Coding for Trusted Storage in Untrusted Networks

Article *in* IEEE Transactions on Information Forensics and Security · December 2012

DOI: 10.1109/TIFS.2012.2217331

5 authors, including:

Paulo F. Oliveira
University of Porto
7 PUBLICATIONS **96** CITATIONS
SEE PROFILE

Tiago Vinhoza
University of Porto
30 PUBLICATIONS **305** CITATIONS
SEE PROFILE

Muriel Médard
Massachusetts Institute of Technology
579 PUBLICATIONS **19,908** CITATIONS
SEE PROFILE

Coding for Trusted Storage in Untrusted Networks

Paulo F. Oliveira, *Member, IEEE*, Luísa Lima, *Member, IEEE*, Tiago T. V. Vinhoza, *Member, IEEE*, João Barros, *Senior Member, IEEE*, and Muriel Médard, *Fellow, IEEE*

***Abstract*—We focus on the problem of secure distributed storage over multiple untrusted clouds or networks. Our main contribution is a low complexity scheme that relies on erasure coding techniques for achieving prescribed levels of confidentiality and reliability. Using matrices that have no singular square submatrices, we subject the original data to a linear transformation. The resulting coded symbols are then stored in different networks. This scheme allows users with access to a threshold number of networks to reconstruct perfectly the original data, while ensuring that eavesdroppers with access to any number of networks smaller than this threshold are unable to decode any of the original symbols. This holds even if the attackers are able to guess some of the missing symbols. We further quantify the achievable level of security, and analyze the complexity of the proposed scheme.**

***Index Terms*—Distributed storage, security, erasure codes.**

I. Introduction

MOTIVATED by the advent and increasing popularity of cloud computing and distributed storage systems [2], we consider a scenario in which a large, confidential file is to be stored securely and in a distributed fashion over multiple networks. These networks are not trustworthy in the sense that an attacker may gain access to some of them, but not to all. Such a scenario is reasonable, since more and more organizations own private clouds and implement security policies that cannot be broken in a straightforward manner [2]. A natural question arises: is it possible to store the file in such a way that attackers with access to a subset of these networks are unable to reconstruct the whole file or any of its parts? A straightforward cryptographic solution would be to encrypt the file using a secret key and then partition the resulting cryptogram into multiple packets that can be spread over the various untrusted networks. However, any approach that encrypts the file using a secret key introduces the need for key management mechanisms [3], which increases the overall complexity and the resources demanded by the system.

We propose a solution that relies on coding techniques rather than in classical cryptography. Although several contributions have uncovered the advantages of erasure codes to ensure superior reliability in distributed storage systems [4], few have addressed their potential to simultaneously provide data confidentiality and resiliency to failures in these applications [5]. By exploiting the benefits of erasure coding techniques, we show that a prescribed level of data confidentiality can be achieved at no additional bandwidth or storage space costs, thus dispensing the need for secret key distribution in certain distributed storage systems. However, our solution requires the location of the networks storing the information to be available to legitimate users and to be kept secret from eavesdroppers. Hence, in our system setup, the information about the location of the data plays the role of a secret key in an encryption scheme. We also demonstrate that these techniques guarantee resiliency to failures albeit at the expense of extra storage space.

Consider the example shown in *Fig. 1*. A large file is to be stored in a distributed fashion in γ untrusted networks (represented by clouds) in such a way that (i) legitimate users with access to any $\gamma - f$ networks are able to reconstruct the entire file, and (ii) eavesdroppers observing any $\gamma_e < \gamma - f$ networks are unable to recover any of its parts. To accomplish these goals, the user splits the file into smaller chunks consisting of n symbols drawn out from a Galois field. Each chunk is then encoded using a linear transformation that maps its n original symbols into $n + r$ coded symbols, where r is the required amount of redundant data that allows our system to be resilient to the removal of up to f networks, which we denote as *network failures*. The next step is to store different subsets of coded symbols in each available network. For simplicity, we assume that each network stores the same amount of data, i.e., $(n + r)/\gamma$ coded symbols per chunk. The location of the networks can be shared with authorized users who have access to the file. In fact, this knowledge can be seen as the advantage a legitimate user has over an attacker, analogous to a secret key in a cryptographic perspective. The proposed coding scheme achieves these goals using encoding matrices that have no singular submatrices (as specified in Section III-B), ensuring that:

- any n out of $n + r$ coded symbols are enough to recover the n original symbols of a chunk;
- an attacker observing strictly less than n coded symbols does not gather information about any individual symbol of a chunk.

This allows the user to mix the original data in such a way that the scheme is resilient to the removal of up to f networks from the system. Simultaneously, an attacker is provably unable to

Manuscript received April 12, 2011; revised August 04, 2012; accepted August 13, 2012. Date of publication September 07, 2012; date of current version November 15, 2012. This work was supported in part by the Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) under Grant SFRH/BD/24718/2005 and Grant SFRH/BD/28946/2006, and in part by the European Commission under Grant FP7-INFSO-ICT-215252 (N-Crave Project). This work was presented in part at the IEEE Global Communications Conference (Globecom'10), Miami, FL, Dec. 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wade Trappe.

P. F. Oliveira and L. Lima are with Instituto de Telecomunicações, Departamento de Ciência de Computadores, Faculdade de Ciências da Universidade do Porto, Porto 4169-007, Portugal (e-mail: pvf@dcc.fc.up.pt; luisalima@dcc.fc.up.pt).

T. T. V. Vinhoza and J. Barros are with Instituto de Telecomunicações, Departamento de Engenharia Electrotécnica e de Computadores, Faculdade de Engenharia da Universidade do Porto, Porto 4200-465, Portugal (e-mail: tiago.vinhoza@ieee.org; jbarros@fe.up.pt).

M. Médard is with Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: medard@mit.edu).

Digital Object Identifier 10.1109/TIFS.2012.2217331

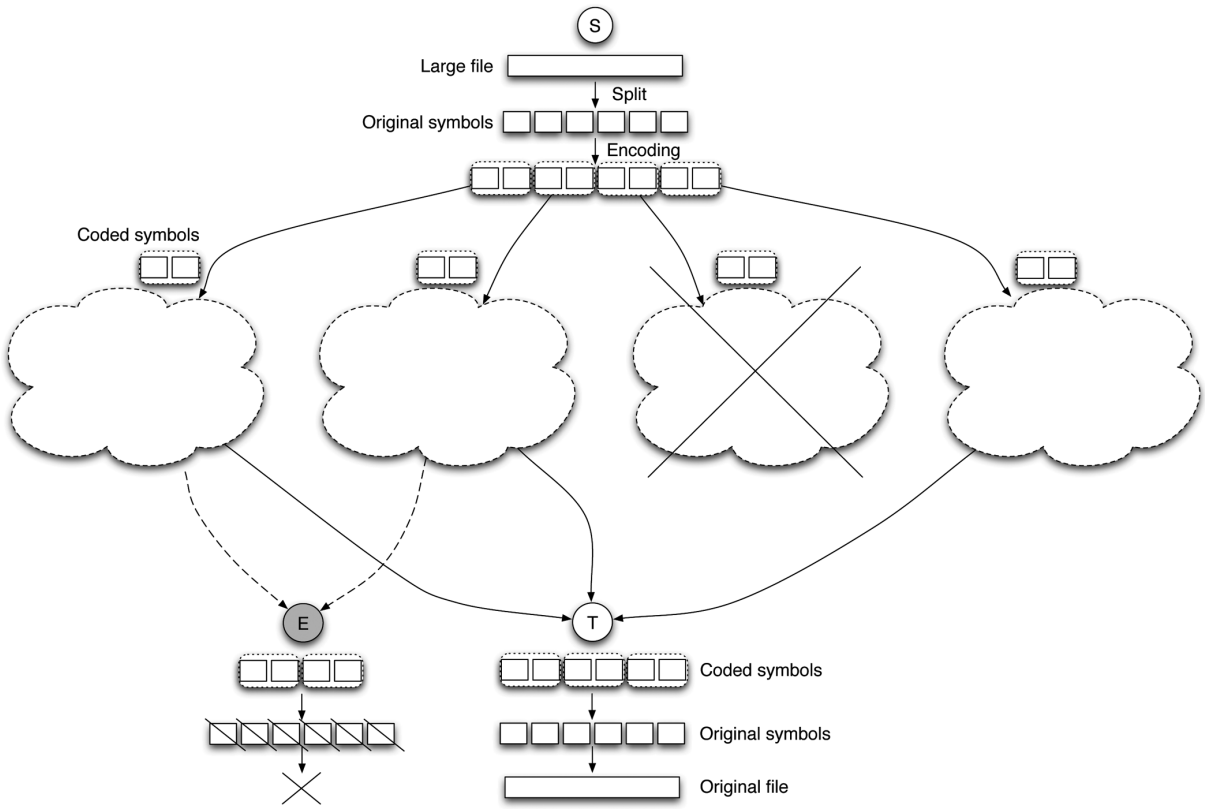

Fig. 1. Example of distributed storage over multiple untrusted networks. A large file is coded and stored by source S using four untrusted networks. Our scheme allows legitimate user T with access to any three networks to reconstruct the original data, while ensuring that eavesdropper E with access only to two networks of his choice is unable to decode any original symbol, even if he is able to guess some of the missing symbols.

