrieval," 2020, *arXiv:2003.13667*.

[67] D. Asonov and J.-C. Freytag, "Repudiative information retrieval," in *Proc. ACM workshop Privacy Electron. Soc.*, vol. 2002, pp. 32–40.

[68] R. R. Toledo, G. Danezis, and I. Goldberg, "Lower-cost-private information retrieval," *Proc. Privacy Enhancing Technol.*, vol. 2016, no. 4, pp. 184–201, Oct. 2016.

[69] B. Chor, E. Kushilevitz, O. Goldreich, and M. Sudan, "Private information retrieval," *J. ACM*, vol. 45, no. 6, pp. 965–981, Nov. 1998.

[70] H.-Y. Lin, S. Kumar, E. Rosnes, A. G. I. Amat, and E. Yaakobi, "Weakly-private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1257–1261.

[71] H.-Y. Lin, S. Kumar, E. Rosnes, A. Graell i Amat, and E. Yaakobi, "The capacity of single-server weakly-private information retrieval," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 415–427, Mar. 2021.

[72] R. Zhou, T. Guo, and C. Tian, "Weakly private information retrieval under the maximal leakage metric," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1089–1094.

[73] T. Guo, R. Zhou, and C. Tian, "On the information leakage in private information retrieval systems," 2019, *arXiv:1909.11605*.

[74] I. Samy, M. A. Attia, R. Tandon, and L. Lazos, "Latent-variable private information retrieval," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1071–1076.

[75] H. Sun and S. A. Jafar, "The capacity of private computation," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3880–3897, Jun. 2018.

[76] M. Mirmohseni and M. A. Maddah-Ali, "Private function retrieval," 2017, *arXiv:1711.04677*.

[77] X. Yao, N. Liu, and W. Kang, "The capacity of private information retrieval under arbitrary collusion patterns," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 1041–1046.

[78] R. Tandon, "The capacity of cache aided private information retrieval," 2017, *arXiv:1706.07035*.

[79] S. Kadhe, B. Garcia, A. Heidarzadeh, S. El Rouayheb, and A. Sprintson, "Private information retrieval with side information," 2017, *arXiv:1709.00112*.

[80] S. Li and M. Gastpar, "Converse for multi-server single-message PIR with side information," 2018, *arXiv:1809.09861*.

[81] Q. Wang and M. Skoglund, "Secure symmetric private information retrieval from colluding databases with adversaries," 2017, *arXiv:1707.02152*.

[82] S. Song and M. Hayashi, "Capacity of quantum private information retrieval with colluding servers," *IEEE Trans. Inf. Theory*, vol. 67, no. 8, pp. 5491–5508, Aug. 2021.

[83] Z. Jia, H. Sun, and S. A. Jafar, "Cross subspace alignment and the asymptotic capacity of X-secure T-private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5783–5798, Sep. 2019.

[84] R. Tajeddine, O. W. Gnilke, D. Karpuk, R. Freij-Hollanti, and C. Hollanti, "Private information retrieval from coded storage systems with colluding, Byzantine, and unresponsive servers," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3898–3906, Jun. 2019.

[85] Z. Jia and S. A. Jafar, "X-secure T-private information retrieval from MDS coded storage with byzantine and unresponsive servers," *IEEE Trans. Inf. Theory*, vol. 66, no. 12, pp. 7427–7438, Dec. 2020.

[86] Y. Gertner, Y. Ishai, E. Kushilevitz, and T. Malkin, "Protecting data privacy in private information retrieval schemes," in *Proc. 30th Annu. ACM Symp. Theory Comput. (STOC)*, 1998, pp. 151–160.

[87] J. Feigenbaum, "Encrypting problem instances," in *Advances in Cryptology*. Springer, 1985, pp. 477–488.

[88] M. Abadi, J. Feigenbaum, and J. Kilian, "On hiding information from an oracle," in *Proc. 19th Annu. ACM Symp. Theory Comput.*, 1987, pp. 195–203.

[89] D. Beaver and J. Feigenbaum, "Hiding instances in multioracle queries," in *STACS*. Springer, 1990, pp. 37–48.

