Secure And Private Information Retrieval (SAPIR) in Cloud Storage Systems

Mohsen Karimzadeh Kiskani† and Hamid R. Sadjadpour†

***Abstract*—An information theoretic approach to security and privacy for cloud data storage systems is introduced. The approach called Secure And Private Information Retrieval (SAPIR) is a random combinations of all contents that are stored across the network by using Random Linear Fountain (RLF) codes. SAPIR achieves asymptotic perfect secrecy. To retrieve a content, a group of servers collaborate with each other to form a Repair Group (RG). Further, an information theoretic Private Information Retrieval (PIR) scheme based on random queries is proposed that ensures a user privately downloads its desired content without the servers knowing the requested content index. The proposed scheme is adaptive and can provide privacy against a significant number of colluding servers.**

***Index Terms*—Cloud Storage, Security, PIR**

I. Introduction

Cloud networks have become a popular platform for data storage during the past decade. Security of the stored data has always been a major concern for many cloud service providers. Many cloud service providers use encryption algorithms to encrypt the data on their servers. Dropbox, for instance, is using Advanced Encryption Standard (AES) to store the contents on its servers[1]. Since the encryption algorithms are *computationally secure*, an adversary may be able to break them with time. For instance, Data Encryption Standard (DES) which was once the official Federal Information Processing Standard (FIPS) in US is not considered secure anymore. An interesting problem in highly sensitive cloud services would then be to look for *information theoretic security* solutions.

To achieve perfect information theoretic secrecy using Shannon cipher [1], the number of keys should be equal to the number of messages. Therefore, to utilize Shannon cipher approach, each user needs to store a huge number of keys which is not practical. In this paper, we propose a technique in which the storage capability of the trusted servers are efficiently used to generate the keys by using Random Linear Fountain (RLF) codes [2]. RLF codes have been shown [3]–[7] to be very useful in distributed storage systems. Our goal is to show that RLF codes can provide perfect secrecy and privacy in distributed cloud storage systems.

On the other hand, in many applications like Peer-to-Peer (P2P) distributed storage systems or distributed storage systems in which some of the servers are under the control of an oppressive government, a user wants to download a content from a pool of distributed servers in a way that the servers cannot determine which content is requested by the user. This is widely known as Private Information Retrieval (PIR) problem.

We propose a novel technique to address the PIR problem in distributed storage systems. In our solution, users use random queries to request data from the servers. These random queries are designed in a way that they can be used to retrieve any desired content while preventing any malicious agent with the knowledge of up to half of the random queries to gain information about the requested content. This is an important feature of the proposed technique that provides privacy in the presence of many colluding servers. Such a feature has not been presented in prior information theoretic PIR approaches like [8] for coded storage systems.

SAPIR has strong capabilities such as a) asymptotic perfect secrecy using RLF codes, b) PIR capability using randomly generated queries, and c) PIR resilience against collusion of a large number of agents.

Our proposed solution can be used in future vehicular wireless network systems. One instance of such application is the case that each vehicle stores some portion of the encoded sensitive data and can only recover this data by requesting the rest from the cloud. The legitimate vehicle can retrieve the content while the malicious neighboring vehicles will not be able to obtain any information even if there is a very large number of such vehicles trying to obtain information.

The rest of the paper is organized as follows. Section II is dedicated to the related work on PIR and security in distributed storage systems. The assumptions and problem formulation are described in section III. We study the security and PIR aspects of SAPIR in sections IV and V, respectively. The simulation results are provided in section VI and the paper is concluded in section VII.

II. Related Works

In this paper, we use *Random Linear Fountain (RLF) codes* [2] to encode the contents within the servers in the network. Significant throughput capacity gains [3]–[7] can be achieved using RLF codes in wireless ad hoc and cellular networks. The application of fountain codes in distributed storage systems was also studied in [9]. The capacity of wireless ad hoc networks with caching was computed in [4] and it was shown that RLF codes can achieve asymptotic perfect secrecy. However, the problem of PIR was not studied.

Erasure codes have been extensively used in storage systems. MDS codes are widely used in storage systems due to their repair capabilities [10], [11]. However, certain requirements are needed to secure the applications that use these

M. K. Kiskani† and H. R. Sadjadpour† are with the Department of Electrical Engineering, University of California, Santa Cruz. Email: {mohsen, hamid}@soe.ucsc.edu

[1]https://www.dropbox.com/en/help/27

codes. Authors in [12] also studied the security of distributed storage systems with MDS codes. Pawar et al. [13] studied the secrecy capacity of MDS codes. The authors in [14] and [15] also proposed security measures for MDS coded storage systems. Shah et al. [16] proposed information-theoretic secure regenerating codes for distributed storage systems. Rawat et al. [17] used Gabidulin codes together with MDS codes to propose optimal locally repairable and secure codes. Unlike all of the references [10]–[17], this paper studies the use of RLF codes to attain asymptotic perfect secrecy for distributed storage systems. We have shown in [18] that another group of sparse codes can also achieve asymptotic perfect secrecy.

Kumar et al. [19] have proposed a construction for repairable and secure fountain codes. Reference [19] achieves security by concatenating Gabidulin codes with Repairable Fountain Codes (RFC). Their specific design allows to use Locally Repairable Fountain Codes for secure repair of the lost data. Unlike [19] which has focused on the security of the repair links using concatenated codes, the current paper presents simultaneous security and privacy of the data by only using RLF codes without any additional code usage.

The idea of PIR was originally introduced in [20] for uncoded databases and it has been studied extensively ever since. Yang et al. [21] proposed a PIR algorithm which is information-theoretically robust against many colluding servers. However, in [21], the client reconstruction algorithm is computationally expensive and does not run in polynomial time. Goldberg [22] improved the scheme in [21] and proposed a hybrid PIR scheme based on secret sharing which provides information-theoretic privacy up to a certain number of colluding servers and provides computational privacy if the number of colluding servers increases. The PIR scheme proposed in this paper is different from [21] and [22] as it is able to asymptotically achieve information-theoretic PIR for up to half of the colluding servers with a polynomial time client reconstruction algorithm.

Recently, there has been a renewed interest in studying PIR for storage systems utilizing different coding techniques. Reference [23] proved that with only one extra bit of download, PIR can be achieved with minimum communication cost. The authors in [24] tried to minimized the storage overhead instead of communication cost. They showed that PIR can be achieved with a storage cost that is arbitrarily close to the optimal value of 1. However, the solutions in [23] and [24] require that the number of servers grows with the data record size. Reference [25] assumed that the number of servers is fixed and established the trade-off between storage and retrieval costs and demonstrated the fundamental limits on the cost of PIR for coded storage systems. The authors in [8] introduced a scheme to achieve PIR in MDS coded databases but the security aspect was not addressed in that paper. They have also assumed that the databases are able to store all the contents which may not be a realistic assumption.

The capacity of PIR for a replication coding based storage system without collusion was studied in [26] and later was extended [27] to the case of colluding servers. For coded storage systems, the non-colluding capacity is found in [28]. The capacity of PIR in coded databases with collusion and fixed number of servers was studied in [29].

Reference [30] proposes and evaluates an information-theoretic PIR scheme for MDS coded distributed storage systems where data is stored using a linear systematic code of rate $R > 0.5$. Their proposed scheme is a generalization of the scheme proposed in [8] for the scenario of a single spy node. They proposed an algorithm to minimize the communication price of privacy using the structure of the underlying linear code.

Unlike all of the prior work in [8], [23], [25], [28]–[30] which have only been focused on PIR, we are interested in achieving simultaneous security and PIR. Further, as far as we know, our work is the first work to study the problem of PIR for a fountain coded-based distributed storage system. Reference [31] has studied the problem of join security and privacy in the presence of one eavesdropper but our proposed PIR scheme is easily scalable to the cases when up to half of the servers are colluding to obtain information about the content or content index which makes this technique very robust against large number of colluding servers.

On the other hand, using coding schemes has been shown to be very efficient from a security point of view. Cai and Young [32] showed that network coding can be used to achieve perfect secrecy. Bhattad et al. [33] studied the problem of "weakly secure" network coding schemes in which even without perfect secrecy, no meaningful information can be extracted from the network. Subsequent to [33] Kadhe et al. studied the problem of weakly secure storage systems in [34], [35]. Yan et al. also proposed algorithms to achieve weak security and also studied weakly secure data exchange with generalized Reed Solomon codes in [36].

The proposed privacy technique in this paper is closely related to the searchable encryption scheme. Searchable encryption is a technique which allows to search on encrypted data without performing decryption and allows a client to store documents on a server in encrypted form. Stored documents can be retrieved selectively while revealing as little information as possible to the server [37]. This is in spirit similar to the concept of private information retrieval but they are indeed different techniques for two different problems. Reference [38] studies a searchable encryption scheme with a deterministic algorithm. In [39], the authors implemented dynamic symmetric searchable encryption schemes that efficiently and privately search server-held encrypted databases with very large quantities of record-keyword pairs. In [40], the authors proposed a searchable encryption scheme that achieves both small leakage and efficiency. Note that our proposed method is an information retrieval problem and is different to searchable encryption. Further, our method is based on information theoretic principles.

When using RLF codes in storage systems, the messages are XORed with each other to create the ciphertext. Hence, the ciphertext will not be independent of the message and the Shannon criteria may not be valid. Therefore it may be intuitive to think that these codes can only achieve weak security as opposed to perfect security. In this paper, we will prove that if the number of messages tends to infinity, the intrinsic weak security of these codes can asymptotically result

in perfect security.