recover any individual information symbol — even if he is able to guess part of the file. Thus, our main contribution is a distributed storage scheme that provides the following guarantees:

- *Quantifiable Level of Security:* We use information-theoretic arguments to show that an eavesdropper with access to any γ_e networks is unable to recover any individual original symbol. The proposed system includes a tunable security parameter k, related to the number of networks the eavesdropper can see. This parameter ensures that no additional symbols are obtained even if an eavesdropper observing chunks containing $n-k$ coded symbols is able to guess up to $k-1$ original symbols of each chunk. Also, we prove that the knowledge of a chunk does not reveal any additional information about the other chunks to the eavesdropper. Our scheme can also be extended to yield perfect secrecy (or unconditional security) albeit at the expense of appending each chunk with random symbols before encoding.
- *Fault-tolerance:* We provide design guidelines for our system to be resilient to f *network failures*. That means that a legitimate user is able to perfectly reconstruct the file even when f networks are unavailable.
- *Recovery Efficiency:* Since the proposed scheme uses maximum distance separable (MDS) codes [6], it guarantees that n original symbols are obtained from any subset of n coded symbols, which is the minimum amount of information required to recover the original data.
- *Low Complexity:* We show that the proposed scheme can be efficiently implemented due to the structure of the matrices used for encoding, which allows for fast computation in terms of algebraic operations.

The work in [1] considers secure distributed storage over two untrusted networks, whereby a Vandermonde matrix is used as a means to achieve a prescribed level of confidentiality. Our contribution differs from the work in [1] in the following aspects: (a) concerned about the fact that Vandermonde matrices are not sufficient to ensure confidentiality with probability one against partial decoding in multiple networks, we focus on the general family of matrices that allow for confidentiality in complex systems composed by multiple networks, while keeping low requirements on the amount of resources needed; (b) we target resilience to network availability failures without compromising neither the performance nor the security of the system; (c) we discuss the design trade-offs when implementing the proposed scheme.

The rest of the paper is organized as follows. Section II provides an overview of relevant related work. The proposed distributed storage system is presented in Section III, which explains the coding and the recovering process of the proposed scheme under the adopted intruder model. Section IV then evaluates the security performance of the proposed scheme. After some discussion on implementation aspects and the security-reliability trade-off in Section V, the paper concludes with Section VI.

II. Related Work

Our work touches a number of areas, including coding, cryptography, and distributed storage. Before proceeding with the

description of the related work, it is worth reviewing common threat models in these areas and how they relate to our distributed storage scenario.

A. Erasure Coding Techniques for Secrecy

Coding techniques are employed in [7] to achieve perfect secrecy over a channel in which an eavesdropper acquires a fraction of the transmitted symbols (wiretap channel II model). In particular, it is shown that a coset scheme achieves the maximum secret rate albeit at the expense of data rate. The maximum number of symbols that can be securely communicated is upper bounded by $n-\phi$, where n is the total number of coded symbols transmitted and ϕ is the number of coded symbols observed by the eavesdropper. A modified version of the wiretap channel II is considered in [8], where the number of erasures at the eavesdropper is fixed. The positions are chosen at random and a coding scheme based on nested MDS codes is shown to achieve the secrecy capacity. Our work can be viewed as a wiretap channel of type II in which the channel eavesdropped by an attacker is worse than the main channel from the source to the legitimate user, ensuring that when the legitimate user receives n coded symbols, the eavesdropper sees $n-k$ coded symbols. In addition, our scheme ensures that even if the eavesdropper is able to guess up to $k-1$ symbols, he does not obtain any other missing symbol. These coding techniques can also be designed from a network point of view, which we overview next.

B. Secure Network Coding

In the context of coded networks, threats are brought about by passive eavesdroppers who observe a subset of nodes or edges of a network carrying coded symbols. In our system setup, this can be viewed as a subset of untrusted networks storing a fraction of coded symbols. The goal in [9] is to build a network code that achieves perfect secrecy under the premise that an eavesdropper only has access to a number of edges in which the sum of their rates is smaller than the network capacity. Lower bounds on the field size that guarantee the existence of a secure network code are derived in [10] by modeling the problem as a network generalization of the wiretap channel of type II. Code constructions that achieve these bounds are also proposed.

A quantifiable security criterion is introduced in [11] to measure the attainable level of secrecy in a multicast scenario. In our security analysis we use this security criterion, which is specified in *Definition 1* on Section III-D. The work in [11] also establishes the requirements to achieve security under the defined security criterion, given the network topology and the network code. These conditions can be met by applying an encoding matrix at the source. However, finding a matrix that satisfies such requirements is computationally complex. The contribution in [12] generalizes this problem by proposing a secure source coding scheme that is independent from both the network topology and the network code. However, as in [11], it is hard to find a matrix that ensures that the system is resistant to an arbitrary number of guesses.

The work in [13] derives bounds for the probability of decoding an individual symbol in a network where Random Linear Network Coding (RLNC) is used. It is shown that RLNC provides inherent security for a threat model in which the intermediate nodes are honest but curious, i.e., they comply with the protocol yet try to decode as much data as possible.

The construction of a secure linear network code for a wiretap network where the wiretapper is allowed to observe a subset of network channels of his choice is proposed in [14]. The code is shown to be optimal in several scenarios and includes secret sharing as a particular case. In [15], provided that the multicast capacity is at least n, a wiretapper observing $n-s$ channels is oblivious to any s components of the source message, where $s \leq n-r$. This approach is also based on secret sharing, which we overview next.

C. Secret Sharing

In a secret sharing scheme, one divides a secret into pieces, called *shares*. Typically, a threshold number of shares is sufficient to recover the original secret, but any number of shares smaller than the threshold reveal *no information* about the secret [16], [17]. There are at least two possible interpretations for the *no information* criterion, based on which current secret sharing schemes are built. In computational security, *no information* means that no information about the secret can be computed in polynomial time on the length of the input, whereas in information-theoretic security the same concept implies that no secret information is revealed, regardless of the computational power of the attacker, that is, the mutual information between the secret and the observed shares is zero. In the (k,n)-threshold secret sharing scheme proposed in [16], a secret is divided into n shares, in such a way that the secret can be recovered from any group of k shares, while $k-1$ shares reveal *no information* about the secret in the information-theoretic sense. In this case, each share requires the same bit-length as the original information, which induces a penalty in terms of rate. In a (k,L,n)-threshold ramp secret sharing scheme [18], the secret information is recovered from any k out of n shares. Compared to [16], the scheme is more efficient since each share is $1/L$ of the size of the secret. However, information about the secret is leaked to an attacker with access to more than $(k-L)$ shares. Our scheme can be viewed as a secret sharing scheme, where each original symbol is a secret and each coded symbol is a share. In our scheme, $n-k$ shares reveal *no information* (in the information-theoretic sense) about any k secrets, and all the n secrets can be recovered from any n shares.