[90] D. Beaver, J. Feigenbaum, J. Kilian, and P. Rogaway, "Locally random reductions: Improvements and applications," *J. Cryptol.*, vol. 10, no. 1, pp. 17–36, 1997.

[91] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, Nov. 1979.

[92] A. Beimel, Y. Ishai, E. Kushilevitz, and I. Orlov, "Share conversion and private information retrieval," in *Proc. IEEE 27th Conf. Comput. Complex.*, Jun. 2012, pp. 258–268.

[93] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, "On the locality of codeword symbols," *IEEE Trans. Inf. Theory*, vol. 58, no. 11, pp. 6925–6934, Nov. 2012.

[94] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A survey on network codes for distributed storage," *Proc. IEEE*, vol. 99, no. 3, pp. 476–489, Mar. 2011.

[95] Y. Birk and T. Kol, "Coding on demand by an informed source (ISCOD) for efficient broadcast of different supplemental data to caching clients," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2825–2830, Jun. 2006.

[96] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, Jul. 2000.

[97] S. Jafar, "Interference alignment: A new look at signal dimensions in a communication network," in *Foundations and Trends in Communication and Information Theory*. 2011, pp. 1–136.

[98] H. Sun and S. A. Jafar, "Blind interference alignment for private information retrieval," 2016, *arXiv:1601.07885*.

[99] S. A. Jafar, "Topological interference management through index coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 529–568, Jan. 2014.

[100] K. Shanmugam and A. G. Dimakis, "Bounding multiple unicasts through index coding and locally repairable codes," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014.

[101] A. Mazumdar, "On a duality between recoverable distributed storage and index coding," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2014, pp. 1977–1981.

[102] S. El Rouayheb, A. Sprintson, and C. Georghiades, "On the index coding problem and its relation to network coding and matroid theory," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3187–3195, Jul. 2010.

[103] M. Effros, S. El Rouayheb, and M. Langberg, "An equivalence between network coding and index coding," 2012, *arXiv:1211.6660*.

[104] A. Mazumdar, "Storage capacity of repairable networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 5810–5821, Nov. 2015.

[105] S. Riis, "Information flows, graphs and their guessing numbers," *Electron. J. Combinatorics*, vol. 14, no. 1, pp. 1–17, Jun. 2007.

[106] V. Cadambe and S. Jafar, "Interference alignment and the degrees of freedom of the K user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008.

[107] V. Cadambe, S. Jafar, H. Maleki, K. Ramchandran, and C. Suh, "Asymptotic interference alignment for optimal repair of mds codes in distributed data storage," *IEEE Trans. Inf. Theory*, vol. 59, no. 5, pp. 2974–2987, May 2013.

[108] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding up distributed machine learning using codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 3, pp. 1514–1529, Mar. 2018.

[109] R. Tandon, Q. Lei, A. G. Dimakis, and N. Karampatziakis, "Gradient coding," 2016, *arXiv:1612.03301*.

[110] Q. Yu, S. Li, M. A. Maddah-Ali, and A. S. Avestimehr, "How to optimally allocate resources for coded distributed computing?" 2017, *arXiv:1702.07297*.

[111] S. Dutta, V. Cadambe, and P. Grover, "Short-dot: Computing large linear transforms distributedly using coded short dot products," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2092–2100.

[112] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Straggler mitigation in distributed matrix multiplication: Fundamental limits and optimal coding," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1920–1933, Mar. 2020.

[113] R. Bitar, P. Parag, and S. El Rouayheb, "Minimizing latency for secure coded computing using secret sharing via staircase codes," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4609–4619, Aug. 2020.

[114] R. Bitar and S. E. Rouayheb, "Staircase-PIR: Universally robust private information retrieval," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[115] Q. Yu, S. Li, N. Raviv, S. M. M. Kalan, M. Soltanolkotabi, and S. A. Avestimehr, "Lagrange coded computing: Optimal design for resiliency, security, and privacy," in *Proceedings of Machine Learning Research*, vol. 89, K. Chaudhuri and M. Sugiyama, Eds. 2018, pp. 1215–1225. [Online]. Available: https://proceedings.mlr.press/v89/yu19b.html

[116] W.-T. Chang and R. Tandon, "On the capacity of secure distributed matrix multiplication," 2018, *arXiv:1806.00469*.