III. Problem Formulation

The network is composed of n servers each capable of storing h contents. These servers are denoted by $\mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_n$. A total number of m contents exist within the network and each content has M bits, i.e., $f_1, f_2, \ldots, f_m$. The servers in the network are divided into Repair Groups (RG). Each RG is able to independently respond to the user file requests and perform the task of repairing.

A. RLF Coding-Based Storage

The contents are randomly encoded and stored on the servers during the data preloading phase. The encoded file in the j^{th} storage location of the i^{th} server for any $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, h$ will have the form

$$c_j^i = \sum_{k=1}^{m} v_k^{i,j} f_k = \mathbf{f}\mathbf{v}_j^i, \tag{1}$$

where[2] $\mathbf{f} = [f_1 \ f_2 \ \ldots f_m]$ denotes the $1 \times m$ vector of all contents and $\mathbf{v}_j^i$ denotes the $m \times 1$ random encoding vector of 0s and 1s. Each content f_i belongs to the Galois Field $\mathbb{F}_{2^M}$, i.e. $\mathbf{f} \in \mathbb{F}_{2^M}^m$. Throughout the paper, unless otherwise stated we assume that all the vector and matrix operations are in $\mathbb{F}_2$. The encoded files stored in server $\mathcal{N}_i$ are $\mathbf{c}_i = [c_1^i \ c_2^i \ \ldots c_h^i]$ where $\mathbf{c}_i \in \mathbb{F}_{2^M}^h$. Note that $\mathbf{c}_i = \mathbf{f}\mathbf{V}_i$ where $\mathbf{V}_i$ is the $m \times h$ random encoding matrix for server $\mathcal{N}_i$.

In RLF all random vectors $\mathbf{v}_j^i$ are chosen independently and uniformly from $\mathbb{F}_2^m$ which results in a random uniform choice of the encoding matrix $\mathbf{V}_i$ where each element can be either 0 or 1 with equal probability. Such an encoding matrix may not necessarily be full rank and may contain linearly dependent rows. This will result in redundant use of storage and may jeopardize the security by revealing more information. Hence, we propose a *full rank encoding* scheme based on RLF codes in which randomly created encoding vectors $\mathbf{v}_j^i$ are discarded if they already exist in the span of the previously selected random vectors. In other words, for each server we select h linearly independent vectors to construct a full rank matrix $\mathbf{V}_i$ of size $m \times h$ for $i = 1, 2, \ldots, n$.

B. Repair Groups (RG)

After the data preloading phase, users can reconstruct their desired contents during content delivery phase. We assume that servers are divided into different *Repair Groups (RGs)*. Each RG is able to independently respond to individual file requests and contains m linearly independent encoded files within all of its servers. Servers within each RG collaborate with each other to retrieve any requested content. The RGs are represented by $\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_u$ and the number of servers within their corresponding RGs by $J_1, J_2, \ldots, J_u$ where, $\sum_{i=1}^{u} J_i = n$.

A desired file f_r can be written as $f_r = \mathbf{f}\mathbf{e}_r$, where $\mathbf{e}_r$ is an all zero $m \times 1$ vector except in the r^{th} location. To retrieve f_r, the user needs to access enough encoded files on the network servers in order to construct $\mathbf{e}_r$ via $\mathbf{v}_j^i$'s. Since codes are constructed in $\mathbb{F}_2^m$, users need m linearly independent encoding vectors to retrieve any of the m contents.

As shown in [5], the average minimum number of RLF codes required to retrieve all the contents is only slightly larger than m. Therefore, for each RG $\mathcal{J}_i$ where $1 \leq i \leq u$, the minimum value of J_i is only slightly larger than $\frac{m}{h}$. Notice that if J_i is smaller than $\frac{m}{h}$, then the servers will not be able to form a full rank matrix to retrieve all desired contents. Note that with exactly m caches, we are able to store m uncoded files which is very close to the proposed RLF technique and demonstrates that RLF-coding based approach efficiently utilizes storage space. For large values of h, i.e. $h \geq m$, each server can become an RG by itself.

C. Content Retrieval

When a user requests to download a file f_r, its request is processed by one of the RGs. This could potentially be the closest RG or the RG that has the minimum load on its servers. The user then sends its request to one of the servers in that RG. We call this server the *coordinating server* and we denote it by $\mathcal{N}_s$.

The goal is to achieve security without encryption. Therefore, we assume that none of the servers uses encryption to store contents. We will show that even with that setting the last hop communication between the coordinating server and the user can be done with asymptotic perfect secrecy.

To achieve secrecy we assume that prior to any communications the coordinating server creates a vector of encoded files $\mathbf{c}_u = \mathbf{f}\mathbf{V}_u$ and send it through a secure channel to the user. This data is then stored on the user and operates similar to the key in Shannon cipher system and it is only sent once. The user will use it over and over to securely download its requested contents. We assume that this part of the data is transmitted to the user using a secure low bandwidth channel. It is assumed that no eavesdropper has this data and it is unique to the user.

Assume that the user requests file f_r from the coordinating server $\mathcal{N}_s$ in RG $\mathcal{J}_k$. RG $\mathcal{J}_k$ has stored $J_k h \geq m - h_u$ randomly encoded files. The matrices $\mathbf{V}_i$ of the J_k servers in the RG form a full rank matrix $\mathbf{V} = [\mathbf{V}_1 \ \mathbf{V}_2 \ \ldots \ \mathbf{V}_{J_k} \ \mathbf{V}_u]_{m \times (J_k h + h_u)}$. Therefore, any content with index r can be retrieved from these servers by solving the linear equation $\mathbf{V}\mathbf{y}_r = \mathbf{e}_r$ in $\mathbb{F}_2$. Since this matrix is full rank, one possible solution can be given as

$$\mathbf{y}_r = \mathbf{V}^T \left(\mathbf{V}\mathbf{V}^T\right)^{-1} \mathbf{e}_r. \tag{2}$$

To solve $\mathbf{V}\mathbf{y}_r = \mathbf{e}_r$, servers within the RG send their corresponding encoding matrices $\mathbf{V}_i$ to server $\mathcal{N}_s$ through low bandwidth secure channels. Server $\mathcal{N}_s$ then creates $\mathbf{V}$ and computes $\mathbf{y}_r$ from the above equation[3]. If $\mathbf{y}_r = [\mathbf{y}_r^1 \ \mathbf{y}_r^2 \ \ldots \ \mathbf{y}_r^{J_k} \ \mathbf{y}_r^u]^T$ is such a solution, where $\mathbf{y}_r^i$ is a $h \times 1$ local decoding vector for server $\mathcal{N}_i$ and $\mathbf{y}_r^u$ is the $h_u \times 1$ user decoding vector, then server $\mathcal{N}_s$ sends $\mathbf{y}_r^i$ to server $\mathcal{N}_i$ and sends $\mathbf{y}_r^u$ down to the user as shown in Figure 1. Server $\mathcal{N}_i$ then sequentially creates $\mathbf{f}\mathbf{V}_i\mathbf{y}_r^i$ and adds it to the previous

[2]Throughout the paper, the vectors are denoted in bold characters.

[3]Notice that the servers of an RG only need to send this information to $\mathcal{N}_s$ once. This could be done even right after the data preloading phase.

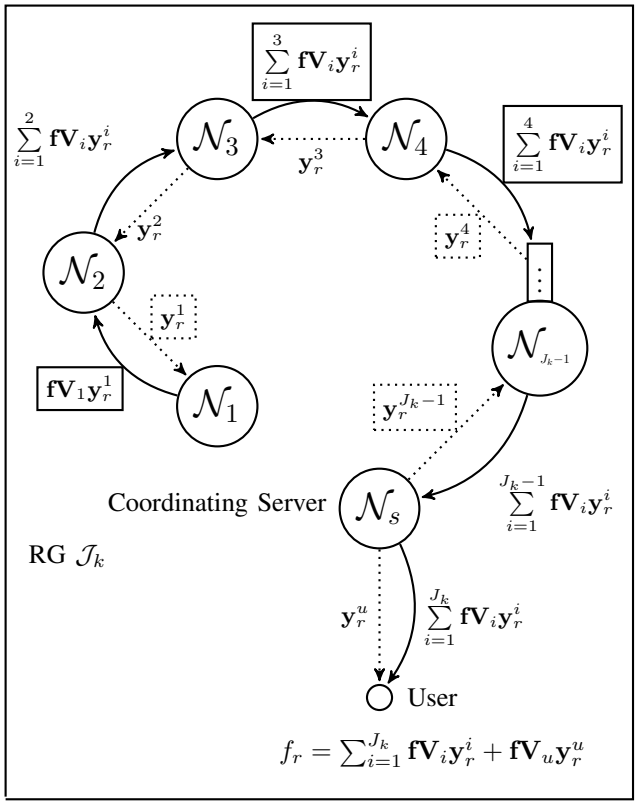

Fig. 1: Diagram of the communications inside RG $\mathcal{J}_k$ when a user requests file f_r. All servers send their data to the coordinating server $\mathcal{N}_s$ using a sequential protocol. The server $\mathcal{N}_s$ then relays all of this data to the user. Dotted lines represent transmissions over secure channel.

encoded file that it has received from previous servers. All of the server responses are then aggregated by the coordinating server $\mathcal{N}_s$ and $\mathcal{N}_s$ will send $\sum_{i=1}^{J_k} \mathbf{f}\mathbf{V}_i\mathbf{y}_r^i$ to the user. The user then adds this to a linear combination of its own data to reconstruct f_r as

$$f_r = \mathbf{f}\mathbf{e}_r = \mathbf{f}\mathbf{V}\mathbf{y}_r = \sum_{i=1}^{J_k} \mathbf{f}\mathbf{V}_i\mathbf{y}_r^i + \mathbf{f}\mathbf{V}_u\mathbf{y}_r^u. \tag{3}$$

We will show that with the above sequential tranmission protocol between the servers, asymptotic perfect secrecy is achievable. This solution reveals the index of the downloaded content to the servers of the RG. This simple solution cannot be used for PIR but we will show in section IV that perfect secrecy can be achieved. A solution to preserve the privacy of the users is presented in section V.