D. All-or-Nothing Transforms

In order to increase the difficulty of a successful brute force search attack to a block cipher, the work in [19] introduces a linear preprocessing scheme, called all-or-nothing transform. This scheme guarantees that an attacker must decrypt the entire ciphertext before it can learn a single message block. The Optimal Asymmetric Encryption Padding introduced in [20] satisfies this definition for RSA encryption, as proved in [21]. The work in [22] considers all-or-nothing transforms and, in particular, addresses unconditional security with respect to a single block of the original message. Our scheme relates to an all-or-nothing transform in the sense that an eavesdropper with partial access to the coded data is unable to invert the linear transformation performed at the source. In particular, all-or-nothing transforms guarantee that the mutual information between any orig-

inal symbol and any $n-1$ coded symbols is zero. Our scheme, in turn, extends this result towards ensuring that the mutual information between any k original symbols and any $n-k$ coded symbols is zero.

E. Information Dispersal Algorithm and Distributed Storage

The work in [23] develops an Information Dispersal Algorithm (IDA) that splits a large file into n pieces such that m pieces are sufficient to recover the original file. This task is accomplished by segmenting the file in chunks of m pieces followed by a $n \times m$ linear transformation. Our scheme can be viewed as an Information Dispersal Algorithm which can be used not only for reliability purposes, but also for enhancing the security of distributed storage systems, since it adds an extra layer of security on top of the IDA fault-tolerance property.

The work in [24] proposes the combined use of IDA and secret sharing and shows how these tools can be used effectively to store a file in a secure and reliable way. The source encrypts the file using a secret key and then partitions the resulting cryptogram (using IDA) into n fragments. Using a secret sharing scheme, the secret key is also divided into n shares. The secret key is recovered from any m out of n shares and the cryptogram is recovered from m out of n cryptogram fragments. The main result in [24] is that although m shares of the secret (i.e., m shares of the encrypted file and m shares of the secret key) are enough to recover the original file, the access to $m-1$ shares gives no (computational) information about the secret. A practical application of the combination of IDA and secret sharing is proposed in [25].

As an example of an application, secure distributed storage in sensor networks is considered in [26]. The main idea is to distribute parts of data by different sensors in such a way that data reconstruction requires access to a certain number of sensors that have parts of the data. A technique to hide information without the presence of an encryption key is presented in [27], where a recursive approach to secret sharing is employed.

III. Problem Setup

We now introduce the notation used in the remainder of the paper. Then, we describe the coding and decoding schemes that satisfy the security criterion under the adopted threat model.

Notation

Vectors are represented by lowercase boldface and matrices are represented by capital boldface letters. $\mathbf{I}_n$ denotes a $n \times n$ identity matrix, and $\mathbf{0}_{n\times m}$ denotes a zero matrix of size $n \times m$. The subvector formed by any k components of a vector $\mathbf{v}$ is denoted by $\mathbf{v}^{(k)}$ and, in a slight abuse of notation, we denote by $\mathbf{M}_{m\times n}$ any $m \times n$ submatrix of $\mathbf{M}$. It will be clear from the context what m rows and n columns of $\mathbf{M}$ are being referred in $\mathbf{M}_{m\times n}$. Finally, we denote the transpose of a vector $\mathbf{v}$ as $\mathbf{v}^T$ and a Galois field with cardinality q as $\mathbb{F}_q$.

A. Coding Scheme

Let the original data $\mathcal{B}$ be a vector composed by N chunks, i.e., $\mathcal{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N)^T$. Each chunk is a vector $\mathbf{b}_c = (b_{c,1}, b_{c,2}, \ldots, b_{c,n})^T$ whose components $b_{c,j}$, $j = 1, \ldots, n$ are independent random variables uniformly distributed over $\mathbb{F}_q$, with entropy $H(b_{c,j}) = H(b)$. We aim to store $\mathcal{B}$ over γ untrusted networks. We also want resiliency against f network failures. To accomplish this, we apply a linear transformation, i.e., a coding operation, in each chunk that maps n input symbols belonging to $\mathbb{F}_q$ to $n+r$ coded symbols. The amount of redundancy r required in each coded chunk depends on the level of reliability a system designer wants.

Let $\mathbf{A}$ be the $(n+r)\times n$ matrix used for performing coding at the source. The elements of $\mathbf{A} = [a_{i,j}]$ belong to a finite field $\mathbb{F}_q$. A mandatory requirement for the choice of $\mathbf{A}$ is that any of its square submatrices must be nonsingular, i.e., all submatrices containing any p rows and any p columns of $\mathbf{A}$ are invertible, for all $p = 2, \ldots, n$. An example of a matrix that satisfies this requirement is $\mathbf{A} = \mathbf{A}_1^{-1} \cdot \mathbf{A}_2$, where $\mathbf{A}_1$ is a $(n+r)\times(n+r)$ Vandermonde matrix whose j-th row is $\left(\alpha_1^{j-1}, \ldots, \alpha_{n+r}^{j-1}\right)$ and $\mathbf{A}_2$ is a $(n+r)\times n$ Vandermonde matrix whose j-th row is $\left(\alpha_{n+r+1}^{j-1}, \ldots, \alpha_{2n+r}^{j-1}\right)$. Note that for both $\mathbf{A}_1$ and $\mathbf{A}_2$, $\alpha_i \neq \alpha_{i'}$ for $i \neq i'$, $i, i' = 1, \ldots, 2n+r$. Alternatively, $\mathbf{A}$ can be a Cauchy matrix, i.e., a matrix whose component (i,j) is of the type $1/(\alpha_i + \beta_j)$, $i = 1, \ldots, n+r$, $j = 1, \ldots, n$, and $\alpha_i \neq \beta_j$ [28]. In either case, the construction of $\mathbf{A}$ requires $2n+r$ nonzero distinct elements. Therefore, the finite field cardinality for constructing the aforementioned matrices must be greater than $2n+r$. Further matrix constructions are outside the scope of this paper.

Upon the generation of matrix $\mathbf{A}$, each chunk $\mathbf{b}_c$ of the original data $\mathcal{B}$ is encoded according to $\mathbf{c}_c = \mathbf{A}\mathbf{b}_c$, where each component of $\mathbf{c}_c = (c_{c,1}, c_{c,2}, \ldots, c_{c,n+r})^T$ is described by $c_{c,i} = \sum_{j=1}^{n} a_{i,j} b_{c,j}$, $1 \leq i \leq n+r$. The coded chunks are then stacked to form the encoded data vector $\mathcal{C} = (\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_N)^T$.

Assuming that each network stores $(n+r)/\gamma$ coded symbols of each chunk, one requires $r = nf/(\gamma - f)$ redundant symbols to be added during the coding process. This allows the recovery of all n original symbols of all N chunks from any $\gamma - f$ available networks, as shown later in *Proposition 1*.

B. Decoding Scheme

We assume that the users know the location of the networks storing the data, either because this information is securely transmitted to them, or because they are the sources of data themselves and store this information locally. To recover the original data $\mathcal{B}$, a legitimate user collects n components of each vector $\mathbf{c}_i$ from the available networks. Assuming that matrix $\mathbf{A}$ is public and that the user has information on which components of vectors $\mathbf{c}_1, \ldots, \mathbf{c}_N$ are stored, vectors $\mathbf{b}_1, \ldots, \mathbf{b}_N$ are obtained by performing $\mathbf{A}_{n\times n}^{-1}\mathbf{c}_i^{(n)}$ for all $i \in \{1, \ldots, N\}$.

Proposition 1: The legitimate user can recover the original data $\mathcal{B}$ by accessing any $\gamma - f$ networks.

Proof: Recall that each network stores $(n+r)/\gamma$ different coded symbols of each chunk of data. With $r = nf/(\gamma - f)$, a user having access to $\gamma - f$ networks is able to collect n different coded symbols of each one of the N chunks. Since, by design, any submatrix of matrix $\mathbf{A}$ is nonsingular, it can be reduced to the identity matrix after elementary row operations. Therefore, by performing $\mathbf{b}_i = \mathbf{A}_{n\times n}^{-1}\mathbf{c}_i^{(n)}$ for all $i \in \{1, \ldots, N\}$, the n original symbols of all N chunks composing $\mathcal{B}$ can be recovered. ■

C. Threat Model and Security Criterion

We consider that during any observation, the goal of an adversary is to recover the original data. We assume the existence of a threat posed by an attacker with the following characteristics:

(i) he observes γ_e out of the γ networks, where γ_e is strictly less than $\gamma - f$, with f denoting the number of tolerated network availability failures, thus obtaining strictly less than n coded symbols of each chunk. For the rest of the paper, we define the security parameter k as the difference between n and the number of coded symbols the eavesdropper observers, that is, $k = (1 - \gamma_e/(\gamma - f))\, n$.