[117] H. Yang and J. Lee, "Secure distributed computing with straggling servers using polynomial codes," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 1, pp. 141–150, Jan. 2019.

[118] J. Kakar, S. Ebadifar, and A. Sezgin, "Rate-efficiency and straggler-robustness through partition in distributed two-sided secure matrix computation," 2018, *arXiv:1810.13006*.

[119] R. G. L. D'Oliveira, S. El Rouayheb, and D. Karpuk, "Gasp codes for secure distributed matrix multiplication," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1107–1111.

[120] H. A. Nodehi, S. R. H. Najarkolaei, and M. A. Maddah-Ali, "Entangled polynomial coding in limited-sharing multi-party computation," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Nov. 2018, pp. 1–5.

[121] M. Aliasgari, O. Simeone, and J. Kliewer, "Distributed and private coded matrix computation with flexible communication load," 2019, *arXiv:1901.07705*.

[122] M. Kim and J. Lee, "Private secure coded computation," 2019, *arXiv:1902.00167*.

[123] J. Kakar, S. Ebadifar, and A. Sezgin, "On the capacity and straggler-robustness of distributed secure matrix multiplication," *IEEE Access*, vol. 7, pp. 45783–45799, 2019.

[124] W.-T. Chang and R. Tandon, "On the upload versus download cost for secure and private matrix multiplication," 2019, *arXiv:1906.10684*.

[125] S. Ebadifar, J. Kakar, and A. Sezgin, "The need for alignment in rate-efficient distributed two-sided secure matrix computation," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[126] H. A. Nodehi and M. A. Maddah-Ali, "Secure coded multi-party computation for massive matrix operations," 2019, *arXiv:1908.04255*.

[127] Z. Jia and S. A. Jafar, "On the capacity of secure distributed batch matrix multiplication," 2019, *arXiv:1908.06957*.

[128] J. Kakar, A. Khristoforov, S. Ebadifar, and A. Sezgin, "Uplink-downlink tradeoff in secure distributed matrix multiplication," 2019, *arXiv:1910.13849*.

[129] M. Aliasgari, O. Simeone, and J. Kliewer, "Private and secure distributed matrix multiplication with flexible communication load," 2019, *arXiv:1909.00407*.

[130] R. G. D'Oliveira, S. El Rouayheb, D. Heinlein, and D. Karpuk, "Degree tables for secure distributed matrix multiplication," Tech. Rep., 2019.

[131] Q. Yu and A. S. Avestimehr, "Entangled polynomial codes for secure, private, and batch distributed matrix multiplication: Breaking the 'cubic' barrier," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 245–250.

[132] Q. Yu, M. A. Maddah-Ali, and A. S. Avestimehr, "Straggler mitigation in distributed matrix multiplication: Fundamental limits and optimal coding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 2022–2026.

[133] M. Bläser, (2013). *Fast Matrix Multiplication*. Graduate Surveys. Theory of Computing Library. [Online]. Available: https://www.theoryofcomputing.org/library.html

[134] Q. Yu and A. S. Avestimehr, "Harmonic coding: An optimal linear code for privacy-preserving gradient-type computation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1102–1106.

[135] Z. Chen, Z. Jia, Z. Wang, and S. A. Jafar, "GCSA codes with noise alignment for secure coded multi-party batch matrix multiplication," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 306–316, Mar. 2021.

[136] N. Raviv, Q. Yu, J. Bruck, and S. Avestimehr, "Download and access trade-offs in Lagrange coded computing," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2019, pp. 1787–1791.

[137] J. So, B. Guler, and A. S. Avestimehr, "CodedPrivateML: A fast and privacy-preserving framework for distributed machine learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 441–451, Mar. 2021.

[138] J. So, B. Guler, and S. Avestimehr, "A scalable approach for privacy-preserving collaborative machine learning," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 8054–8066. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/5bf8aaef51c6e0d363cbe554acaf3f20-Paper.pdf

[139] J. So, B. Guler, and A. S. Avestimehr, "Turbo-aggregate: Breaking the quadratic aggregation barrier in secure federated learning," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 479–489, Mar. 2021.