IV. Security

This section is dedicated to the study of security of our approach. We assume that the low bandwidth communication between the users and the RGs is done over secure channels. Therefore, the adversary will not be able to wiretap the information sent over the secure link between the user and the coordinating server. On the other hand, we assume that the servers are not using any sort of encryption algorithm and yet perfect secrecy can be achieved over the high bandwidth communication links within a specific RG.

Perfect secrecy was originally introduced by Shannon in [1]. Assume that a transmitter wants to secretly send a message $\mathfrak{M} \in \mathbb{M}$ to the receiver. The transmitter chooses a key $\mathfrak{K} \in \mathbb{K}$ which is independent of message and encodes the message with an encoding function $\mathfrak{e} : \mathbb{M} \times \mathbb{K} \to \mathbb{C}$ and sends the codeword $\mathfrak{C} = e(\mathfrak{M}, \mathfrak{K}) \in \mathbb{C}$ over the channel. The legitimate receiver then uses a decoding function $\mathfrak{d} : \mathbb{C} \times \mathbb{K} \to \mathbb{M}$ to decode the message as $\mathfrak{M} = \mathfrak{d}(\mathfrak{C}, \mathfrak{K})$.

Definition 1. An encoding scheme is said to achieve perfect secrecy if $H(\mathfrak{M}|\mathfrak{C}) = H(\mathfrak{M})$, or equivalently $I(\mathfrak{M}; \mathfrak{C}) = 0$.

We will prove that asymptotic perfect secrecy can be achieved between the coordinating server and the user. We define the idea of *asymptotic perfect secrecy* as follows.

Definition 2. An encoding scheme is said to achieve asymptotic perfect secrecy if it can achieve perfect secrecy for the communication link between the coordinating server and the user when the number of files tends to infinity.

Notice that if the number of the messages is finite, we cannot achieve perfect secrecy with our approach. That scenario will result in a weakly secure scheme [33]. Our contribution is that the weakly secure scheme will result in perfect security when the number of messages goes to infinity.

Assume that the user sends the request $\mathbf{e}_r$ to the coordinating server $\mathcal{N}_s$ in RG $\mathcal{J}_k$ to download f_r. Under this scenario, the adversary knows the requested content index but we will prove that it is still unable to reduce its equivocation about the requested content. We assume that an adversary can wiretap any of the high bandwidth links between servers.

When the query $\mathbf{e}_r$ is received by the coordinating server $\mathcal{N}_s$, it uses the matrix $\mathbf{V}$ to solve the linear equation $\mathbf{V}\mathbf{y}_r = \mathbf{e}_r$ to find the decoding vector $\mathbf{y}_r$ and then breaks $\mathbf{y}_r$ into local decoding gains $\mathbf{y}_r^i$ and send them to other servers[4]. Then, based on the proposed sequential protocol depicted in Figure 1, starting from server $\mathcal{N}_1$ each server $\mathcal{N}_i$ multiplies the decoding gain $\mathbf{y}_r^i$ that it has received from the coordinating server to its locally stored information $\mathbf{f}\mathbf{V}_i$ to get $\mathbf{f}\mathbf{V}_i\mathbf{y}_r^i$ and then it XORs it to the information that it has received from the previous server and relays it to the next server until it reaches the coordinating server $\mathcal{N}_s$ which then transmits it to the user.

In this paper, we only study the security of the last communication link between the coordinating server $\mathcal{N}_s$ and the user. Notice that this is the most vulnerable communication link since it includes all the aggregated data from all the servers. A similar approach can be used for all other links and prove that perfect secrecy can be achieved in any of other links. We assume that the eavesdropper wiretaps the link between $\mathcal{N}_s$ and the user. Therefore, using equation (3), the first part of this equation is known to the eavesdropper while the second

[4]Note that the information in matrix $\mathbf{V}$ has previously been communicated to the coordinating server. Also, notice that the size of the decoding information $\mathbf{y}_r^i$ and the encoding matrix $\mathbf{V}$ is much smaller than the size of the file chunks.

part is a secret. Let's define

$$S_r \triangleq \sum_{i=1}^{J_k} \mathbf{f}\mathbf{V}_i \mathbf{y}_r^i, \tag{4}$$

$$T_r \triangleq \mathbf{f}\mathbf{V}_u \mathbf{y}_r^u. \tag{5}$$

Then the requested content can be written as $f_r = S_r + T_r$ and since all operations are in $\mathbb{F}_2$, we have

$$S_r = f_r + T_r. \tag{6}$$

This is similar to the Shannon cipher system [1] in which f_r, T_r, and S_r can be regarded as the message, key, and codeword respectively. The eavesdropper knows the encoded file S_r but it cannot obtain any information about the message f_r if a unique key T_r with uniform distribution is used for each message.

Theorem 1 provides the necessary and sufficient condition to obtain perfect secrecy and the proof can be found in [41].

Theorem 1. If $|\mathbb{M}| = |\mathbb{K}| = |\mathbb{C}|$, a coding scheme achieves perfect secrecy if and only if for each pair $(\mathfrak{M}, \mathfrak{C}) \in (\mathbb{M} \times \mathbb{C})$, there exists a unique key $\mathfrak{K} \in \mathbb{K}$ such that $\mathfrak{C} = \mathfrak{e}(\mathfrak{M}, \mathfrak{K})$ and the key $\mathfrak{K}$ is uniformly distributed in $\mathbb{K}$.

To use this theorem, first we prove that for large enough values of m, the key T_r is uniformly distributed.

Lemma 1. The asymptotic distribution of bits of coded files stored on the user tends to uniform.

Proof. All files have M bits and they may have a distribution different from uniform. We will prove that each randomly encoded file will be uniformly distributed for large values of m. Let us denote the k^{th} bit of file f_l by f_l^k where $1 \leq k \leq M$ and $1 \leq l \leq m$. Assume that $\mathbb{P}(f_l^k = 1) = p_l^k = 1 - \mathbb{P}(f_l^k = 0)$. Further, we assume that the bits of files (f_l^k) are independent. The k^{th} bit of the coded file in the j^{th} coded file of the user can be represented as $c_{u,j}^k = \sum_{l=1}^m v_l^{u,j} f_l^k$. $v_l^{i,j}$ is a binary value with uniform distribution and independent of all other bits. Using regular summation (not over $\mathbb{F}_2$) and denoting $\tau_l^{j,k} \triangleq v_l^{u,j} f_l^k$, we define $H^{j,k} \triangleq \sum_{l=1}^m \tau_l^{j,k}$. Therefore, $\mathbb{P}[c_{u,j}^k = 0] = \mathbb{P}[H^{j,k} \overset{2}{\equiv} 0]$. Therefore, the k^{th} bit of the coded file is equal to 0 if an even number of terms in $H^{j,k}$ is equal to 1. The probability distribution of $H^{j,k}$ can be computed using *probability generating functions*. Since $\tau_l^{j,k}$ is a Bernoulli random variable with probability $\frac{1}{2}p_l^k$, its probability generating function is equal to

$$G_l^{j,k}(z) = \left(1 - \frac{1}{2}p_l^k\right) + \frac{1}{2}p_l^k z. \tag{7}$$

Since $v_l^{u,j}$ and f_l^k are independent random variables, $\tau_l^{j,k}$ will become independent random variables. Therefore, the probability generating function of $H^{j,k}$ denoted by $G_H^{j,k}(z)$ is the product of all probability generating functions.

$$G_H^{j,k}(z) = \prod_{l=1}^m \left(\left(1 - \frac{1}{2}p_l^k\right) + \frac{1}{2}p_l^k z\right) \tag{8}$$

Denoting the probability distribution of $H^{j,k}$ as $\mathfrak{h}(.)$, the probability of $H^{j,k}$ being even is

$$\mathbb{P}[H^{j,k} \overset{2}{\equiv} 0] = \sum_{u=0}^{\lfloor \frac{m}{2} \rfloor} \mathfrak{h}(2u) = \sum_{u=0}^{\lfloor \frac{m}{2} \rfloor} \mathfrak{h}(2u) z^{2u} \Big|_{z=1}$$

$$= \frac{1}{2}\left[\sum_{u=0}^{m} \mathfrak{h}(u) z^u + \sum_{u=0}^{m} \mathfrak{h}(u)(-z)^u\right]_{z=1}$$

$$= \frac{1}{2}G_H^{j,k}(1) + \frac{1}{2}G_H^{j,k}(-1) = \frac{1}{2}\prod_{l=1}^m \left(\left(1 - \frac{1}{2}p_l^k\right) + \frac{1}{2}p_l^k\right)$$

$$+ \frac{1}{2}\prod_{l=1}^m \left(\left(1 - \frac{1}{2}p_l^k\right) - \frac{1}{2}p_l^k\right) = \frac{1}{2}\left(1 + \prod_{l=1}^m \left(1 - p_l^k\right)\right)$$

Therefore,

$$\lim_{m\to\infty} \mathbb{P}[c_{u,j}^k = 0] = \lim_{m\to\infty} \frac{1}{2}\left(1 + \prod_{l=1}^m \left(1 - p_l^k\right)\right) = \frac{1}{2} +$$

$$\frac{1}{2}\lim_{m\to\infty} \prod_{l=1}^m \left(1 - p_l^k\right) = \frac{1}{2} + \frac{1}{2}\lim_{m\to\infty} \left(1 - \inf\{p_l^k\}\right)^m = \frac{1}{2}.$$

This proves the lemma. □

This lemma paves the way to prove the following theorem.