(ii) he has full information about the encoding and decoding schemes, as well as knowledge of matrix $\mathbf{A}$.

(iii) he is capable of guessing the correct value of up to $k-1$ original symbols of a chunk. This assumption empowers the attacker, since a correct guess of a variable in an underdetermined system of equations may leak the value of other variables.

The assumption that the information about the location of the stored data is securely distributed and that its secrecy is maintained, combined with the fact that the data is stored in a very large network, such as the Internet, poses a challenge to an attacker trying to locate the $\gamma - f$ networks required to access a subset of n coded symbols. Note that, although we consider only one eavesdropper, the latter can result from the collusion of several eavesdroppers, which are allowed to cooperate and together obtain $n-k$ coded symbols from a set of networks with size γ_e.

We adopt an information-theoretic secrecy criterion inspired on the work in [11]. Let $\mathbf{X}$ be the vector of original data of size n and $\mathbf{Y}$ be an $(n-k)\times 1$ coded block observed by an attacker.

Definition 1 (Secrecy Criterion (From [11])): The coded block $\mathbf{Y}$ is considered to be secure with respect to m components of $\mathbf{X}$ if the mutual information between $\mathbf{Y}$ and any subset of $\mathbf{X}$ of size m is zero, that is, $I(\mathbf{Y}; \mathbf{X}^{(m)}) = 0$.

That means that any individual symbol is resistant up to $m-1$ guesses [12].

Our goal is to prove that the proposed scheme satisfies the secrecy criterion in *Definition 1*, while ensuring that a legitimate user is able to recover the complete information.

IV. Security Analysis

We now perform the security analysis of our scheme. To accomplish this task, we measure the confidentiality level of the stored information through the mutual information between the original data to be stored and all the linear combinations observed by an attacker. First, we show in *Lemma 1* that the knowledge of $\{\mathbf{c}_1^{(n-k)}, \ldots, \mathbf{c}_{i-1}^{(n-k)}, \mathbf{c}_{i+1}^{(n-k)}, \ldots, \mathbf{c}_N^{(n-k)}\}$ does not increase the information that an attacker has about $\mathbf{b}_i$. That is, the amount of information about the i-th chunk leaked by $n-k$ components of all coded chunks is equal to the amount of information leaked by $n-k$ components of the i-th coded chunk. Then, we demonstrate in *Lemma 2* that an eavesdropper observing any $n-k$ components of a coded chunk $\mathbf{c}_i$ is unable to recover any isolated symbol of $\mathbf{b}_i$, even if he guesses $k-1$ symbols. Finally, *Theorem 1* states, from an information-theoretic perspective, that any linear combination of any $n-k$ components of $\mathbf{c}_i$ does not provide information about any k components of $\mathbf{b}_i$, thus maintaining the property of being secure against up to $k-1$ guesses. Consequently, any extra coding method can be employed over the untrusted networks, while still preserving the security properties of our scheme.

Lemma 1: The mutual information between any subset of m components of a chunk $\mathbf{b}_i$, i.e., $\mathbf{b}_i^{(m)}$, and any $n-k$ components of all the coded chunks, i.e., $\mathcal{C}^{(n-k)} = (\mathbf{c}_1^{(n-k)}, \ldots, \mathbf{c}_N^{(n-k)})^T$, is given by

$$I(\mathbf{b}_i^{(m)}; \mathcal{C}^{(n-k)}) = I(\mathbf{b}_i^{(m)}; \mathbf{c}_i^{(n-k)}). \tag{1}$$

Proof: By applying the definition of mutual information to both the left and right-hand sides of (1), the result holds if and only if $H(\mathbf{b}_i^{(m)}|\mathcal{C}^{(n-k)}) = H(\mathbf{b}_i^{(m)}|\mathbf{c}_i^{(n-k)})$. Let $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^T$ denote a realization of the random vector $\mathcal{C}^{(n-k)}$, i.e., $\mathbf{c}_1^{(n-k)} = \mathbf{x}_1, \ldots, \mathbf{c}_N^{(n-k)} = \mathbf{x}_N$ and $\mathbf{z}_i$ a realization of the random vector $\mathbf{b}_i^{(m)}$. By the definition of conditional entropy, we have that

$$H(\mathbf{b}_i^{(m)}|\mathcal{C}^{(n-k)}) = \sum_{\mathcal{C}^{(n-k)}} H(\mathbf{b}_i^{(m)}|\mathcal{C}^{(n-k)} = \mathcal{X}) P(\mathcal{C}^{(n-k)} = \mathcal{X}). \tag{2}$$

By the definition of entropy we have that

$$H(\mathbf{b}_i^{(m)}|\mathcal{C}^{(n-k)} = \mathcal{X}) = -\sum_{\mathbf{b}_i^{(m)}} P(\mathbf{b}_i^{(m)} = \mathbf{z}_i|\mathcal{C}^{(n-k)} = \mathcal{X}) \log P(\mathbf{b}_i^{(m)} = \mathbf{z}_i|\mathcal{C}^{(n-k)} = \mathcal{X}). \tag{3}$$

Since $\mathbf{c}_j^{(n-k)} = \mathbf{A}_{(n-k)\times n} \cdot \mathbf{b}_j$ and the chunks $\mathbf{b}_j$ are independent for different values of j, we have that $\mathbf{c}_j^{(n-k)}$ and $\mathbf{b}_i^{(m)}$ are independent if $i \neq j$, and that $\mathbf{c}_i^{(n-k)}$ and $\mathbf{c}_j^{(n-k)}$ are also independent if $i \neq j$. Hence,

$$P(\mathbf{b}_i^{(m)} = \mathbf{z}_i|\mathcal{C}^{(n-k)} = \mathcal{X}) = P(\mathbf{b}_i^{(m)} = \mathbf{z}_i|\mathbf{c}_i^{(n-k)} = \mathbf{x}_i). \tag{4}$$

After replacing (4) in (3), we have

$$H(\mathbf{b}_i^{(m)}|\mathcal{C}^{(n-k)} = \mathcal{X}) = H(\mathbf{b}_i^{(m)}|\mathbf{c}_i^{(n-k)} = \mathbf{x}_i). \tag{5}$$

In turn, replacing (5) in (2) leads to

$$H(\mathbf{b}_i^{(m)}|\mathcal{C}^{(n-k)}) = \sum_{\mathbf{c}_i^{(n-k)}} H(\mathbf{b}_i^{(m)}|\mathbf{c}_i^{(n-k)} = \mathbf{x}_i) P(\mathbf{c}_i^{(n-k)} = \mathbf{x}_i) = H(\mathbf{b}_i^{(m)}|\mathbf{c}_i^{(n-k)}),$$

and the result follows. ∎

The result in *Lemma 1* shows that an eavesdropper can only obtain information about chunk $\mathbf{b}_i$ through coded chunk $\mathbf{c}_i$. This also means that an eavesdropper that guesses a whole chunk is unable to use it to obtain information about any other chunk. Thus, we now focus on the information that coded chunk $\mathbf{c}_i$ leaks about an original chunk of data $\mathbf{b}_i$. For simplicity, without loss of generality, we use $\mathbf{b}$ and $\mathbf{c}$ to denote any chunk of original data $\mathbf{b}_i$ and the corresponding encoded data vector $\mathbf{c}_i$.

Lemma 2: Let $\mathbf{c}^{(n-k)}$ be the subvector formed by any $n-k$ components of a coded chunk $\mathbf{c}$. Let μ be the number of symbols that an attacker observing $\mathbf{c}^{(n-k)}$ could guess. Then, if $\mu \leq k-1$, the attacker is unable to recover any additional symbols.

Proof: By observing $n-k$ components of $\mathbf{c}$, the attacker obtains the system of linear equations $\mathbf{c}^{(n-k)} = \mathbf{A}_{(n-k)\times n} \cdot \mathbf{b}$ to solve, where $\mathbf{b}$ is the unknown. Now, suppose that the attacker is able to guess any $k-1$ symbols. Note this is the worst-case scenario – if an attacker cannot obtain additional symbols by guessing any $\mu < k$ symbols, then he cannot recover additional symbols by guessing any $1, \ldots, \mu - 1$ symbols as well. Hence, the cases in which the attacker guesses up to $k-2$ symbols are encompassed in the case that we analyze now. After guessing $k-1$ symbols, the system observed by the eavesdropper $\mathbf{c}^{(n-k)} = \mathbf{A}_{(n-k)\times n} \cdot \mathbf{b}$ can be rewritten as $\mathbf{c}'^{(n-k)} = \mathbf{V} \cdot \mathbf{b}^{(n-k+1)}$, with $k-1$ less unknowns, where $\mathbf{V}$ is of size $(n-k) \times (n-k+1)$. Since $\mathbf{V}$ is a submatrix of $\mathbf{A}$, it preserves the property that any of its square submatrices is nonsingular. Matrix $\mathbf{V}$ has the following reduced row echelon form (RREF):

$$\mathrm{RREF}(\mathbf{V}) = \left[\begin{array}{c|c} & d_1 \\ \mathbf{I}_{n-k} & \vdots \\ & d_{n-k} \end{array}\right].$$

The attacker is able to recover a symbol if and only if there exists $d_i = 0$, for any $i = 1, 2, \ldots, n-k$. We prove by contradiction that this is not the case.

Let us assume that $\exists i, d_i = 0$. Then, the square submatrix $\mathbf{V}'$ of size $(n-k) \times (n-k)$, obtained from the deletion of the i-th column from $\mathrm{RREF}(\mathbf{V})$ is of the form:

$$\mathbf{V}' = \left[\begin{array}{c|c|c} & & d_1 \\ \mathbf{I}_{i-1} & \mathbf{0}_{(i-1)\times(n-k-i)} & \vdots \\ & & d_{i-1} \\ 0 \ldots 0 & 0 \ldots 0 & 0 \\ & & d_{i+1} \\ \mathbf{0}_{(n-k-i)\times(i-1)} & \mathbf{I}_{n-k-i} & \vdots \\ & & d_{n-k} \end{array}\right].$$

Since the i-th row of matrix $\mathbf{V}'$ is a null vector, matrix $\mathbf{V}'$ is singular, and consequently matrix $\mathbf{A}$ contains a singular square submatrix, which is a contradiction. It follows that $\forall i = 1, \ldots, n-k, d_i \neq 0$, and thus, an eavesdropper is unable to recover any other original symbols even after guessing up to $k-1$ symbols of his choice. ∎

Theorem 1: The mutual information between any subset of m components of vector $\mathbf{b}$, i.e., $\mathbf{b}^{(m)}$, and any $n-k$ components of the corresponding coded data vector $\mathbf{c}$, i.e., $\mathbf{c}^{(n-k)}$, is given by

$$I(\mathbf{b}^{(m)}; \mathbf{c}^{(n-k)}) = \begin{cases} 0, & \text{if } m \leq k, \\ (m-k)H(b), & \text{if } m > k. \end{cases} \quad (6)$$

Proof: Since we assume that matrix $\mathbf{A}$ is of public knowledge, the only unknowns for the eavesdropper are the components of $\mathbf{b}$. First, we have that

$$I(\mathbf{b}^{(m)}; \mathbf{c}^{(n-k)}) = H(\mathbf{b}^{(m)}) - H(\mathbf{b}^{(m)}|\mathbf{c}^{(n-k)}).$$

We are now ready to analyze $H(\mathbf{b}^{(m)}|\mathbf{c}^{(n-k)})$ by resorting to the chain rule for entropy. Without loss of generality, we assume that the subset of m components of $\mathbf{b}$ is composed by the first m components of $\mathbf{b}$. Then,

$$\begin{aligned} H(\mathbf{b}^{(m)}|\mathbf{c}^{(n-k)}) &= H(b_1, \ldots, b_m|\mathbf{c}^{(n-k)}) \\ &= \sum_{j=1}^{m} H(b_j|\mathbf{c}^{(n-k)}, b_{j-1}, \ldots, b_1). \end{aligned} \quad (7)$$

Let us first analyze the case in which $m \leq k$. If $j = k$, we have the term

$$H(b_k|\mathbf{c}^{(n-k)}, b_1, \ldots, b_{k-1}). \quad (8)$$

The conditional part of (8) forms a system of linear equations with $(n-k)$ equations and $(n-k+1)$ unknowns. After putting the system in the reduced row echelon form, the i-th equation, where $i = 1, \ldots, n-k$ is now of the type $y_i = b_{k+i-1} + d'_i b_n$, where each y_i results from each c_i after substituting the guessed variables $b_1, \ldots, b_{k-1}$, and after applying the elementary row operations to form the RREF. Note that, according to *Lemma 2*, $d'_i \neq 0$, $i = 1, \ldots, n-k$.

Since b_k and b_j are independent for $k \neq j$, b_k is uniformly distributed in $\mathbb{F}_q$ and $d'_1 \neq 0$, we have that $b_k + d'_1 b_n$ is independent of b_k and therefore, (8) is written as

$$H(b_k|y_1, \ldots, y_{n-k}) = H(b_k|y_1) = H(b_k|b_k + d'_1 b_n) = H(b_k). \quad (9)$$

For $j < k$,

$$H(b_j|\mathbf{c}^{(n-k)}, b_1, \ldots, b_{j-1}) \geq H(b_k|\mathbf{c}^{(n-k)}, b_1, \ldots, b_{k-1}). \quad (10)$$

Since the right-hand side of (10) is equal to $H(b)$, and the left-hand side of (10) must be less than or equal to $H(b)$, $H(b_j|\mathbf{c}^{(n-k)}, b_1, \ldots, b_{j-1}) = H(b)$ for all $j < k$.

For $m > k$, the first k terms are equal to $H(b)$ and the last $(m-k)$ terms of the sum in (7) are equal to zero, since the attacker can form a system with more equations than unknowns. It follows that

$$H(\mathbf{b}^{(m)}|\mathbf{c}^{(n-k)}) = \begin{cases} mH(b), & \text{if } m \leq k, \\ kH(b), & \text{if } m > k. \end{cases} \quad (11)$$

Since b_j, $j = 1, \ldots, m$ are i.i.d. random variables, then $H(b_1, \ldots, b_m) = mH(b)$ and (6) holds. ∎

The proposed scheme can be extended to allow for perfect secrecy as shown in *Corollary 1*.

Corollary 1: Suppose now that $\mathbf{b}$ is of the form $\mathbf{b} = (b_1, b_2, \ldots, b_k, r_1, \ldots, r_{n-k})^T$ where $n-k$ random symbols are appended after k original symbols. The random symbols r_i are i.i.d. uniformly distributed random variables over $\mathbb{F}_q$ and are also independent from the original symbols. Then, if an eavesdropper sees any $n-k$ coded symbols by our

scheme, the set of k original symbols in vector $\mathbf{b}$ is perfectly secure.

Proof: Recall, from *Theorem 1*, that our scheme guarantees that $I(\mathbf{b}^{(m)}; \mathbf{c}^{(n-k)}) = 0$, if $m \leq k$. Since $m = k$, our result holds. ■

Our security scheme also satisfies the notion of block security defined in [29], as shown in *Corollary 2*.

Corollary 2: When the attacker observes $n - k$ coded symbols of a chunk and $k - t$ data symbols are leaked, he does not obtain any information about any additional t components of the corresponding data chunk.

Proof: We show that the entropy of $\mathbf{b}^{(t)}$ given any $n - k$ components of the coded chunk and any $k - t$ guessed symbols is equal to $tH(b)$. Without loss of generality, suppose that the attacker is interested in the first t components of $\mathbf{b}$. The $k - t$ leaked symbols are represented by vector $\mathbf{b}^{(k-t)}$. By applying the chain rule for entropy to $H(\mathbf{b}^{(t)}|\mathbf{c}^{(n-k)}, \mathbf{b}^{(k-t)})$, we have that

$$H(\mathbf{b}^{(t)}|\mathbf{c}^{(n-k)}, \mathbf{b}^{(k-t)}) = \sum_{j=1}^{t} H(b_j|\mathbf{c}^{(n-k)}, \mathbf{b}^{(k-t)}, b_{j-1}, \ldots, b_1). \quad (12)$$

By direct application of *Theorem 1* to (12), each of the t terms in the summation is equal to $H(b)$ and, thus, the result follows. ■

V. Discussion

In this section we discuss several aspects pertaining our security scheme. First, we analyze the implications of tweaking the design parameters n, γ, f on the achieved level of security and reliability. Then, we address the resource allocation and the computational complexity of our scheme. We close this section discussing the implications of having nonuniform input data and the analogies between our work and similar strategies in the literature.

To support our discussion, we present a numerical example. Fig. 2 illustrates the encoding process of a 200 megabyte file to be stored in $\gamma = 12$ networks. The system is designed to be resilient against $f = 2$ network failures. That means that 1/6 of the network contents represent the storage overhead required by fault-tolerance, and hence, 40 megabytes of redundancy symbols have to be added to the 200 megabyte file during the encoding process. First, we split the large file into chunks, each consisting of $n = 160$ symbols drawn uniformly from $\mathbb{F}_q$ with cardinality $q = 2^{16}$. This step generates 655.360 chunks of 320 bytes each. Second, we encode each chunk with matrix $\mathbf{A}$, adding $r = (n \cdot f)/(\gamma - f) = 32$ redundancy symbols. Finally, we store $(n + r)/\gamma = 16$ coded symbols of each chunk in each untrusted network. After repeating this process for all chunks, each network stores 20 megabytes of data. Any user accessing $\gamma - f = 10$ out of 12 networks is able to collect 160 coded symbols of each chunk and, therefore, is able to reconstruct all the chunks required to recover the original file.

A. Security and Reliability Trade-Offs

We now analyze the trade-off between security and reliability when implementing a distributed storage system.

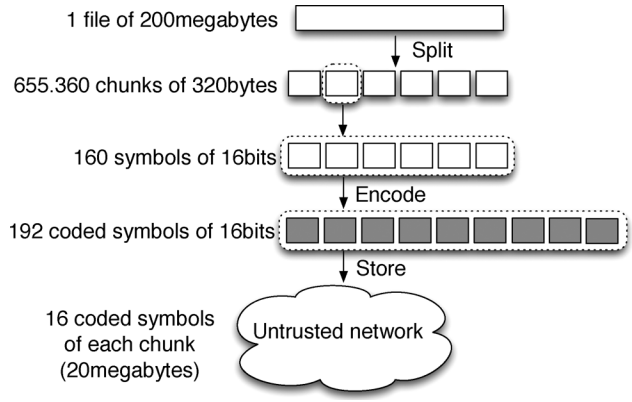

Fig. 2. Encoding setup for storing a 200-megabyte file in a distributed storage system composed by $\gamma = 12$ untrusted networks and resilient to 2 network failures.

TABLE I
Security Level for the Example in Fig. 2 When Varying the Design Parameters, for $\gamma_e = 5$ Eavesdropped Networks

	Fig.2	f	γ	n
Chunk size (n)	160	160	160	**320**
Number of Networks (γ)	12	12	**18**	12
Allowed network failures (f)	2	**4**	2	2
Required redundancy (r)	32	80	20	64
Information stored in each network (MB)	20	25	12.5	20
Security parameter (k)	80	60	110	160

An eavesdropper with access to γ_e networks misses $k = (1 - \gamma_e/(\gamma - f))\, n$ coded symbols from each chunk. Thus, each symbol of a chunk is protected by means of k other symbols, and this is true for all the chunks of coded data, as we showed in Section IV. Equation (13) shows the relationships between the design parameters n, γ, f and the achieved level of security k when an attacker has access to a fixed number of networks γ_e.

$$k = \left(1 - \frac{\gamma_e}{\gamma - f}\right) n. \quad (13)$$

As we increase the number of available networks to store the data, the security also increases, since the attacker has access to a smaller fraction of networks. Also, the amount of storage space required per network decreases. By designing the system to be more fault-tolerant, we decrease the level of security provided, since γ_e becomes closer to the threshold number of networks necessary to recover all of the data. The increase in reliability also implies an increase in the redundancy required, which adds complexity to the encoding procedure. This leads to larger block sizes for the coded symbols, increasing the required storage space. Furthermore, the size of the chunk has a direct relationship with security. An increase in n means an increase in the absolute number of symbols that need to be guessed in order to reveal all of the chunk's data. Like any other coding scheme, large values of n lead to more complexity at the encoder/decoder. Table I illustrates these trade-offs when varying the design parameters n, γ, f on example described in Fig. 2.

These results show that depending on the application requirements, the design parameters can be tweaked to achieve a certain trade-off between security and fault-tolerance. Furthermore, our scheme can be applied on top of any network protocol, including those in which network nodes introduce redundancy (such as

network error correction [30] or fountain codes [31]). This is shown not to decrease the security of the system.

B. Storage Overhead

We recall that each network stores $(n+r)/\gamma$ coded symbols from each one of the N chunks. Thus, under the assumption that each symbol is represented in $\mathbb{F}_q$, each one of the γ networks requires $(n+r)N\log_2(q)/\gamma$ bits of storage space for the actual linear combinations.

In addition to that, each network must also store information about the encoding matrix. Our scheme uses an encoding matrix of size $(n+r)\times n$ without singular square submatrices, which is generated from $2n+r$ distinct coefficients. This information must be stored in $f+1$ networks, since to recover the data, a user accesses $\gamma-f$ networks. Also, each coded symbol corresponds to a specific row of the encoding matrix, which is identified by $\log_2(n+r)$ bits. Thus, we conclude that each one of the γ networks requires $((n+r)/\gamma)\lceil\log_2(n+r)\rceil$ bits of overhead, and $f+1$ of them require an extra of $(2n+r)\log_2(q)$ bits of overhead. Note that the overhead can be easily made negligible by increasing the number of chunks N in which the original data is split.

In the example depicted in *Fig. 2*, 5632 bits are enough to recover all the $n+r$ rows of the encoding matrix, and 128 bits are required to identify $(n+r)/\gamma$ of those rows. Since, for the chosen design parameters, each of the networks stores 20 megabytes of data, this storage overhead accounts for less than 0.0035% of the total amount of required storage space in $f+1$ networks, and less than $7.63\times10^{-5}\%$ in the remaining $\gamma-f-1$ networks.

C. Complexity Analysis

Our system is based on matrices with no singular square submatrices. In particular we analyze two types of matrices that satisfy this property: (i) the result of the multiplication of the inverse of a Vandermonde by a Vandermonde matrix, and (ii) a Cauchy matrix.

The use of structured matrices reduces the computational complexity of inversion and multiplication by vectors. In particular, Vandermonde matrices are parity check matrices for MDS codes, and in that context their structure is used to exploit the lower complexity of matrix and matrix-vector multiplication [28]. For the purpose of our analysis, we consider the algorithms in [32] and we apply them to our scenario along the lines discussed in [28].

Due to the structure of Vandermonde matrices, multiplying the original data vector first by a Vandermonde matrix and then by the inverse of a Vandermonde matrix is more efficient than directly multiplying the original data vector by the product of the two matrices. The multiplication of a Vandermonde (inverse of Vandermonde) matrix by a vector can be translated into a polynomial evaluation (interpolation) problem. There exist efficient algorithms for evaluation (interpolation) of polynomials (points). Taking this relationship into consideration, the product of a Vandermonde matrix by a vector takes $O(n\log^2 n)$ operations. This complexity can be further reduced to $O(n\log n)$ operations when a Fast Fourier transform is employed. The multiplication of the inverse of a Vandermonde by a vector takes $O(n\log^2 n)$. By taking these benchmarks into account, the computational overhead at the source is $O(n\log^2 n)$. At the sink, the decoding complexity of one vector also takes $O(n\log^2 n)$ operations [28]. In general, the use of a Cauchy matrix needs more operations than the use of a system based on Vandermonde matrices, as shown in [28].

D. Non-Uniform Input Data

In the security analysis presented in this paper, we consider that the components of the original data chunks are independent random variables uniformly distributed over $\mathbb{F}_q$, and therefore, there is no redundant information. However, in scenarios in which this assumption is not applicable, e.g., when the information symbols are uncompressed plaintext, the eavesdropper may obtain the original chunk vector $\mathbf{b}$ even if he does not have access to the threshold number of networks required for the uniform case. Note that we can see the eavesdropper observation as the output of a binary erasure channel with a given capacity C. In this context, the joint source and channel coding theorem states that, for large n and if $H(\mathbf{b})<C$, the uncertainty at the eavesdropper can be reduced to zero by using joint source and channel decoding [33].

E. Parallels With Other Techniques

The security analysis from Section IV shows that the proposed scheme satisfies the weak security criterion introduced in the network coding literature [11]. We can trace a parallel to this work by viewing the file encoding step in our scheme as the precoder adopted at the source node in [11]. Our work can also be viewed as:

- the general case of [1], which is a scheme for secure distributed storage where $\gamma=2$ and $\gamma_e=1$. Unlike the work presented in this paper, the scheme proposed in [1] is not tolerant to network failures and does not ensure confidentiality with probability one against partial decoding in a distributed storage system composed by multiple networks.
- a wiretap channel of type II [7] in which the channels from the source to the legitimate user and to the eavesdropper are such that when the legitimate user receives n coded symbols, the eavesdropper observes $n-k$ coded symbols.
- a secret sharing scheme [16], [18] in which the number of secrets is not one but n, and each share has the same bit-length as each secret. Access to any n out of $n+r$ shares reveals all the secrets. However, access to $n-k$ shares does not leak information about any set of k secrets.
- an all-or-nothing transform [22] such that the mutual information between any $n-k$ output values and any k input values is zero.
- the general case of the physical access attack performed on the mobile node used to bootstrap the network considered in [34], from which the adversary obtains $n-1$ coded symbols resulting from the combination of n original symbols in $\mathbb{F}_2$. Thus, in our model, the compromised node from [34] is an untrusted network storing $n-k$ coded symbols, where the security parameter is $k=1$.

VI. Conclusion

We proposed an encoding scheme that exploits the algebraic properties of structured matrices to ensure both confidentiality and fault-tolerance over a set of untrusted and unreliable networks. The scheme uses part of the original information to protect the other part and vice versa. We showed that the proposed approach can be implemented efficiently and is easily applicable to the problem of reliable distributed storage over multiple untrusted networks. Specifically, our theoretical results show that any attack based on the knowledge of k coded symbols (available in a subset of networks) requires the eavesdropper to guess the remaining $n-k$ symbols (stored in the other networks). This is true even if the eavesdropper is interested in acquiring only one information symbol.

First, we showed that an attacker can only obtain information about a chunk through the observation of the corresponding coded chunk. Second, by resorting to the algebraic properties of matrices with no singular square submatrices, we proved the inability of an attacker to perform a Gaussian elimination attack on the observed data. Third, by resorting to information-theoretic arguments, we showed that, for a given chunk, the mutual information between any k original symbols and any $n-k$ symbols observed by the eavesdropper is zero.

In addition, the proposed scheme allows us to share the data with any number of legitimate users. As in network coding based schemes, if one or more users have access to a subset of the untrusted networks (storing k coded symbols of each chunk) but already possess $n-k$ symbols that are linearly independent of the k stored symbols, the presented scheme allows for perfect reconstruction of the original file even if the linear combinations available to the various users are different.

Finally, it is worth noting that the proposed scheme can be used in typical scenarios where a single user wishes to store a file over some public networks that he does not trust, while keeping part of the data in his own local machine. In this case, the user can use the parameter k to tweak simultaneously the level of security and the amount of data that he keeps. Also, our scheme can be applied in content distribution scenarios where users own (possibly different) parts of a file, and wish to receive the remainder of the file in a secure and reliable way via an untrusted and unreliable channel.

Our future work targets the adoption of a stronger threat model involving Byzantine attackers, capable of corrupting some of the coded data stored in multiple untrusted networks.

Acknowledgment

The authors are grateful to Prof. D. Silva (Universidade Federal de Santa Catarina, Brazil), and Dr. T. Abrudan and R. A. Costa (Universidade do Porto, Portugal) for helpful and valuable discussions.

References

[1] P. F. Oliveira, L. Lima, T. T. V. Vinhoza, M. Médard, and J. Barros, "Trusted storage over untrusted networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Miami, FL, Dec. 2010.

[2] M. Armbrust *et al.*, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[3] M. Bloch and J. Barros, *Physical-Layer Security: From Inform. Theory to Security Eng.*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[4] A. G. Dimakis, V. Prabhakaran, and K. Ramchandran, "Decentralized erasure codes for distributed networked storage," *IEEE/ACM Trans. Netw.*, vol. 52, no. 6, pp. 2809–2816, Jun. 2006.

[5] S. Pawar, S. E. Rouayheb, and K. Ramchandran, "On secure distributed data storage under repair dynamics," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Austin, TX, Jul. 2010, pp. 2543–2547.

[6] T. K. Moon, *Error Correction Coding: Math. Methods and Algorithms*. Hoboken, NJ: Wiley, 2005.

[7] L. H. Ozarow and A. D. Wyner, "Wire-tap channel II," *AT&T Bell Labs. Tech. J.*, vol. 63, no. 10, pp. 2135–2157, Dec. 1984.

[8] A. Subramanian and S. W. McLaughlin, MDS Codes on the Erasure-Erasure Wiretap Channel, Arxiv, 2009 [Online]. Available: arXiv:0902.3286

[9] N. Cai and R. W. Yeung, "Secure network coding," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Lausanne, Switzerland, Jul. 2002.

[10] S. E. Rouayheb, E. Soljanin, and A. Sprintson, "Secure network coding for wiretap networks of type II," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1361–1371, Mar. 2012.

[11] K. Bhattad and K. R. Narayanan, "Weakly secure network coding," in *Proc. First Workshop on Network Coding, Theory, and Applicat. (NetCod)*, Riva del Garda, Italy, Apr. 2005.

[12] D. Silva and F. R. Kschischang, "Universal weakly secure network coding," in *Proc. IEEE Inform. Theory Workshop (ITW)*, Volos, Greece, Jun. 2009, pp. 281–285.

[13] L. Lima, M. Médard, and J. Barros, "Random linear network coding: A free cipher?," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Nice, France, Jun. 2007.

[14] N. Cai and R. W. Yeung, "Secure network coding on a wiretap network," *IEEE Trans. Inf. Theory*, vol. 57, no. 1, pp. 424–435, Jan. 2011.

[15] K. Harada and H. Yamamoto, "Strongly secure linear network coding," *IEICE Trans. Fund. Electron., Commun., Comput. Sci.*, vol. 91, no. 10, pp. 2720–2728, 2008.

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

[17] G. Blakley, "Safeguarding cryptographic keys," in *Proc. Nat. Comput. Conf.*, Montvale, NJ, 1979, vol. 48, pp. 313–317, AFIPS Press.

[18] H. Yamamoto, "Secret sharing system using (k, L, n) threshold scheme," *Electron. Commun. Jpn. (Part I: Commun.)*, vol. 69, no. 9, pp. 46–54, 1986.

[19] R. Rivest, "All-or-nothing encryption and the package transform," in *Fast Software Encryption*. New York: Springer, 1997, pp. 210–218.

[20] M. Bellare and P. Rogaway, "Optimal asymmetric encryption," in *Proc. Advances in Cryptology (EUROCRYPT'94)*, 1995, pp. 92–111, Springer.

[21] V. Boyko, "On the security properties of OAEP as an all-or-nothing transform," in *Proc. Advances in Cryptology (Crypto'99)*, 1999, pp. 783–783, Springer.

[22] D. R. Stinson, "Something about all or nothing (transforms)," *Designs, Codes and Cryptography*, vol. 22, no. 2, pp. 133–138, 2001.

[23] M. O. Rabin, "Efficient dispersal of information for security, load balancing, and fault tolerance," *J. ACM*, vol. 36, no. 2, pp. 335–348, Apr. 1989.

[24] H. Krawczyk, "Secret sharing made short," in *Proc. Advances in Cryptology (CRYPTO'93)*, 1994, pp. 136–146, Springer.

[25] A. Bessani, M. Correia, B. Quaresma, F. André, and P. Sousa, "DepSky: Dependable and secure storage in a cloud-of-clouds," in *Proc. EuroSys'11*, Salzburg, Austria, Apr. 2011.

[26] A. Parakh and S. Kak, "A distributed data storage scheme for sensor networks," in *Security and Privacy in Mobile Inform. and Commun. Syst.*. New York: Springer, 2009, pp. 14–22.

[27] A. Parakh and S. Kak, "A distributed data storage scheme for sensor networks," in *Proc. IEEE 3rd Int. Symp. on Advanced Networks and Telecommun. Syst. (ANTS)*, 2009, pp. 1–3.

[28] J. Lacan and J. Fimes, "Systematic MDS erasure codes based on Vandermonde matrices," *IEEE Commun. Lett.*, vol. 8, no. 9, pp. 570–572, Sep. 2004.

[29] S. H. Dau, V. Skachek, and Y. M. Chee, "On the security of index coding with side information," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3975–3988, Jun. 2012.

[30] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4413–4430, Oct. 2004.

[31] M. Luby, "LT codes," in *Proc. 43rd Symp. Found. Comput. Sci.*, Washington, DC, Nov. 2002, pp. 271–280, IEEE Computer Society.

[32] I. Gohberg and V. Olshevsky, "Fast algorithms with preprocessing for matrix-vector multiplication problems," *J. Complexity*, vol. 10, no. 4, pp. 411–427, Dec. 1994.

[33] A. B. Carleial and M. E. Hellman, "A note on Wyner's wiretap channel (Corresp.)," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 387–390, May 1977.

[34] P. F. Oliveira and J. Barros, "A network coding approach to secret key distribution," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 3, pp. 414–423, Sep. 2008.

Paulo F. Oliveira (S'08–M'10) received both the Licenciatura and the Ph.D. degrees in computer science from the Universidade do Porto, Portugal.

He was awarded the Prize Engenheiro António de Almeida (best student award for the Licenciatura degree), and a Doctoral Scholarship from the Portuguese Foundation for Science and Technology. During his Ph.D., he was a researcher at Instituto de Telecomunicações, and a visiting researcher at the Massachusetts Institute of Technology and at the Carnegie Mellon University. His research interests include communication networks, security, network coding, and information theory.

Luísa Lima (S'06–M'12) received the European Ph.D. in computer science (*summa cum laude*) for her work in network coding security and received a best student award for the Licenciatura in Computer Science and Network Engineering, both from Universidade do Porto, Portugal.

She worked as a visiting researcher at the Massachusetts Institute of Technology, USA, Technische Universitaet Muenchen, Germany, and Telefónica Investigación and Desarrollo, Spain. She currently works as a postdoctoral researcher at Instituto de Telecomunicações (IT), where she does research and technology transfer in the area of vehicular networks. Her research interests include wireless systems, vehicular networks, network coding, and security.

Tiago T. V. Vinhoza (S'01–M'07) received the diploma, M.Sc., and Ph.D. degrees in electrical engineering from the Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) in 1999, 2003, and 2007, respectively.

Since 2008, he has been a Research Fellow with the Instituto de Telecomunições (IT), located at the University of Porto. In June 2009 he was awarded a "Ciência 2008" contract by the Portuguese Foundation for Science and Technology (FCT). His research interests include the areas of information and communications theory, signal processing, and their applications in wireless communications and security.

João Barros (S'98–M-04–SM'11) is an Associate Professor with Habilitation ("Agregação") in Electrical and Computer Engineering at the University of Porto and the Founding Director of the Institute for Telecommunications in Porto (IT Porto), Portugal, which includes almost 100 active members. He was a Fulbright scholar at Cornell University and has been a Visiting Professor with the Massachusetts Institute of Technology (MIT) since 2008.

In the past three and half years, he has served as National Director of the Carnegie Mellon Portugal Program, a five-year international partnership funded by the Portuguese Foundation of Science and Technology. The program is focused on information and communication technologies and fosters collaborative research and advanced training among 9 Portuguese universities and 5 associate laboratories, Carnegie Mellon University, and more than 80 companies. In recent years, he has been Principal Investigator (PI) and Co-PI of numerous national, European, and industry-funded projects, coauthoring one book and more than 140 research papers in the fields of networking, information theory and security, with a special focus on smart city technologies, network coding, physical-layer security, sensor networks, and intelligent transportation systems.

Dr. Barros has received several awards, including the 2010 IEEE Communications Society Young Researcher Award for the Europe, Middle East, and Africa region, the 2011 IEEE ComSoC and Information Theory Society Joint Paper Award, and a state-wide best teaching award by the Bavarian State Ministry of Sciences, Research and the Arts. He is frequently invited as an expert speaker by international organizations such as the European Commission, OECD, ITU, EuroDIG, and IEEE. He is a cofounder of Veniam Works', a startup specializing in vehicular mesh networking technologies. He also works as an independent consultant for various organizations and projects. He received his undergraduate education in electrical and computer engineering from the Universidade do Porto (UP), Portugal, and Universitaet Karlsruhe, Germany, a performing arts degree in flute from the Music Conservatory of Porto, and the Ph.D. degree in electrical engineering and information technology from the Technische Universitaet Muenchen (TUM), Germany.

Muriel Médard (S'91–M'95–SM'02–F'08) is a Professor in the Electrical Engineering and Computer Science Department at Massachusetts Institute of Technology (MIT). She was previously an Assistant Professor in the Electrical and Computer Engineering Department and a member of the Coordinated Science Laboratory at the University of Illinois Urbana-Champaign. From 1995 to 1998, she was a Staff Member at MIT Lincoln Laboratory in the Optical Communications and the Advanced Networking Groups. She received B.S. degrees in EECS and in mathematics in 1989, the B.S. degree in humanities in 1990, the M.S. degree in EE in 1991, and the Sc.D. degree in EE in 1995, all from MIT, Cambridge.

She has served as an Associate Editor for the Optical Communications and Networking Series of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, as an Associate Editor in Communications for the IEEE TRANSACTIONS ON INFORMATION THEORY, and as an Associate Editor for the *OSA Journal of Optical Networking*. She has served as a Guest Editor for the IEEE JOURNAL OF LIGHTWAVE TECHNOLOGY, the Joint special issue of the IEEE TRANSACTIONS ON INFORMATION THEORY and the IEEE/ACM TRANSACTIONS ON NETWORKING ON NETWORKING AND INFORMATION THEORY and the IEEE TRANSACTIONS ON INFORMATION FORENSIC AND SECURITY: SPECIAL ISSUE ON STATISTICAL METHODS FOR NETWORK SECURITY AND FORENSICS. She serves as an associate editor for the IEEE/OSA JOURNAL OF LIGHTWAVE TECHNOLOGY. She serves on the board of Governors of the IEEE Information Theory Society as well as being the first Vice President and the President. She has served as TPC cochair of ISIT, WiOpt, and CONEXT. Her research interests are in the areas of network coding and reliable communications, particularly for optical and wireless networks.

Prof. Médard was awarded the 2009 Communication Society and Information Theory Society Joint Paper Award for the paper: Tracey Ho, Muriel Médard, Rolf Kotter, David Karger, Michelle Effros Jun Shi, Ben Leong, "A Random Linear Network Coding Approach to Multicast," IEEE TRANSACTIONS ON INFORMATION THEORY, vol. 52, no. 10, pp. 4413–4430, October 2006. She was awarded the 2009 William R. Bennett Prize in the Field of Communications Networking for the paper: Sachin Katti, Hariharan Rahul, Wenjun Hu, Dina Katabi, Muriel Médard, Jon Crowcroft, "XORs in the Air: Practical Wireless Network Coding," IEEE/ACM TRANSACTIONS ON NETWORKING, vol. 16, no. 3, pp. 497–510, June 2008. She was awarded the IEEE Leon K. Kirchmayer Prize Paper Award 2002 for her paper, "The Effect Upon Channel Capacity in Wireless Communications of Perfect and Imperfect Knowledge of the Channel," IEEE TRANSACTIONS ON INFORMATION THEORY, vol. 46, no. 3, pp. 935–946, May 2000. She was co-awarded the Best Paper Award for G. Weichenberg, V. Chan, M. Médard, "Reliable Architectures for Networks Under Stress", Fourth International Workshop on the Design of Reliable Communication Networks (DRCN 2003), October 2003, Banff, Alberta, Canada. She received an NSF Career Award in 2001 and was co-winner of the 2004 Harold E. Edgerton Faculty Achievement Award, established in 1982 to honor junior faculty members "for distinction in research, teaching and service to the MIT community." In 2007, she was named a Gilbreth Lecturer by the National Academy of Engineering.