[140] M. Kim and J. Lee, "Private secure coded computation," *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 1918–1921, Nov. 2019, doi: 10.1109/LCOMM.2019.2934436.

[141] W. Chang and R. Tandon, "On the upload versus download cost for secure and private matrix multiplication," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Visby, Sweden, Aug. 2019, pp. 1–5.

[142] W.-T. Chang and R. Tandon, "On the capacity of secure distributed matrix multiplication," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[143] Z. Jia and S. A. Jafar, "On the capacity of secure distributed batch matrix multiplication," 2019, *arXiv:1908.06957*.

[144] Z. Jia and S. A. Jafar, "Cross subspace alignment codes for coded distributed batch computation," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2821–2846, May 2021.

[145] Z. Jia and S. Jafar, "X-secure t-private federated submodel learning with elastic dropout resilience," Oct. 2020, *arXiv:2020.01059*.

[146] Y. Lu, Z. Jia, and S. A. Jafar, "Double blind T-Private information retrieval," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 428–440, Mar. 2021.

[147] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, vol. 54, A. Singh and J. Zhu, Eds. Fort Lauderdale, FL, USA: PMLR, Apr. 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[148] P. Kairouz *et al.*, "Advances and open problems in federated learning," in *Foundations and Trends in Machine Learning*, vol. 1, nos. 1–2. NOW Publishers, 2021, pp. 1–210.

[149] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning: Revisited and enhanced," in *Applications and Techniques in Information Security*, L. Batten, D. S. Kim, X. Zhang, and G. Li, Eds. Singapore: Springer, 2017, pp. 100–110.

[150] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai, "Privacy-preserving deep learning via additively homomorphic encryption," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 5, pp. 1333–1345, May 2018.

[151] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients—How easy is it to break privacy in federated learning?" in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 16937–16947. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/c4ede56bbd98819ae6112b20ac6bf145-Paper.pdf

[152] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf

[153] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2019, pp. 2512–2520.

[154] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.

[155] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, New York, NY, USA, Oct. 2016, pp. 308–318, doi: 10.1145/2976749.2978318.

[156] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar, and B. McMahan, "CpSGD: Communication-efficient and differentially-private distributed SGD," in *Advances in Neural Information Processing Systems*, vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/21ce689121e39821d07d04faab328370-Paper.pdf

[157] S. Song, K. Chaudhuri, and A. D. Sarwate, "Stochastic gradient descent with differentially private updates," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2013, pp. 245–248.

[158] B. McMahan, D. Ramage, K. Talwar, and L. Zhang, "Learning differentially private recurrent language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–14. [Online]. Available: https://openreview.net/pdf?id=BJ0hF1Z0b

[159] S. Asoodeh, W.-N. Chen, F. P. Calmon, and A. Ozgur, "Differentially private federated learning: An information-theoretic perspective," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2021, pp. 1–8. [Online]. Available: http://federated-learning.org/fl-icml-2020/

[160] R. Bassily, A. Smith, and A. Thakurta, "Private empirical risk minimization: Efficient algorithms and tight error bounds," in *Proc. IEEE 55th Annu. Symp. Found. Comput. Sci.*, Oct. 2014, pp. 464–473, doi: 10.1109/FOCS.2014.56.

[161] U. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta, "Amplification by shuffling: From local to central differential privacy via anonymity," in *Proc. 30th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2019, pp. 2468–2479.

[162] A. Bittau *et al.*, "Prochlo: Strong privacy for analytics in the crowd," in *Proc. Symp. Operating Syst. Princ. (SOSP)*, 2017, pp. 441–459.

[163] A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev, "Distributed differential privacy via shuffling," 2018, *arXiv:1808.01394*.

[164] G. Andrew, O. Thakkar, H. B. McMahan, and S. Ramaswamy, "Differentially private learning with adaptive clipping," 2019, *arXiv:1905.03871*.

[165] H. Ono and T. Takahashi, "Locally private distributed reinforcement learning," 2020, *arXiv:2001.11718*.

[166] Y. Li, T.-H. Chang, and C.-Y. Chi, "Secure federated averaging algorithm with differential privacy," in *Proc. IEEE 30th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2020, pp. 1–6.