Theorem 2. For the proposed full rank encoding scheme if m is large but $m < 2^{h_u}$, then the proposed encoded strategy provides asymptotic perfect secrecy against any eavesdropper wiretapping the last hop communication link between the coordinating server $\mathcal{N}_s$ and the user.

Proof. We formulated this problem as a Shannon cipher system assuming that $\mathfrak{M} = f_r$, $\mathfrak{K} = T_r$, and $\mathfrak{C} = S_r$. The condition $m < 2^{h_u}$ ensures that a unique vector $\mathbf{y}_r^u$ exists for each requested message. Therefore, since full rank encoding scheme is used, then $\mathbf{V}_u$ will be full rank and T_r guarantees that a unique key exists for each requested message f_r. Notice that if the size of the RG is large enough, then the unique choice of the key does not affect the solvability of the linear equation $\mathbf{V}\mathbf{y}_r = \mathbf{e}_r$. Therefore, for any pair $(\mathfrak{m}, \mathfrak{C}) \in (\mathbb{M}, \mathbb{C})$, a unique key $\mathfrak{K} \in \mathbb{K}$ exists such that $\mathfrak{C} = \mathfrak{m} + \mathfrak{K}$. Further, we are guaranteed to have $|\mathbb{M}| = |\mathbb{K}| = |\mathbb{C}|$.

Notice that the key $\mathfrak{K} = T_r$ belongs to the set of all possible bit strings with M bits. Lemma 1 proves that each encoded file is uniformly distributed among all M-bit strings. Hence each key which is a unique summation of such encoded files is uniformly distributed among the set of all M-bit strings. In other words, regardless of the distribution of the bits in files, T_r can be any bit string with equal probability for large values of m. Therefore, the conditions in Theorem 1 are met. □

Remark 1. Note that the key $\mathbf{c}_u = \mathbf{f}\mathbf{V}_u$ is stored on the user during the data preloading phase securely. Therefore, the eavesdropper does not have any knowledge about the key T_r.

V. Private Information Retrieval

In PIR, the goal is to provide conditions that when a user downloads the content f_r with index $r \in \{1, 2, \ldots, m\}$, the content index remains a secret to all of the servers. To achieve PIR, users send queries to the servers and servers respond to users based on those queries. These queries should be

designed in a way that reveal no information to the servers about the requested content index. To formally define the *information theoretic PIR*, let R be a random variable denoting the requested content index and let $\mathcal{Q}_l$ be a subset of at most l queries. We have the following definition.

Definition 3. A PIR scheme is capable of achieving perfect information theoretic PIR against i colluding servers if for the set $\mathcal{Q}_l$ of all queries available to all of these servers and any number of contents we have $I(R;\mathcal{Q}_l)=0$, where $I(.)$ is the mutual information function.

A. Random Query Generation

To achieve PIR, the user chooses a fixed $\epsilon>0$ and sets $A^\epsilon \triangleq m+\lceil \log_2(\frac{1}{\epsilon})\rceil$. Then it picks A^ϵ query vectors from $\mathbb{F}_2^m$ uniformly at random and statistically independent of each other. These will be the set of random queries. Therefore, we will have a set $\mathcal{Q}^\epsilon=\{\mathbf{q}_1,\mathbf{q}_2,\dots,\mathbf{q}_{A^\epsilon}\}$ of i.i.d. random query vectors. In the following, we will prove that with a probability of at least $1-\epsilon$, these random vectors span the whole m-dimensional space of $\mathbb{F}_2^m$.

Theorem 3. Let $\mathbf{Q}$ be a matrix of size $m\times l$ whose elements are independent random variables taking the values 0 and 1 with equal probability and let $\rho_m(l)$ be the rank of the matrix $\mathbf{Q}$ in $\mathbb{F}_2$. Let $s\geq 0$ and c be fixed integers, $c+s\geq 0$. If $m\to\infty$ and $l=m+c$, then

$$\mathbb{P}[\rho_m(l)=m-s] \to 2^{-s(s+c)}\prod_{i=s+1}^{\infty}\left(1-\frac{1}{2^i}\right)\prod_{j=1}^{s+c}\left(1-\frac{1}{2^j}\right)^{-1}, \quad (9)$$

where the last product equals 1 for $c+s=0$.

Proof. This is Theorem 3.2.1 in page 126 of [42]. □

Corollary 1. For $l=m+c$ where $c\geq 0$, if $m\to\infty$ we have

$$\mathbb{P}[\rho_m(l)=m]\to\prod_{i=c+1}^{\infty}\left(1-\frac{1}{2^i}\right) \quad (10)$$

Proof. The proof follows for $s=0$ in Theorem 3. □

In the following, we will use these results for our proofs.

Definition 4. We define the random variable A as the minimum number of random query vectors $\mathbf{q}_1,\mathbf{q}_2,\dots,\mathbf{q}_i$ to span the whole space of $\mathbb{F}_2^m$.

Lemma 2. The probability of the event that $A<m$ is zero and for any $c\geq 0$ we have

$$\mathbb{P}[A\leq m+c]\to\prod_{i=c+1}^{\infty}\left(1-\frac{1}{2^i}\right) \quad (11)$$

Proof. This is a direct result of Corollary 1. □

Lemma 3. The probability of the event that $A=m+c$ is less than 2^{-c} for any $c\geq 0$.

Proof. Let $F(c)\triangleq\mathbb{P}[A\leq m+c]$. It is easy to verify from equation (11) that for $m\to\infty$ we have

$$F(c)\to\frac{F(c-1)}{1-\frac{1}{2^c}}. \quad (12)$$

Since $F(c)\leq 1$, from equation (12) we arrive at

$$F(c-1)\leq 1-2^{-c}. \quad (13)$$

Hence,

$$\begin{aligned}\mathbb{P}[A=m+c]&=F(c)-F(c-1) \quad (14)\\ &\to F(c-1)\left(\frac{1}{1-\frac{1}{2^c}}-1\right)=F(c-1)\left(\frac{1}{1-\frac{1}{2^c}}-1\right)\\ &=\frac{F(c-1)}{2^c-1}\leq\frac{1-2^{-c}}{2^c-1}=2^{-c}\end{aligned}$$

□

Lemma 4. The probability of the event that $A\leq m+c$ is at least $1-2^{-c}$ and at most $1-2^{-(c+1)}$ for any $c\geq 0$. i.e.

$$1-2^{-c}\leq F(c)\leq 1-2^{-(c+1)} \quad (15)$$

Proof. The upper bound is already proved in equation (13). From Lemma 3 we have,

$$\begin{aligned}F(c)&=\mathbb{P}[A\leq m+c]=1-\mathbb{P}[A>m+c]\\ &=1-\sum_{i=c+1}^{\infty}\mathbb{P}[A=m+i]\geq 1-\sum_{i=c+1}^{\infty}2^{-i}=1-2^{-c}\end{aligned}$$

□

Theorem 4. With a probability of at least $1-\epsilon$, the set of random queries $\mathcal{Q}^\epsilon=\{\mathbf{q}_1,\mathbf{q}_2,\dots,\mathbf{q}_{A^\epsilon}\}$ where $A^\epsilon=m+\lceil\log_2(\frac{1}{\epsilon})\rceil$ spans the whole m-dimensional space of $\mathbb{F}_2^m$.

Proof. From Lemma 4, we have

$$\mathbb{P}\left[A\leq A^\epsilon=m+\lceil\log_2(\frac{1}{\epsilon})\rceil\right]\geq 1-2^{-\lceil\log_2(\frac{1}{\epsilon})\rceil}\geq 1-\epsilon$$

This proves the theorem. □

Theorem 4 states that the probability of spanning the m-dimensional space can arbitrarily go to 1 provided that the number of random vectors increases logarithmically with $\frac{1}{\epsilon}$. For example, to span the m-dimensional space with a probability of at least 0.99, it is enough to only have $m+7$ random vectors. Using these random query vectors, we can now show that even with a large number of colluding servers no information about the requested content index can be obtained. To prove this result, we need to prove some lemmas.

Let $\mathbf{Q}^\epsilon\triangleq[\mathbf{q}_1\ \mathbf{q}_2\ \dots\ \mathbf{q}_{A^\epsilon}]$ be the matrix of size $m\times A^\epsilon$ whose columns are random query vectors. Matrix $\mathbf{Q}^\epsilon$ contains A^ϵ statistically independent random vectors. Let $B_{\mathbf{x}}^r$ be the event that for a specific vector $\mathbf{x}\in\mathbb{F}_2^{A^\epsilon}$ and a specific base vector $\mathbf{e}_r$, we have $\mathbf{Q}^\epsilon\mathbf{x}=\mathbf{e}_r$.