[167] B. Balle, G. Barthe, and M. Gaboardi, "Privacy amplification by subsampling: Tight analyses via couplings and divergences," in *Advances in Neural Information Processing Systems*, vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/3b5020bb891119b9f5130f1fea9bd773-Paper.pdf

[168] M. A. Heikkilä, A. Koskela, K. Shimizu, S. Kaski, and A. Honkela, "Differentially private cross-silo federated learning," 2020, *arXiv:2007.05553*.

[169] B. Balle, J. Bell, A. Gascón, and K. Nissim, "The privacy blanket of the shuffle model," in *Advances in Cryptology*, A. Boldyreva and D. Micciancio, Eds. Cham, Switzerland: Springer, 2019, pp. 638–667.

[170] B. Ghazi, R. Pagh, and A. Velingker, "Scalable and differentially private distributed aggregation in the shuffled model," 2019, *arXiv:1906.08320*.

[171] B. Balle, P. Kairouz, B. McMahan, O. Thakkar, and A. G. Thakurta, "Privacy amplification via random check-ins," in *Advances in Neural Information Processing Systems*, vol. 33, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Red Hook, NY, USA: Curran Associates, 2020, pp. 4623–4634. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/313f422ac583444ba6045cd122653b0e-Paper.pdf

[172] A. Girgis, D. Data, S. Diggavi, P. Kairouz, and A. Theertha Suresh, "Shuffled model of differential privacy in federated learning," in *Proc. 24th Int. Conf. Artif. Intell. Statist.*, vol. 130, A. Banerjee and K. Fukumizu, Eds., Apr. 2021, pp. 2521–2529. [Online]. Available: https://proceedings.mlr.press/v130/girgis21a.html

[173] A. Smith, A. Thakurta, and J. Upadhyay, "Is interaction necessary for distributed private learning?" in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 58–77.

[174] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci.*, Oct. 2013, pp. 429–438.

[175] M. Seif, R. Tandon, and M. Li, "Wireless federated learning with local differential privacy," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 2604–2609.

[176] A. Sonee and S. Rini, "Efficient federated learning over multiple access channel with differential privacy constraints," 2020, *arXiv:2005.07776*.

[177] Y. Koda, K. Yamamoto, T. Nishio, and M. Morikura, "Differentially private AirComp federated learning with power adaptation harnessing receiver noise," 2020, *arXiv:2004.06337*.

[178] D. Liu and O. Simeone, "Privacy for free: Wireless federated learning via uncoded transmission with adaptive power control," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 170–185, Jan. 2021.

[179] B. Hasircioglu and D. Gunduz, "Private wireless federated learning with anonymous over-the-air computation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5195–5199.

[180] M. Seif, W.-T. Chang, and R. Tandon, "Privacy amplification for federated learning via user sampling and wireless aggregation," 2021, *arXiv:2103.01953*.

[181] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," 2016, *arXiv:1603.01887*.

[182] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory Cryptography*, M. Hirt and A. Smith, Eds. Berlin, Germany: Springer, 2016, pp. 635–658.

[183] I. Mironov, "Renyi differential privacy," 2017, *arXiv:1702.07476*.

[184] A. Triastcyn and B. Faltings, "Federated learning with Bayesian differential privacy," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 2587–2596.

[185] S. Guo, T. Zhang, G. Xu, H. Yu, T. Xiang, and Y. Liu, "Topology-aware differential privacy for decentralized image classification," 2020, *arXiv:2006.07817*.

[186] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.

[187] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 739–753.

[188] L. Zhu and S. Han, "Deep leakage from gradients," in *Federated Learning*. Springer, 2020, pp. 17–31.

[189] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting gradients - how easy is it to break privacy in federated learning?" 2020, *arXiv:2003.14053*.

[190] K. Bonawitz *et al.*, "Practical secure aggregation for privacy-preserving machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 1175–1191.

[191] S. Kadhe, N. Rajaraman, O. O. Koyluoglu, and K. Ramchandran, "FastSecAgg: Scalable secure aggregation for privacy-preserving federated learning," 2020, *arXiv:2009.11248*.