Lemma 5. For any specific non-zero vector $\mathbf{x}\in\mathbb{F}_2^{A^\epsilon}$ we have

$$\mathbb{P}[B_{\mathbf{x}}^r]=\mathbb{P}[\mathbf{Q}^\epsilon\mathbf{x}=\mathbf{e}_r]=2^{-m}. \quad (16)$$

Proof. Lets assume vector $\mathbf{x}$ has k ones. If $\mathbf{Q}^\epsilon\mathbf{x} = \mathbf{e}_r$, then k vectors from the set of all vectors $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{A^\epsilon}$ are added together to create $\mathbf{e}_r$. Let's denote these vectors by $\mathbf{q}_{e_1}, \mathbf{q}_{e_2}, \ldots, \mathbf{q}_{e_k}$. Let $q_r^{e_j}$ denotes the r^{th} element of vector $\mathbf{q}_{e_j}$. Since the vectors $\mathbf{q}_{e_1}, \mathbf{q}_{e_2}, \ldots, \mathbf{q}_{e_k}$ are independent and their elements are also mutually independent, using binary summations in $\mathbb{F}_2$ we have

$$\mathbb{P}\left[B_{\mathbf{x}}^r\right] = \mathbb{P}\left[\sum_{j=1}^{k} q_r^{e_j} = 1\right] \prod_{\substack{l'=1 \\ l' \neq r}}^{m} \mathbb{P}\left[\sum_{j=1}^{k} q_{l'}^{e_j} = 0\right]. \tag{17}$$

We can easily prove that $\mathbb{P}[\sum_{j=1}^{k} q_r^{e_j} = 1] = \frac{1}{2}$. To prove this, we can use induction on k. This equation is valid for the base case $k = 1$. Assume that it is valid for $k-1$. We have

$$\mathbb{P}\left[\sum_{j=1}^{k} q_r^{e_j} = 1\right] = \mathbb{P}\left[q_r^{e_k} = 1\right]\mathbb{P}\left[\sum_{j=1}^{k-1} q_r^{e_j} = 0\right] + \mathbb{P}\left[q_r^{e_k} = 0\right]\mathbb{P}\left[\sum_{j=1}^{k-1} q_r^{e_j} = 1\right] = \frac{1}{2}.$$

Similarly, it is easy to prove that $\mathbb{P}[\sum_{j=1}^{k} q_{l'}^{e_j} = 0] = \frac{1}{2}$. Hence, equation (17) can be simplified to $\mathbb{P}[B_{\mathbf{x}}^r] = \mathbb{P}[\mathbf{Q}^\epsilon\mathbf{x} = \mathbf{e}_r] = 2^{-m}$. □

Lemma 6. The following inequalities hold for $1 \leq j \leq i$,

$$\frac{1}{i+1} 2^{iH(\frac{j}{i})} \leq \binom{i}{j} \leq 2^{iH(\frac{j}{i})} \tag{18}$$

where $H(\alpha)$ denotes the binary entropy function, i.e. $H(\alpha) = -\alpha\log_2(\alpha) - (1-\alpha)\log_2(1-\alpha)$.

Proof. The proof can be found in the appendix of [43]. □

We are now ready to prove the following theorem which shows that accessing a significant number of random queries in $\mathcal{Q}^\epsilon$ cannot help in reconstructing any of the base vectors for large m.

Theorem 5. Consider the set $\mathcal{Q}^\epsilon = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{A^\epsilon}\}$ of $A^\epsilon = m + \lceil \log_2(\frac{1}{\epsilon}) \rceil$ statistically independent random uniform query vectors. For large enough values of m with probability arbitrarily close to 1, none of the base vectors exist in the span of any subset $\mathcal{Q}_l \subset \mathcal{Q}^\epsilon$ with cardinality of at most $l = \lfloor \delta m \rfloor$ where $\delta < 0.5$.

Proof. Consider any base vector $\mathbf{e}_r$ and a non-zero vector $\mathbf{x} \in \mathbb{F}_2^{A^\epsilon}$. For this vector, computing $\mathbf{Q}^\epsilon\mathbf{x}$ in $\mathbb{F}_2$ is equivalent to adding a subset of columns of $\mathbf{Q}^\epsilon$ whose set of indices is equal to the set of indices of non-zero elements in $\mathbf{x}$. If for some $\mathbf{x} \in \mathbb{F}_2^{A^\epsilon}$ we have $\mathbf{Q}^\epsilon\mathbf{x} = \mathbf{e}_r$, then any subset $\mathcal{Q}_l \subset \mathcal{Q}^\epsilon$ of random vectors which contains all of those column vectors of $\mathbf{Q}^\epsilon$ whose set of indices is equal to the set of indices of non-zero elements in $\mathbf{x}$, also spans $\mathbf{e}_r$. In fact, the number of non-zero elements of $\mathbf{x}$ or Hamming weight of $\mathbf{x}$ (i.e., Ham($\mathbf{x}$)) is equal to the number of vectors that should be added to reconstruct $\mathbf{e}_r$.

Consider all vectors $\mathbf{x} \in \mathbb{F}_2^{A^\epsilon}$ with Hamming weight less than or equal to $l = \lfloor \delta m \rfloor$ where $\delta < 0.5$. Lemma 5 shows that for any $\mathbf{x}$, we have $\mathbb{P}[B_{\mathbf{x}}^r] = 2^{-m}$. Therefore, the asymptotic probability of existence of a subset $\mathcal{Q}_l \subset \mathcal{Q}^\epsilon$ with a cardinality of at most $l = \lfloor \delta m \rfloor$ which spans $\mathbf{e}_r$ for large values of m can be found as

$$\begin{aligned}
&\lim_{m\to\infty} \mathbb{P}\left[\exists \mathcal{Q}_l \subseteq \mathcal{Q}^\epsilon | \text{card}\{\mathcal{Q}_l\} \leq l = \lfloor \delta m \rfloor, \mathbf{e}_r \in \text{span}\{\mathcal{Q}_l\}\right] \\
&= \\
&\lim_{m\to\infty} \mathbb{P}\left[\bigcup_{\mathbf{x}\in\mathbb{F}_2^{A^\epsilon},\ \text{Ham}(\mathbf{x})\leq l} B_{\mathbf{x}}^r\right] \overset{(a)}{\leq} \lim_{m\to\infty} \sum_{\mathbf{x}\in\mathbb{F}_2^{A^\epsilon},\ \text{Ham}(\mathbf{x})\leq l} \mathbb{P}[B_{\mathbf{x}}^r] \\
&\overset{(b)}{=} \lim_{m\to\infty} \sum_{i=1}^{l} \binom{A^\epsilon}{i} 2^{-m} \overset{(c)}{\leq} \lim_{m\to\infty} l\binom{A^\epsilon}{l} 2^{-m}, \\
&\overset{(d)}{\leq} \lim_{m\to\infty} l\binom{m}{l} 2^{-m} = \lim_{m\to\infty} \lfloor \delta m \rfloor \binom{m}{\lfloor \delta m \rfloor} 2^{-m}, \\
&\overset{(e)}{\leq} \lim_{m\to\infty} \delta m 2^{mH(\frac{\lfloor \delta m \rfloor}{m})} 2^{-m} \overset{(f)}{\leq} \lim_{m\to\infty} \delta m 2^{-m(1-H(\delta))}, \\
&\overset{(g)}{=} 0,
\end{aligned}$$

where inequality (a) comes from the *union bound*, (b) holds by using Lemma 5 and counting all the vectors $\mathbf{x}$ with Hamming weight less than $l = \lfloor \delta m \rfloor$, and inequality (e) comes from Lemma 6. Notice that (c), (d), (e) and (f) are only valid for $\delta < 0.5$. Therefore, the probability of existence of any desired base in the span of any subset of vectors with cardinality less than $\lfloor \delta m \rfloor$ goes to zero as m grows if $\delta < 0.5$. □

The proof of Theorem 5 is critically dependent on the assumption that $\delta < 0.5$. Using a similar approach, one can show that if $\delta > 0.5$, then the probability of existence of at least one base in the span of the random vectors goes to 1 as $m \to \infty$. Further, it is also possible to prove that if $\delta > 1$, then all of the base vectors exist in the span of the random vectors with probability 1 as $m \to \infty$.

In the following theorem we will use the result of Theorem 5 to prove that accessing a large number of queries cannot reveal any information about the requested content index.