[192] Y. Zhao and H. Sun, "Information theoretic secure aggregation with user dropouts," 2021, *arXiv:2101.07750*.

[193] J. H. Bell, K. A. Bonawitz, A. Gascón, T. Lepoint, and M. Raykova, "Secure single-server aggregation with (poly) logarithmic overhead," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2020, pp. 1253–1269.

[194] J. So, R. E. Ali, B. Güler, and A. S. Avestimehr, "Secure aggregation for buffered asynchronous federated learning," 2021, *arXiv:2110.02177*.

[195] C.-S. Yang, J. So, C. He, S. Li, Q. Yu, and S. Avestimehr, "LightSecAgg: Rethinking secure aggregation in federated learning," 2021, *arXiv:2109.14236*.

[196] A. R. Elkordy and A. S. Avestimehr, "Secure aggregation with heterogeneous quantization in federated learning," Tech. Rep., 2020.

[197] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 118–128.

[198] J. So, B. Guler, and A. S. Avestimehr, "Byzantine-resilient secure federated learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2168–2181, Jul. 2020.

[199] L. He, S. P. Karimireddy, and M. Jaggi, "Secure byzantine-robust machine learning," 2020, *arXiv:2006.04747*.

[200] S. Prakash, H. Hashemi, Y. Wang, M. Annavaram, and S. Avestimehr, "Byzantine-resilient federated learning with heterogeneous data distribution," 2020, *arXiv:2010.07541*.

[201] Y. Khazbak, T. Tan, and G. Cao, "MLGuard: Mitigating poisoning attacks in privacy preserving distributed collaborative learning," in *Proc. 29th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Aug. 2020, pp. 1–9.

[202] B. Pejó and G. Biczók, "Quality inference in federated learning with secure aggregation," 2020, *arXiv:2007.06236*.

[203] J. So, R. E. Ali, B. Guler, J. Jiao, and S. Avestimehr, "Securing secure aggregation: Mitigating multi-round privacy leakage in federated learning," 2021, *arXiv:2106.03328*.

[204] K. Banawan, A. Arafa, and S. Ulukus, "Timely private information retrieval," in *Proc. IEEE ISIT*, Jul. 2021.

[205] Q. Yu, M. Maddah-Ali, and S. Avestimehr, "Polynomial codes: An optimal design for high-dimensional coded matrix multiplication," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4406–4416. [Online]. Available: http://papers.neurips.cc

[206] M. Soleymani, R. E. Ali, H. Mahdavifar, and A. S. Avestimehr, "List-decodable coded computing: Breaking the adversarial toleration barrier," 2021, *arXiv:2101.11653*.

[207] M. Soleymani, H. Mahdavifar, and A. S. Avestimehr, "Analog Lagrange coded computing," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 283–295, Mar. 2021.

[208] M. Soleymani, H. Mahdavifar, and A. S. Avestimehr, "Privacy-preserving distributed learning in the analog domain," 2020, *arXiv:2007.08803*.

[209] T. Tang, R. E. Ali, H. Hashemi, T. Gangwani, S. Avestimehr, and M. Annavaram, "Verifiable coded computing: Towards fast, secure and private distributed machine learning," 2021, *arXiv:2107.12958*.

[210] Z. Ghodsi, T. Gu, and S. Garg, "SafetyNets: Verifiable execution of deep neural networks on an untrusted cloud," 2017, *arXiv:1706.10268*.
[211] R. E. Ali, J. So, and A. S. Avestimehr, "On polynomial approximations for privacy-preserving and verifiable ReLU networks," 2020, *arXiv:2011.05530*.
[212] A. M. Subramaniam, A. Heidarzadeh, and K. R. Narayanan, "Collaborative decoding of polynomial codes for distributed computation," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Aug. 2019, pp. 1–5.
[213] T. Jahani-Nezhad and M. A. Maddah-Ali, "Berrut approximated coded computing: Straggler resistance beyond polynomial computing," 2020, *arXiv:2009.08327*.
[214] M. Soleymani, R. E. Ali, H. Mahdavifar, and A. S. Avestimehr, "ApproxIFER: A model-agnostic approach to resilient and robust prediction serving systems," 2021, *arXiv:2109.09868*.

Sennur Ulukus (Fellow, IEEE) received the B.S. and M.S. degrees in electrical and electronics engineering from Bilkent University and the Ph.D. degree in electrical and computer engineering from WINLAB, Rutgers University. She is currently the Anthony Ephremides Professor in Information Sciences and Systems with the Department of Electrical and Computer Engineering, University of Maryland, College Park, where she holds a joint appointment with the Institute for Systems Research (ISR). Prior to joining UMD, she was a Senior Technical Staff Member with AT&T Labs Research. She is also a Distinguished Scholar-Teacher with the University of Maryland. Her research interests are in information theory, wireless communications, machine learning, signal processing, and networks; with recent focus on private information retrieval, age of information, group testing, distributed coded computing, machine learning for wireless, energy harvesting communications, physical layer security, and wireless energy and information transfer. She received the 2003 IEEE Marconi Prize Paper Award in Wireless Communications, the 2019 IEEE Communications Society Best Tutorial Paper Award, the 2020 IEEE Communications Society Women in Communications Engineering (WICE) Outstanding Achievement Award, the 2020 IEEE Communications Society Technical Committee on Green Communications and Computing (TCGCC) Distinguished Technical Achievement Recognition Award, the 2005 NSF CAREER Award, the 2011 ISR Outstanding Systems Engineering Faculty Award, and the 2012 ECE George Corcoran Outstanding Teaching Award. She was a Distinguished Lecturer of the IEEE Information Theory Society from 2018 to 2019.

Salman Avestimehr (Fellow, IEEE) received the B.S. degree in electrical engineering from the Sharif University of Technology in 2003 and the M.S. and Ph.D. degrees in electrical engineering and computer science from the University of California at Berkeley in 2005 and 2008, respectively.

He is currently the Dean's Professor, the Inaugural Director of the USC-Amazon Center on Secure and Trusted Machine Learning (Trusted AI), and the Director of the Information Theory and Machine Learning (vITAL) Research Laboratory, Electrical and Computer Engineering Department, University of Southern California. He is also an Amazon Scholar with Alexa AI. His research interests include information theory, large-scale distributed computing and machine learning, secure and private computing/learning, and federated learning. He has received a number of awards for his research, including the James L. Massey Research & Teaching Award from the IEEE Information Theory Society, the Information Theory Society and Communication Society Joint Paper Award, the Presidential Early Career Award for Scientists and Engineers (PECASE) from the White House (President Obama), the Young Investigator Program (YIP) Award from the U.S. Air Force Office of Scientific Research, the National Science Foundation CAREER Award, the David J. Sakrison Memorial Prize, and several best paper awards at conferences. He has been an Associate Editor of IEEE TRANSACTIONS ON INFORMATION THEORY and the General Co-Chair of the 2020 International Symposium on Information Theory (ISIT).

Michael Gastpar (Fellow, IEEE) received the Dipl.El.-Ing. degree from Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland, in 1997, the M.S. degree in electrical engineering from the University of Illinois at Urbana-Champaign, Urbana, IL, USA, in 1999, and the Doctorat ès Science degree from the École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2002.

He was also a Student in engineering and philosophy with The University of Edinburgh and the University of Lausanne. From 2003 to 2011, he was an Assistant and a tenured Associate Professor with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley. Since 2011, he has been a Professor with the School of Computer and Communication Sciences, EPFL. He was also a Professor with the Delft University of Technology, The Netherlands; and a Researcher with the Mathematics of Communications Department, Bell Labs, Lucent Technologies, Murray Hill, NJ, USA. His research interests are in network information theory and related coding, and signal processing techniques, with applications to sensor networks and neuroscience. He received the IEEE Communications Society and Information Theory Society Joint Paper Award in 2013 and the EPFL Best Thesis Award in 2002. He was an Associate Editor of Shannon Theory for the IEEE TRANSACTIONS ON INFORMATION THEORY from 2008 to 2011 and has served as the Technical Program Committee Co-Chair for the 2010 and 2021 International Symposia on Information Theory (Austin, TX, USA; and Melbourne, Australia). He was an Information Theory Society Distinguished Lecturer from 2009 to 2011.