Theorem 6. For every subset $\mathcal{Q}_l \subset \mathcal{Q}^\epsilon$ with cardinality at most $l = \lfloor \delta m \rfloor$ where $\delta < 0.5$ we have

$$\lim_{m\to\infty} I(R\ ;\ \mathcal{Q}_l) = 0. \tag{19}$$

Proof. Let D_l be the event that none of the base vectors exist in the span of any subset $\mathcal{Q}_l \subset \mathcal{Q}^\epsilon$ with cardinality at most l. Let $\mathbf{e}_R$ be the equivalent random base vector which is uniquely defined by the requested content index R. If D_l happens, then for every subset $\mathcal{Q}_l \subset \mathcal{Q}^\epsilon$ there should exist $k_l \geq 1$ random vectors $\mathbf{q}_{o_1}, \mathbf{q}_{o_2}, \ldots, \mathbf{q}_{o_{k_l}} \in \mathcal{Q}^\epsilon - \mathcal{Q}_l$ such that $\mathbf{e}_R \notin \text{span}\{\mathcal{Q}_l\}$ but $\mathbf{e}_R \in \text{span}\{\mathcal{Q}_l \cup \{\mathbf{q}_{o_i}, \mathbf{q}_{o_2}, \ldots, \mathbf{q}_{o_{k_l}}\}\}$. Hence, for any $\mathcal{Q}_l \subset \mathcal{Q}^\epsilon$, any representation of $\mathbf{e}_R$ in terms of random queries should have a form of

$$\mathbf{e}_R = \sum_{\mathbf{q}_j \in \mathcal{Q}_l} \theta_j \mathbf{q}_j + \sum_{i=1}^{k_l} \mathbf{q}_{o_i} \tag{20}$$

where θ_i's are equal to zero or one. Hence, for every $\mathcal{Q}_l \subset \mathcal{Q}^\epsilon$

we should have

$$I(R;\mathcal{Q}_l|D_l) = H(R|D_l) - H(R|\mathcal{Q}_l, D_l) \overset{(a)}{=} H(R|D_l) - H(R|D_l) = 0 \quad (21)$$

where (a) comes from the independence of the content index (or its equivalent base vector $\mathbf{e}_R$) and the random queries. Using the chain rule for mutual information, we arrive at

$$I(R;\mathcal{Q}_l, D_l) = I(R;D_l) + I(R;\mathcal{Q}_l|D_l) = I(R;\mathcal{Q}_l) + I(R;D_l|\mathcal{Q}_l). \quad (22)$$

Combining equations (21) and (22) results in

$$\begin{aligned} I(R;\mathcal{Q}_l) &= I(R;D_l) - I(R;D_l|\mathcal{Q}_l) \\ &= H(D_l) - H(D_l|R) - H(D_l|\mathcal{Q}_l) + H(D_l|R,\mathcal{Q}_l) \\ &\leq H(D_l) + H(D_l|R,\mathcal{Q}_l) \leq 2H(D_l). \end{aligned} \quad (23)$$

Theorem 5 shows that when $m \to \infty$ with probability approaching 1, none of the base vectors exist in the span of any subset $\mathcal{Q}_l \subset \mathcal{Q}^\epsilon$ of cardinality less than or equal to $l = \lfloor \delta m \rfloor$ for $\delta < 0.5$. Hence, the event D_l will happen with probability 1 when $m \to \infty$. Therefore, $\lim_{m\to\infty} H(D_l) = 0$ and thus for any subset $\mathcal{Q}_l \subset \mathcal{Q}^\epsilon$ of cardinality less than or equal to $l = \lfloor \delta m \rfloor$ for $\delta < 0.5$ we have

$$\lim_{m\to\infty} I(R;\mathcal{Q}_l) \leq 2 \lim_{m\to\infty} H(D_l) = 0 \quad (24)$$

This proves the theorem. □

Remark 2. In practice the user generates enough number of random vectors $\mathcal{Q} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_A\}$ of $A \geq m$ to span the whole m-dimensional space. Then it chooses a subset $\mathcal{Q}^{\text{full}} = \{\mathbf{q}_{t_1}, \mathbf{q}_{t_2}, \ldots, \mathbf{q}_{t_m}\} \subseteq \mathcal{Q}$ of m linearly independent vectors from them and use this subset as its set of useful query vectors. This way it is guaranteed that the m chosen queries will span the whole space of $\mathbb{F}_2^m$ and any base vector $\mathbf{e}_r$ can be represented in terms of these independent query vectors. Let the decoding gain d_k be equal to 1 if query $\mathbf{q}_{t_k}$ is added to other queries in the representation of $\mathbf{e}_r$ and let it be zero otherwise. Then, $\mathbf{e}_r$ can be represented as

$$\mathbf{e}_r = \sum_{k=1}^{m} d_k \mathbf{q}_{t_k}. \quad (25)$$

The following lemma shows that the average number of random vectors to span $\mathbb{F}_2^m$ is very close to m. Therefore, the required number of random queries to span $\mathbb{F}_2^m$ is slightly more than m.

Lemma 7. If $\mathbf{q}_j$ is a random vector belonging to $\mathbb{F}_2^m$ with elements having uniform distribution, the average minimum number of vectors $\mathbf{q}_j$ to span the whole space of $\mathbb{F}_2^m$ equals

$$\mathbb{E}_q = m + \sum_{i=1}^{m} \frac{1}{2^i - 1} = m + \gamma, \quad (26)$$

where γ asymptotically approaches the Erdős–Borwein constant (≈ 1.6067).

Proof. The proof can be found in [5]. □

Remark 3. Since $\mathcal{Q}^{\text{full}}$ is a subset of $\mathcal{Q}$, if any vector $\mathbf{e}_r$ does not exist in the span of any subset of cardinality l of $\mathcal{Q}$, then this vector will not also exist in the span of any subset of cardinality l of $\mathcal{Q}^{\text{full}}$ too. Therefore, Theorems 5 and 6 remain valid for this choice of random queries too. This means that in practice, every base vector is guaranteed to exist in the span of the m query vectors but none of the base vectors exist in the span of any subset $\mathcal{Q}_l \subset \mathcal{Q}$ with probability close to one if $l < \lfloor \delta m \rfloor$ for $\delta < 0.5$.

Remark 4. Throughout lemmas and theorems in this section, we were able to prove Theorem 6 which shows that if the ratio of the number of compromised random queries to the total number of random queries that we utilize as our expansion vectors is less than 0.5, then no information about the requested content can be obtained from the compromised set of queries. This means that the collusion of up to half of all the servers would asymptotically reveal no information about the requested content index.

B. Responding to Queries

In this section, we assume that the user has chosen m linearly independent random query vectors in $\mathcal{Q}^{\text{full}}$ and wants to download the r^{th} content. Since $\mathcal{Q}^{\text{full}}$ is a set of vectors which spans the whole space of $\mathbb{F}_2^m$, the user can expand the base vector $\mathbf{e}_r$ in terms of the query vectors in $\mathcal{Q}^{\text{full}}$ as mentioned in (25). Hence, the requested content can be expanded in terms of query vectors as

$$f_r = \mathbf{f}\mathbf{e}_r = \mathbf{f}\left(\sum_{k=1}^{m} d_k \mathbf{q}_{t_k}\right) = \sum_{k=1}^{m} d_k \mathbf{f}\mathbf{q}_{t_k}, \quad (27)$$

where $d_k \in \mathbb{F}_2$ is either 0 or 1. Based on equation (27) the user requests some parts of the desired content from each RG so that none of the RGs can understand any information about the requested content.

To accomplish PIR, the user partitions the set of random queries $\mathbf{q}_{t_k}$ whose corresponding decoding gains d_k are non-zero into a disjoint subsets $\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_a$. The choice of the number of subsets (i.e. a) depends on the number of colluding servers. Each subset of queries is then sent to a different RG as depicted in Figure 2. Therefore, the requested content can be retrieved as

$$f_r = \sum_{\substack{\mathbf{q}_{t_k} \in \mathcal{Q}^{\text{full}} \\ d_k \neq 0}} \mathbf{f}\mathbf{q}_{t_k} = \sum_{i=1}^{a} \sum_{\mathbf{q}_{t_k} \in \mathcal{Q}_i} \mathbf{f}\mathbf{q}_{t_k} \quad (28)$$

The ultimate goal in PIR is to prevent any colluding group of servers to gain information about the requested content index. Assume that the number of colluding servers is b. If any two colluding servers lie within the same RG, they receive the same subset of queries from the user. Therefore, without loss of generality we consider the worst scenario in which all the colluding servers lie within different RGs and all these b colluding servers are able to collaboratively obtain all the queries $\mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_b$. Based on Theorems 5 and 6, if the number of all query vectors in $\mathcal{Q}_l = \cup_{i=1}^{b} \mathcal{Q}_i$ is less than $\lfloor \delta m \rfloor$ for some $\delta < 0.5$, then no information can be achieved about the requested content index. This provides significant PIR capability for this technique.

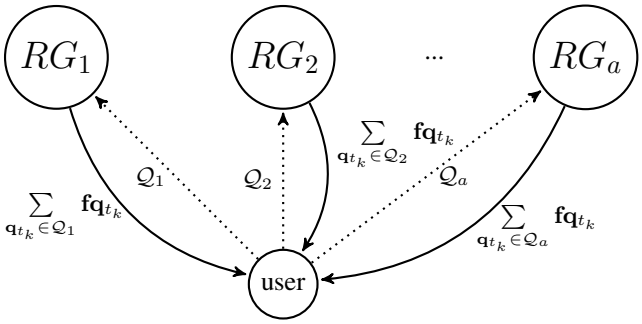

Fig. 2: Multiple RGs respond to queries sent from the user. This allows the user to privately download its desired content while a significant number of colluding servers can achieve no information about the downloaded content.

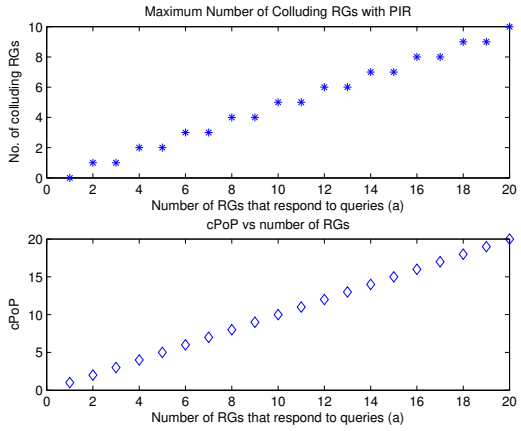

Fig. 3: The average communication Price of Privacy (cPoP) increases as the maximum number of colluding RGs increases.

Notice that since RGs have full rank encoding matrices, they can respond to any query that they receive. Assume that RG $\mathcal{J}_i$ with the full rank encoding matrix $\mathbf{V}=[\mathbf{V}_1\ \mathbf{V}_2\ \dots\ \mathbf{V}_{J_i}]$ receives the set of queries $\mathcal{Q}_i$. This RG needs to send $\sum_{\mathbf{q}_{t_k}\in\mathcal{Q}_i}\mathbf{fq}_{t_k}$ to the user. It can solve the linear equation

$$\mathbf{V}\mathbf{p}_i=\sum_{\mathbf{q}_{t_k}\in\mathcal{Q}_i}\mathbf{q}_{t_k} \tag{29}$$

in Galois Field $\mathbb{F}_2$ for $\mathbf{p}_i$ as

$$\mathbf{p}_i=\mathbf{V}^T\left(\mathbf{V}\mathbf{V}^T\right)^{-1}\left(\sum_{\mathbf{q}_{t_k}\in\mathcal{Q}_i}\mathbf{q}_{t_k}\right). \tag{30}$$

Similar to before the server $\mathcal{N}_s$ in the RG $\mathcal{J}_i$ which has already acquired all the information in matrix $\mathbf{V}$, computes the overall query decoding solution $\mathbf{p}_i$ which is a vector of size $J_ih\times 1$. If this vector is divided into J_i equal size pieces as $\mathbf{p}_i=[\mathbf{p}_i^1\ \mathbf{p}_i^2\ \dots\ \mathbf{p}_i^{J_i}]^T$, then the server $\mathcal{N}_s$ sends the j^{th} portion of $\mathbf{p}_i$ to server $\mathcal{N}_j$ in the RG $\mathcal{J}_i$. More precisely, server $\mathcal{N}_j$ receives a query response vector $\mathbf{p}_i^j$ of size $h\times 1$ from $\mathcal{N}_s$ for each $j=1,2,\dots,J_i$. Then the server $\mathcal{N}_j$ sends $\mathbf{fV}_j\mathbf{p}_i^j$ back to the coordinating server $\mathcal{N}_s$. The coordinating server $\mathcal{N}_s$ then aggregates all the data received from multiple servers in the RG to construct $\sum_{\mathbf{q}_{t_k}\in\mathcal{Q}_i}\mathbf{fq}_{t_k}$ as

$$\sum_{\mathbf{q}_{t_k}\in\mathcal{Q}_i}\mathbf{fq}_{t_k}=\mathbf{fV}\mathbf{p}_i=\sum_{j=1}^{J_i}\mathbf{fV}_j\mathbf{p}_i^j. \tag{31}$$

The coordinating server $\mathcal{N}_s$ in the RG $\mathcal{J}_i$ then transmits $\sum_{\mathbf{q}_{t_k}\in\mathcal{Q}_i}\mathbf{fq}_{t_k}$ to the user.

Each RG only transmits one encoded file to the user. However all the servers within an RG need to collaborate with each other prior to responding to the queries sent from the user. Notice that communication between the servers are carried using high bandwidth fiber optic links while transmissions from the servers to the user are performed over low bandwidth links. In our computation of communication cost for achieving PIR, the cost of sending queries are ignored because it is assumed that the size of contents are significantly higher than the size of the queries.

Remark 5. It is worth mentioning that in our paper, the coordinations between servers in an RG is necessary because the servers do not have full storage capacity to store all the contents. In fact if we also assume each server has high storage capacity similar to [8], then each server can act as an RG and there will be no communications between servers.

C. Trade-off Between Communication Cost and Privacy Level

In order to achieve PIR, each user needs to download more information. This additional bandwidth utilization is referred to as *communication Price of Privacy (cPoP)* [8] which is defined as follows.

Definition 5. The communication Price of Privacy (cPoP) is the ratio of the total number of bits downloaded by the user from the servers to the size of the requested file.

To explain the trade-off between communication cost and level of privacy, assume that the user divides the queries into a equal size groups of queries and sends each group of queries to a different RG. Each RG should respond to at most $\lceil\frac{m}{a}\rceil$ queries. If b RGs collude to gain some information about the requested content index, then they will have access to a total of at most $b\lceil\frac{m}{a}\rceil$ queries. We proved in Theorem 6 that knowing $\lfloor\delta m\rfloor$ queries gives no information about the requested content index if $\delta<0.5$. Hence, if $b<\frac{a}{2}$, then the colluding RGs will get no information about the requested content index. Therefore, if less than half of the RGs collude to gain some information about the requested content index, they cannot gain any information. We can increase a to get the maximum possible level of privacy. However, the downside of increasing a is the increase in communication Price of Privacy (cPoP).

As discussed earlier, if the queries are sent to a RGs then a responses from these RGs are required to retrieve a content. Since each RG transmits an encoded file of size M bits to the user the total number of bits downloaded by the user will be equal to aM and therefore the cPoP will be equal to $aM/M=a$. Figure 3 shows that as a increases, both the PIR strength and the cPoP increase linearly.

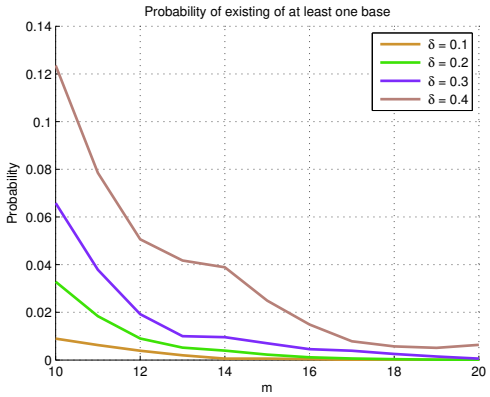

Fig. 4: Probability of the event that at least one base exists in the span of any subset of $l = \lfloor \delta m \rfloor$ random vectors.

D. Full Size Servers

Assume that the servers have large storage capability such that each RG is only composed of 1 server. Our assumption of full rank encoding scheme guarantees that servers with storage ability of $h \geq m$ encoded files can be used to retrieve any desired content. In [8], the authors studied the use of MDS codes for PIR. They considered full size storage systems with MDS codes and they considered the case when only one of the databases is compromised. They proposed a PIR technique in which a cPoP of $\frac{1}{1-R}$ can be achieved in full size databases where R is the MDS code rate. To compare our results with [8], notice that if we assume that there is only one malicious server in the cloud, then we can choose any two servers and send half of the queries to each one of them. This way we have a cPoP of 2 which is better than the results in [8] for $R > 1/2$.

VI. Simulation

To numerically verify the results proved in section V, we created m linearly independent random query vectors which are used to expand the bases. Figure 4 demonstrates the probability of the event that at least one of the base vectors exists in the span of $l = \lfloor \delta m \rfloor$ vectors for $\delta = 0.1, 0.2, 0.3$ and 0.4. Consistent with our results in section V, the probability of the event that a base exists in the span of any set of $l = \lfloor \delta m \rfloor$ vectors goes quickly to zero.

It is proved [44] that the problem of finding the minimum spanning set of vectors is NP-Complete. It is even proved [45] that this problem is NP-Hard to approximate. Therefore, in general it is NP-Hard to find out if a given base exists in the span of at most $l = \lfloor \delta m \rfloor$ vectors out of the m vectors. For our simulations we have used a brute force approach to check if a given base exists in the span of at most $l = \lfloor \delta m \rfloor$ vectors out of the m existing random query vectors where $m \leq 20$.

VII. Conclusions

In this paper, we have studied the problems of security and private information retrieval in distributed storage systems which are using a full rank encoding scheme based on Random Linear Fountain (RLF) codes. We have proposed an approach based on uniform random queries to achieve information theoretic PIR property. We have proved that our proposed technique can asymptotically achieve perfect secrecy for a distributed storage system. Our proposed solution is robust against a significant number of colluding servers in the network. We have also shown that our technique can outperform MDS codes for storage systems in terms of PIR cost for certain regimes.

References

[1] Claude E Shannon. Communication theory of secrecy systems*. *Bell system technical journal*, 28(4):656–715, 1949.

[2] David JC MacKay. Fountain codes. *IEE Proceedings-Communications*, 152(6):1062–1068, 2005.

[3] Mohsen Karimzadeh Kiskani and Hamid R Sadjadpour. Capacity of cellular networks with femtocache. In *Computer Communications Workshops (INFOCOM WKSHPS), 2016 IEEE Conference on*, pages 9–14. IEEE, 2016.

[4] Mohsen Karimzadeh Kiskani and Hamid R. Sadjadpour. Secure coded caching in wireless ad-hoc networks. In *International Conference on Computing, Networking and Communications (ICNC)*, January 2017.

[5] Mohsen Karimzadeh Kiskani and Hamid R Sadjadpour. Throughput analysis of decentralized coded content caching in cellular networks. *IEEE Transactions on Wireless Communications*, 16(1):663–672, 2017.

[6] Mohsen Karimzadeh Kiskani and Hamid R Sadjadpour. A secure approach for caching contents in wireless ad hoc networks. *IEEE Transactions on Vehicular Technology*, 66(11):10249–10258, 2017.

[7] Mohsen Karimzadeh Kiskani and Hamid R. Sadjadpour. Secure and private cloud storage systems with random linear fountain codes. In *IEEE SmartWorld, Ubiquitous Intelligence and Computing, Advanced and Trusted Computed, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, Aug 2017.

[8] Razan Tajeddine and Salim El Rouayheb. Private information retrieval from MDS coded data in distributed storage systems. *arXiv preprint arXiv:1602.01458*, 2016.

[9] Alexandros G Dimakis, Vinod Prabhakaran, and Kannan Ramchandran. Distributed fountain codes for networked storage. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 5, pages V–V. IEEE, 2006.

[10] Alexandros G Dimakis, P Brighten Godfrey, Yunnan Wu, Martin J Wainwright, and Kannan Ramchandran. Network coding for distributed storage systems. *IEEE Transactions on Information Theory*, 56(9):4539–4551, 2010.

[11] Alexandros G Dimakis, Kannan Ramchandran, Yunnan Wu, and Changho Suh. A survey on network codes for distributed storage. *Proceedings of the IEEE*, 99(3):476–489, 2011.

[12] Theodoros K Dikaliotis, Alexandros G Dimakis, and Tracey Ho. Security in distributed storage systems by communicating a logarithmic number of bits. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1948–1952. IEEE, 2010.

[13] Sameer Pawar, Salim El Rouayheb, and Kannan Ramchandran. On secure distributed data storage under repair dynamics. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 2543–2547. IEEE, 2010.

[14] Sameer Pawar, Salim El Rouayheb, and Kannan Ramchandran. Securing dynamic distributed storage systems against eavesdropping and adversarial attacks. *IEEE Transactions on Information Theory*, 57(10):6734–6753, 2011.

[15] Sameer Pawar, Salim El Rouayheb, and Kannan Ramchandran. Securing dynamic distributed storage systems from malicious nodes. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 1452–1456. IEEE, 2011.

[16] Nihar B Shah, KV Rashmi, and P Vijay Kumar. Information-theoretically secure regenerating codes for distributed storage. In *Global Telecommunications Conference (GLOBECOM 2011), 2011 IEEE*, pages 1–5. IEEE, 2011.

[17] Ankit Singh Rawat, Onur Ozan Koyluoglu, Natalia Silberstein, and Sriram Vishwanath. Optimal locally repairable and secure codes for distributed storage systems. *IEEE Transactions on Information Theory*, 60(1):212–236, 2014.

[18] Mohsen Karimzadeh Kiskani, Hamid R. Sadjadpour, Mohammad Reza Rahimi, and Fred Etemadieh. Low Complexity Secure Code (LCSC) design for big data in cloud storage systems. In *IEEE International Conference on Communications (ICC)*. IEEE, May 2018.

[19] Siddhartha Kumar, Eirik Rosnes, and Alexandre Graell i Amat. Secure repairable fountain codes. *IEEE Communications Letters*, 20(8):1491–1494, 2016.

[20] Benny Chor, Oded Goldreich, Eyal Kushilevitz, and Madhu Sudan. Private information retrieval. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 41–50. IEEE, 1995.

[21] Erica Y Yang, Jie Xu, and Keith H Bennett. Private information retrieval in the presence of malicious failures. In *Computer Software and Applications Conference, 2002. COMPSAC 2002. Proceedings. 26th Annual International*, pages 805–810. IEEE, 2002.

[22] Ian Goldberg. Improving the robustness of private information retrieval. In *Security and Privacy, 2007. SP'07. IEEE Symposium on*, pages 131–148. IEEE, 2007.

[23] Nihar B Shah, KV Rashmi, and Kannan Ramchandran. One extra bit of download ensures perfectly private information retrieval. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 856–860. IEEE, 2014.

[24] Arman Fazeli, Alexander Vardy, and Eitan Yaakobi. PIR with low storage overhead: coding instead of replication. *arXiv preprint arXiv:1505.06241*, 2015.

[25] Terence H Chan, Siu-Wai Ho, and Hirosuke Yamamoto. Private information retrieval for coded storage. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 2842–2846. IEEE, 2015.

[26] Hua Sun and Syed A Jafar. The capacity of private information retrieval. *IEEE Transactions on Information Theory*, 2017.

[27] Hua Sun and Syed A Jafar. The capacity of robust private information retrieval with colluding databases. *arXiv preprint arXiv:1605.00635*, 2016.

[28] Karim Banawan and Sennur Ulukus. The capacity of private information retrieval from coded databases. *arXiv preprint arXiv:1609.08138*, 2016.

[29] Ragnar Freij-Hollanti, Oliver Gnilke, Camilla Hollanti, and David Karpuk. Private information retrieval from coded databases with colluding servers. *arXiv preprint arXiv:1611.02062*, 2016.

[30] Siddhartha Kumar, Eirik Rosnes, et al. Private information retrieval in distributed storage systems using an arbitrary linear code. *arXiv preprint arXiv:1612.07084*, 2016.

[31] Heecheol Yang, Wonjae Shin, and Jungwoo Lee. Private information retrieval for secure distributed storage systems. *IEEE Transactions on Information Forensics and Security*, 13(12):2953–2964, 2018.

[32] Ning Cai and Raymond W Yeung. Secure network coding. In *Information Theory, 2002. Proceedings. 2002 IEEE International Symposium on*, page 323. IEEE, 2002.

[33] Kapil Bhattad, Krishna R Narayanan, et al. Weakly secure network coding. *NetCod, Apr*, 104, 2005.

[34] Swanand Kadhe and Alex Sprintson. On a weakly secure regenerating code construction for minimum storage regime. In *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*, pages 445–452. IEEE, 2014.

[35] Swanand Kadhe and Alex Sprintson. Weakly secure regenerating codes for distributed storage. In *Network Coding (NetCod), 2014 International Symposium on*, pages 1–6. IEEE, 2014.

[36] Muxi Yan, Alex Sprintson, and Igor Zelenko. Weakly secure data exchange with generalized reed solomon codes. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 1366–1370. IEEE, 2014.

[37] Peter Van Liesdonk, Saeed Sedghi, Jeroen Doumen, Pieter Hartel, and Willem Jonker. Computationally efficient searchable symmetric encryption. In *Workshop on Secure Data Management*, pages 87–100. Springer, 2010.

[38] Mihir Bellare, Alexandra Boldyreva, and Adam O'Neill. Deterministic and efficiently searchable encryption. In *Annual International Cryptology Conference*, pages 535–552. Springer, 2007.

[39] David Cash, Joseph Jaeger, Stanislaw Jarecki, Charanjit S Jutla, Hugo Krawczyk, Marcel-Catalin Rosu, and Michael Steiner. Dynamic searchable encryption in very-large databases: Data structures and implementation. In *NDSS*, volume 14, pages 23–26. Citeseer, 2014.

[40] Emil Stefanov, Charalampos Papamanthou, and Elaine Shi. Practical dynamic searchable encryption with small leakage. In *NDSS*, volume 71, pages 72–75, 2014.

[41] Matthieu Bloch and Joao Barros. *Physical-layer security: from information theory to security engineering*. Cambridge University Press, 2011.

[42] Valentin Fedorovich Kolchin. *Random graphs*. Number 53. Cambridge University Press, 1999.

[43] David JC MacKay. Good error-correcting codes based on very sparse matrices. *IEEE transactions on Information Theory*, 45(2):399–431, 1999.

[44] Alexander Vardy. The intractability of computing the minimum distance of a code. *IEEE Transactions on Information Theory*, 43(6):1757–1766, 1997.

[45] Ilya Dumer, Daniele Micciancio, and Madhu Sudan. Hardness of approximating the minimum distance of a linear code. *IEEE Transactions on Information Theory*, 49(1):22–37, 2003.

Mohsen Karimzadeh Kiskani received his bachelors degree in mechanical engineering from Sharif University of Technology in 2008. He got his Masters degree in electrical engineering from Sharif University of Technology in 2010. He got his PhD in electrical engineering from University of California Santa Cruz in September 2017. He has been a postdoctoral researcher at University of California Santa Cruz since December 2017. He has done research and published 15 papers in areas related to wireless communications, security, machine learning and information theory. He also obtained a masters degree in computer science from University of California Santa Cruz in 2016. His main areas of interest in computer science include machine learning and deep learning.

Hamid Sadjadpour (S'94–M'95–SM'00) received the B.S. and M.S. degrees from the Sharif University of Technology, and the Ph.D. degree from the University of Southern California at Los Angeles, Los Angeles, CA. In 1995, he joined the AT&T Research Laboratory, Florham Park, NJ, USA, as a Technical Staff Member and later as a Principal Member of Technical Staff. In 2001, he joined the University of California at Santa Cruz, Santa Cruz, where he is currently a Professor. He has authored over 170 publications. He holds 17 patents. His research interests are in the general areas of wireless communications and networks. He has served as a Technical Program Committee Member and the Chair in numerous conferences. He is a co-recipient of the best paper awards at the 2007 International Symposium on Performance Evaluation of Computer and Telecommunication Systems and the 2008 Military Communications conference, and the 2010 European Wireless Conference Best Student Paper Award. He has been a Guest Editor of EURASIP in 2003 and 2006. He was a member of the Editorial Board of Wireless Communications and Mobile Computing Journal (Wiley), and the Journal Of Communications and Networks.