Syed A. Jafar (Fellow, IEEE) received the B.Tech. degree from IIT Delhi, India, in 1997, the M.S. degree from Caltech, USA, in 1999, and the Ph.D. degree from Stanford University, USA, in 2003, all in electrical engineering.

His industry experience includes positions at Lucent Bell Labs and Qualcomm. He is currently a Chancellor's Professor of electrical engineering and computer science with the University of California at Irvine, Irvine, CA, USA. His research interests include multiuser information theory, wireless communications, and network coding. He was a recipient of the New York Academy of Sciences Blavatnik National Laureate in Physical Sciences and Engineering, the NSF CAREER Award, the ONR Young Investigator Award, the UCI Academic Senate Distinguished Mid-Career Faculty Award for Research, the School of Engineering Mid-Career Excellence in Research Award, and the School of Engineering Maseeh Outstanding Research Award. His co-authored articles have received the IEEE Information Theory Society Paper Award, the IEEE Communication Society and Information Theory Society Joint Paper Award, the IEEE Communications Society Best Tutorial Paper Award, the IEEE Communications Society Heinrich Hertz Award, the IEEE Signal Processing Society Young Author Best Paper Award, and various conference best paper awards. He received the UC Irvine EECS Professor of the Year Award six times from the Engineering Students Council, the School of Engineering Teaching Excellence Award, and the Senior Career Innovation in Teaching Award. He is a Thomson Reuters/Clarivate Analytics Highly Cited Researcher. He served as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS from 2004 to 2009, IEEE COMMUNICATIONS LETTERS from 2008 to 2009, and IEEE TRANSACTIONS ON INFORMATION THEORY from 2009 to 2012. He served as the Technical Program Committee Co-Chair for 2018 ISIT, Vail, CO, USA. He was an University of Canterbury Erskine Fellow in 2010, an IEEE Communications Society Distinguished Lecturer from 2013 to 2014, and an IEEE Information Theory Society Distinguished Lecturer from 2019 to 2020.

Ravi Tandon (Senior Member, IEEE) received the B.Tech. degree in electrical engineering from IIT Kanpur in 2004 and the Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park (UMCP), in 2010. He is currently the Litton Industries John M. Leonis Distinguished Associate Professor with the Department of ECE, The University of Arizona. Prior to joining The University of Arizona in Fall 2015, he was a Research Assistant Professor with Virginia Tech, with positions at the Bradley Department of ECE, the Hume Center for National Security and Technology, and the Discovery Analytics Center, Department of Computer Science. From 2010 to 2012, he was a Post-Doctoral Research Associate with Princeton University. His current research interests include information theory and its applications to wireless networks, communications, security and privacy, machine learning, and data mining. He was a recipient of the 2018 Keysight Early Career Professor Award, the NSF CAREER Award in 2017, and the Best Paper Award at IEEE GLOBECOM 2011. He also serves as an Editor for IEEE TRANSACTIONS ON INFORMATION THEORY, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, and IEEE TRANSACTIONS ON COMMUNICATIONS.

Chao Tian (Senior Member, IEEE) received the B.E. degree in electronic engineering from Tsinghua University, Beijing, China, in 2000, and the M.S. and Ph.D. degrees in electrical and computer engineering from Cornell University, Ithaca, NY, USA, in 2003 and 2005, respectively.

He was a Post-Doctoral Researcher with the École Polytechnique Fédérale de Lausanne (EPFL) from 2005 to 2007; a member of technical staff-research with AT&T Labs Research, NJ, USA, from 2007 to 2014; and an Associate Professor with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, from 2014 to 2017. He joined the Department of Electrical and Computer Engineering, Texas A&M University, in 2017. His research interests include data storage systems, multi-user information theory, joint source-channel coding, signal processing, and compute algorithms. He received the Liu Memorial Award at Cornell University in 2004 and the AT&T Key Contributor Award in 2010, 2011, and 2013. His authored and/or coauthored articles received the 2014 IEEE ComSoc DSTC Data Storage Best Paper Award and the 2017 IEEE Jack Keil Wolf ISIT Student Paper Award. He was an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS from 2012 to 2014 and an Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS from 2016 to 2021. He is currently an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY.