Trusted Storage over Untrusted Networks

Paulo F. Oliveira, Luísa Lima, Tiago T. V. Vinhoza, João Barros, and Muriel Médard

***Abstract*—We consider distributed storage over two untrusted networks, whereby coding is used as a means to achieve a prescribed level of confidentiality. The key idea is to exploit the algebraic structure of the Vandermonde matrix to mix the input blocks, before they are stored in different locations. The proposed scheme ensures that eavesdroppers with access to only one of the networks are unable to decode any symbol even if they are capable of guessing some of the missing blocks. Information-theoretic techniques allow us to quantify the achievable level of confidentiality. Moreover, the proposed approach is shown to offer low complexity and optimal rate.**

I. Introduction

Suppose that a user wants to store and share a large file in a distributed fashion yet only has access to multiple networks that he does not trust. A natural question arises: is it possible to store the file in such a way that attackers who only have access to a subset of these networks are unable to reconstruct the file or any of its parts? A standard 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. Such an approach has two obvious drawbacks: (a) computational security does not yet offer provably secure cryptographic primitives, (b) the secret key must be shared with any user who has the right to retrieve the file.

We propose a different technique that relies on coding rather than classical cryptography. Although several contributions have uncovered the advantages of coding techniques in ensuring superior resiliency and flexibility in distributed storage systems [1], few have addressed its potential to provide data confidentiality in untrusted networks. Inspired by recent advances in network coding [2] [3], we show that the aforementioned goal can be achieved without requiring more bandwidth or storage space, while ensuring quantifiable confidentiality levels and dispensing with the need for secret key distribution.

Consider the example shown in *Figure 1*. The aforementioned large file is to be stored in a distributed fashion in two untrusted networks (represented by clouds). Eavesdroppers E_1 and E_2, who are assumed not to collude, only have access to one of the networks yet can observe all of the traffic carried in each of them. To ensure that eavesdroppers E_1 and E_2 are unable to reconstruct the file or any of its parts, our user splits the file into n blocks b_i and encodes each block in a special way. The coding scheme exploits the structure of the Vandermonde matrix, which allows the user to mix the original data in such a way that an attacker is provably unable to recover any individual information symbol — even if it is able to guess part of the file. The next step is to pick $n-k$ code blocks c_i and store them in the first network (shown on the left). The coding scheme ensures that these $n-k$ blocks are protected by means of the other k blocks that are stored in the other network (shown on the right). In turn, these remaining k blocks are protected by the $n-k$ blocks that are only available in the first network. The locations of the blocks can be shared publicly with authorized users who have access to both networks. Since the eavesdroppers E_1 and E_2 only have access to one of the networks albeit different ones, the properties of the Vandermonde matrix prevent them from acquiring data contained in the original file.

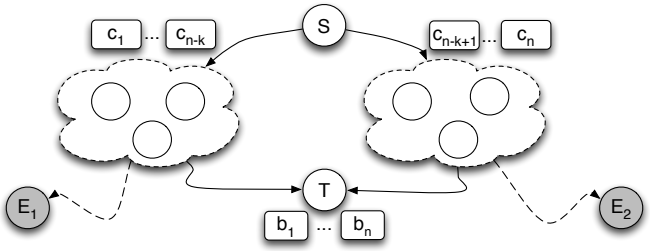

Fig. 1. Example of distributed storage over two untrusted networks.

Our main contribution is thus a distributed storage scheme that exploits the algebraic structure of the Vandermonde matrix to provide the following guarantees:

- *Rate Optimality:* Provided that the original data is compressed in an optimal way, the proposed scheme does not incur on any redundant communication or storage, as is the case with any reference system that requires secret key distribution.
- *Low Complexity:* We demonstrate that the proposed scheme is efficient, in particular in scenarios where network coding is already employed.
- *Widely Applicable:* The coding scheme is not specific to any particular network topology. Furthermore, it can be applied on top of any network protocol, including those in which network nodes introduce redundancy (such as redundant network coding [2] or fountain codes [4]). This is shown not to decrease the security of the system.
- *Quantifiable Level of Security:* Although our scheme does

P. F. Oliveira (pvf@dcc.fc.up.pt) and L. Lima (luisalima@dcc.fc.up.pt) are with Instituto de Telecomunicações, Departamento de Ciência de Computadores, Faculdade de Ciências da Universidade do Porto, Porto, Portugal. T. T. V. Vinhoza (tiago.vinhoza@ieee.org) and J. Barros (jbarros@fe.up.pt) are with Instituto de Telecomunicações, Departamento de Engenharia Electrotécnica e de Computadores, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal. M. Médard (medard@mit.edu) is with Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, United States. This work was partly supported by the Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) under grants SFRH/BD/24718/2005 and SFRH/BD/28946/2006.

not yield perfect secrecy (or unconditional security), we are able to use information-theoretic security arguments to show that the eavesdropper with access to a subset of the coded packets is not able to recover any symbol individually, as introduced in [5]. The proposed system includes a tunable security parameter k which ensures that the encoding scheme is such that no additional symbols are obtained from the publicly known system of equations, even if an eavesdropper observing $n-k$ coded blocks is able to guess up to $k-1$ of the original n blocks.

The rest of the paper is organized as follows. Section II provides a detailed description of relevant related work. The basic methodology is scrutinized in Section III, which explains the coding and the recovering process of the proposed scheme under the adopted intruder model. Section IV then elaborates on the consequences of having compromised coded information and proves that the attacker with access to the untrusted network is indeed ignorant about the original information. After some notes on the system aspects described in Section V, the paper concludes with Section VI.

II. Related Work

Coding techniques were used in [6] to achieve strong secrecy over a channel, in which an eavesdropper acquires a fraction of the transmitted symbols. It was shown that a coset scheme achieves the maximum secret rate albeit at the expense of data rate and constraints on the field size. The maximum number of symbols that can be securely communicated is upper bounded by $n-\gamma$, where n is the total number of symbols transmitted and γ is the number of symbols observed by the eavesdropper. A modified version of the wiretap channel II is considered in [7], 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. A similar problem is considered in [8] in the context of coded networks. The goal is to build a network code that achieves information-theoretic security under the premise that an eavesdropper only has access to a subset of edges in the network that is smaller than the capacity of the network. New bounds for this problem are derived in [9] by modeling the problem as a network generalization of the wiretap channel of type II [6].

A quantifiable security criterion is introduced in [5] to measure the attainable level of secrecy in a multicast scenario, in which an attacker only observes linear combinations of data packets and not the data packets themselves. The main contribution of [5] is a topology dependent scheme in which an encoding matrix at the source is chosen such that an eavesdropper with access to some packets cannot obtain any information about the original plaintext. Although the mutual information between the original information and the coded packets is different from zero, the mutual information measured between a single coded symbol and the original plaintext is shown to be zero. The contribution in [3] generalizes this problem by proposing a fixed outer coding scheme that achieves secure capacity and is universal in the sense that any feasible network code can be used internally without making any assumption about the network. Our problem setup can be viewed as a wiretap channel II in which the user controls which blocks can be intercepted by the eavesdropper. In the following we adopt the security criterion in [5].

As an example of an application, secure distributed storage in sensor networks is considered in [10]. The main idea is to distribute parts of data by different sensors in such a way that each partition is implicitly secure, i.e., reconstruction of the data requires access to a threshold number of sensors that store the data partition. The scheme uses the 3 and 9 roots of a number in a cubic transformation to provide a low complexity confidentiality solution. A technique to hide information without the presence of an encryption key is presented in [11]. The hidden information may be used for validation of shares at the time of secrets reconstruction. The proposed protocol provides methods to share large secrets by dividing the secret in smaller pieces and recursively hiding them in the shares. The problem of determining the secrecy capacity of distributed storage systems against a passive eavesdropper observing a fixed number of nodes is considered in [12]. The problem described in [12] can be translated in an instance of [7] with the diference that in [12] only certain erasure patterns can occur.

III. Problem Setup

We now introduce the notation used in the remainder of the paper. Vectors are represented by lowercase boldface and matrices are represented by capital boldface letters, $\mathrm{diag}\,(x_1, x_2, \ldots, x_n)$ denotes $n \times n$ diagonal matrix with $x_1, x_2, \ldots, x_n$ in the diagonal, and $\mathbf{I}_n$ denotes a $n \times n$ identity matrix. The realization of a random variable x is denoted by $\tilde{x}$. For compactness, we write row i of a matrix $\mathbf{M}$ as $\mathbf{M}_i$. The set of rows ranging from i to l of matrix $\mathbf{M}$ is represented as $\mathbf{M}_{i:l}$, the subvector formed by the positions ranging from i to l of a vector $\mathbf{v}$ is represented as $\mathbf{v}_{i:l}$ and a subset containing any k components of $\mathbf{v}$ is denoted by $\mathbf{v}^{(k)}$.

Let $\mathbf{A}$ (whose elements are $[A_{i,j}] = (a_j^{i-1})$) be a $n \times n$ Vandermonde matrix used for performing coding at the source, where all the coefficients a_i are distributed over all non-zero elements of a finite field $\mathbb{F}_q$, $q = 2^u > n$, and are different from each other, i.e.,$\forall i, l \in \{1, \cdots, n\}, i \neq l \Rightarrow a_i \neq a_l$. Let the original data, or plaintext, be a vector $\mathbf{b} = (b_1, b_2, \ldots, b_n)^T$ whose components b_i, $i = 1, \ldots, n$ are independent random variables uniformly distributed over $\mathbb{F}_q$, with entropy $H(b_i) = H(b)$. Each encoded data vector is represented by $\mathbf{c} = (c_1, \ldots, c_n)^T = \mathbf{A}\mathbf{b}$, where $c_i = \sum_{j=1}^{n} a_j^{i-1} b_j$.

To recover the original information, a legitimate user receives $n-k$ contiguous components of $\mathbf{c}$ from an untrusted network. The remaining k components of $\mathbf{c}$ (or alternatively, k linear combinations of $\mathbf{b}$ that are independent from the combinations present in the first untrusted network) are obtained from the remaining network. Matrix $\mathbf{A}$ is public. To obtain $\mathbf{b}$, the user performs $\mathbf{A}^{-1}\mathbf{c}$.

We consider that during any observation, the ultimate goal of an adversary is to discover the original data. We assume the threat posed by a passive attacker that (i) is able to listen to all the exchanged traffic over the untrusted network and (ii) has full information about the encoding and decoding schemes, as well as knowledge of matrix $\mathbf{A}$.

To explain the security metric, we first trace a parallel between our scheme and a classical cryptographic framework. The original data $\mathbf{b}$ is first encoded with matrix $\mathbf{A}$, and then divided in two different ciphertexts $\mathbf{A}_{1:n-k} \cdot \mathbf{b} = \mathbf{c}_{1:n-k}$ and $\mathbf{A}_{n-k+1:n} \cdot \mathbf{b} = \mathbf{c}_{n-k+1:n}$. The first ciphertext can be viewed as the result of the encryption of a subset of $n-k$ components of vector $\mathbf{b}$ (i.e., $\mathbf{b}^{(n-k)}$) using a *key* which is a function $f(\mathbf{b}^{(k)})$ of the remaining $\mathbf{b}^{(k)} = \mathbf{b}^{(n)} \setminus \mathbf{b}^{(n-k)}$ original symbols. The interpretation for the second ciphertext is similar, except that in this case, a *key* of size $n-k$ is protecting k symbols.

We adopt an information-theoretic secrecy criterion inspired by [5]. Let $\mathbf{X}$ be the vector of original data of size n and $\mathbf{Y}$ be an $(n-k) \times 1$ ciphertext vector.

Definition 1 (Secrecy Criterion (from [5])): The ciphertext $\mathbf{Y}$ 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 [3].

The goal of the problem 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. First, we show in *Lemma 1* that an eavesdropper observing any $n-k$ contiguous components of $\mathbf{c}$ is unable to recover any isolated symbol, even if it guesses $k-1$ symbols. Then, we perform an information-theoretic analysis of the scheme in *Theorem 1*.

Lemma 1: Let y be the number of symbols that an attacker observing $\mathbf{c}_{p+1:p+n-k}$ could guess, where $0 \leq p \leq k$. Then, if $y \leq k-1$, the attacker is unable to recover any additional symbols.

Proof: First, assume that an eavesdropper observes the first $n-k$ rows of $\mathbf{c}$. Then, it obtains the system of linear equations $\mathbf{A}_{1:n-k} \cdot \mathbf{b} = \mathbf{c}_{1:n-k}$ to solve, where

$$\mathbf{A}_{1:n-k} = \begin{bmatrix} a_1^0 & a_2^0 & \dots & a_n^0 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{n-k-1} & a_2^{n-k-1} & \dots & a_n^{n-k-1} \end{bmatrix}.$$

Now, suppose that the attacker is able to guess $k-1$ symbols. Note this is the worst-case scenario – if an attacker cannot obtain additional symbols by guessing any x symbols, then it cannot recover additional symbols by guessing any $x-1, \dots, 1$ symbols as well. Hence, the cases in which the attacker guesses $0 \dots k-2$ symbols are encompassed in the case that we analyze now. For each one of the $\binom{n}{k-1}$ possible combinations resulting from $k-1$ guesses, the attacker obtains a sub-matrix $\mathbf{V}$ of size $(n-k) \times (n-k+1)$, which preserves the structure of the Vandermonde matrix. Thus, after guessing $k-1$ symbols, the system observed by an eavesdropper $\mathbf{A}_{1:n-k} \cdot \mathbf{b} = \mathbf{c}_{1:n-k}$ can be rewritten as $\mathbf{V} \cdot \mathbf{b}' = \mathbf{c}'_{1:n-k}$, with $k-1$ less unknowns. Without loss of generality, we assume that the attacker guesses the last $k-1$ symbols of vector $\mathbf{b}$ (from component $n-k+2$ to n):

$$\mathbf{V} = \begin{bmatrix} a_1^0 & a_2^0 & \dots & a_{n-k+1}^0 \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{n-k-1} & a_2^{n-k-1} & \dots & a_{n-k+1}^{n-k-1} \end{bmatrix}.$$

Then, we take the reduced row echelon form of $\mathbf{V}$:

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

where

$$d_m = \prod_{\substack{i=1 \\ i \neq m}}^{n-k} (a_i + a_{n-k+1}) \Big/ \prod_{\substack{i=1 \\ i \neq m}}^{n-k} (a_i + a_m)\ . \tag{1}$$

The result in (1) follows from the fact that a vector of the form $\mathbf{d} = [d_1, d_2, \dots, d_{n-k}, -1]^T$ lies in the nullspace of RREF($\mathbf{V}$). Hence it also lies in the nullspace of $\mathbf{V}$. If $\forall m \in \{1, \cdots, n-k\}, d_m$ is given by (1) then the condition that $\mathbf{Vd} = 0$ is satisfied. The reduced row echelon form is unique and if the attacker was able to recover any symbol, the reduced row echelon form of $\mathbf{V}$ would include at least one row of the form $[0, \dots 0, 1, 0 \dots, 0]$. We now show that this is impossible. Each element a_j in (1) is different from zero and $a_i \neq a_j \forall i, j$. Since the characteristic of the field is 2, d_i could only be 0 if $a_i = a_{n-k+1}$, thus $d_i \neq 0, \forall i$. It follows that an eavesdropper is unable to recover any other original symbols if it can guess up to $k-1$ symbols. Suppose now that, instead of having access to the first $n-k$ rows of $\mathbf{c}$, an eavesdropper observes any $n-k$ contiguous components of $\mathbf{c}$, obtaining the system of linear equations $\mathbf{A}_{p+1:p+n-k} \cdot \mathbf{b} = \mathbf{c}_{p+1:p+n-k}$, where $0 \leq p \leq k$ and

$$\mathbf{A}_{p+1:p+n-k} = \begin{bmatrix} a_1^p & a_2^p & \dots & a_n^p \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{p+n-k-1} & a_2^{p+n-k-1} & \dots & a_n^{p+n-k-1} \end{bmatrix}.$$

Once again, we can consider, without loss of generality, that the attacker is able to guess the last $k-1$ components of vector b. Then, he can eliminate the last $k-1$ columns of matrix $\mathbf{A}_{p+1:p+n-k}$, obtaining matrix

$$\mathbf{V}' = \begin{bmatrix} a_1^p & a_2^p & \dots & a_{n-k+1}^p \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{p+n-k-1} & a_2^{p+n-k-1} & \dots & a_{n-k+1}^{p+n-k-1} \end{bmatrix},$$

which can be written as a function of matrix $\mathbf{V}$ as $\mathbf{V}' = \mathbf{VE}$, where $\mathbf{E}$ is a matrix of size $(n-k+1) \times (n-k+1)$, with

the following structure:

$$\mathbf{E} = \text{diag}\left(a_1^p, a_2^p, \ldots, a_{n-k}^p, a_{n-k+1}^p\right).$$

Consider a matrix $\mathbf{Q} = \mathbf{Q}_1\mathbf{Q}_2$ such that $\text{RREF}(\mathbf{V}') = \mathbf{QV}' = \mathbf{QVE}$ and $\text{RREF}(\mathbf{V}) = \mathbf{Q}_2\mathbf{V}$. Thus, $\text{RREF}(\mathbf{V}') = \mathbf{Q}_1\text{RREF}(\mathbf{V})\mathbf{E}$. Let r_j be the j^{th} row of matrix $\text{RREF}(\mathbf{V})\mathbf{E}$. From the definition of matrix $\mathbf{E}$, we can easily show that by only applying the following elementary row operation $r_j = r_j/a_j^p$, $j = 1, \ldots, n-k$, we obtain its reduced row echelon form

$$\text{RREF}(\mathbf{V}') = \left[\begin{array}{c|c} & \frac{a_{n-k+1}^p d_1}{a_1^p} \\ \mathbf{I}_{n-k} & \vdots \\ & \frac{a_{n-k+1}^p d_{n-k}}{a_{n-k}^p} \end{array}\right].$$

Since $\forall j \in \{1, \ldots, n-k\}$ both a_j^p and d_j are non-zero and since a_{n-k+1}^p is also non-zero, we have $\frac{a_{n-k+1}^p d_j}{a_j^p} \neq 0$. ■

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

$$I(\mathbf{b}^{(m)}; \mathbf{c}_{p+1:p+n-k}) = \begin{cases} 0, & \text{if } m \leq k \\ (m-k)H(b), & \text{if } m > k \end{cases} \quad (2)$$

Proof: Without loss of generality, we pick the first $n-k$ rows of $\mathbf{c}$. We also assume that matrix $\mathbf{A}$ is of public knowledge, so the only unknowns for the eavesdropper are the components of $\mathbf{b}$. First, we have that

$$I(\mathbf{b}^{(m)}; \mathbf{c}_{1:n-k}) = H(\mathbf{b}^{(m)}) - H(\mathbf{b}^{(m)}|\mathbf{c}_{1:n-k}).$$

We are now ready to analyze $H(\mathbf{b}^{(m)}|\mathbf{c}_{1:n-k})$ 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:

$$\begin{aligned} H(\mathbf{b}^{(m)}|\mathbf{c}_{1:n-k}) &= H(b_1, \cdots, b_m|\mathbf{c}_{1:n-k}) \\ &= \sum_{j=1}^{m} H(b_j|\mathbf{c}_{1:n-k}, \tilde{b}_{j-1}, \cdots, \tilde{b}_1) \quad (3) \end{aligned}$$

where conditioning on $\tilde{b}_i$ means conditioning on the random variable b_i being equal to $\tilde{b}_i$. Let us first analyze the case in which $m \leq k$. If $j = k$, we have the term

$$H(b_k|\mathbf{c}_{1:n-k}, \tilde{b}_1, \cdots, \tilde{b}_{k-1}). \quad (4)$$

The conditional part of (4) forms the following system of equations:

$$\left[\begin{array}{ccc} a_k^0 & \cdots & a_n^0 \\ \vdots & \ddots & \vdots \\ a_k^{n-k-1} & \cdots & a_n^{n-k-1} \end{array}\right] \left[\begin{array}{c} b_k \\ \vdots \\ b_n \end{array}\right] = \left[\begin{array}{c} z_1 \\ \vdots \\ z_{n-k} \end{array}\right], \quad (5)$$

where $z_i = c_i - (a_1^{i-1}\tilde{b}_1 + \cdots + a_{k-1}^{i-1}\tilde{b}_{k-1})$. Putting the system in the reduced row echelon form, the i-th equation, $i = 1, \ldots, n-k$, is now of the type $y_i = b_{k+i-1} + d_i' b_n$, where each y_i results from the elementary row operations to form the RREF and $d_1' \cdots d_{n-k}'$ can be obtained via Lemma 1. Now, we have that

$$H(b_k|y_1, \cdots, y_{n-k}) = H(b_k|y_1) = H(b_k|b_k + d_1' b_n) = H(b_k) \quad (6)$$

since b_i and b_j are independent for $i \neq j$ and b_k is uniformly distributed in $\mathbb{F}_q$. Therefore, $b_k + d_1' b_n$ is independent of b_k. For $j < k$,

$$H(b_j|\mathbf{c}_{1:n-k}, \tilde{b}_1, \cdots, \tilde{b}_{j-1}) \geq H(b_k|\mathbf{c}_{1:n-k}, \tilde{b}_1, \cdots, \tilde{b}_{k-1}). \quad (7)$$

Since the RHS of (7) is equal to $H(b)$, equality holds. Hence, $H(b_j|\mathbf{c}_{1:n-k}, \tilde{b}_1, \cdots, \tilde{b}_{j-1}) = H(b)$ for all $j < k$. For $m > k$, the first k terms are equal to $H(b)$ and the last $(m-k)$ terms of equation (3) are equal to zero, since the attacker can form a system with more equations than unknowns. It follows that:

$$H(\mathbf{b}^{(m)}|\mathbf{c}_{1:n-k}) = \begin{cases} mH(b) & \text{if } m \leq k \\ kH(b) & \text{if } m > k \end{cases} \quad (8)$$

Since b_j, $j = 1, \ldots, m$ are i.i.d. random variables, then $H(b_1, \ldots, b_m) = mH(b)$ and (2) holds. ■

Lemma 1 shows that an attacker observing $n-k$ contiguous positions of a vector encoded with matrix $\mathbf{A}$ is unable to perform Gaussian elimination on the matrix to recover any symbol even if he uses up to $k-1$ guesses. It is also important to note that the size of the field needs to be strictly greater than the size of the matrix n. Moreover, it is easy to see from *Theorem 1* that any linear combination of any contiguous $n-k$ components of $\mathbf{c}$ maintains the property of being secure against up to $k-1$ guesses. Thus, any network coding method can be employed in the untrusted network, while still preserving the security properties of our scheme.

V. System Aspects

We now discuss several system aspects pertaining our security scheme. We analyze the computational complexity of our scheme and then compare it to other weak security strategies.

A. Computational Complexity

The use of a Vandermonde matrix reduces the computational complexity of inversion and multiplication by vectors. Note that Vandermonde matrices are parity check matrices for MDS codes, and in that context its structure was used to feature lower complexity matrix and matrix-vector multiplication [13]. For the purposes of our analysis, we consider the algorithms in [14]. The complexity is measured in algebraic operations. Inversion takes $O(n^2)$ operations for a $n \times n$ matrix. Matrix-vector multiplication takes $O(n \log^2 n)$ operations. By taking these benchmarks into account, the computational overhead at the source is $O(n^2)$. The generation of the first two rows of the Vandermonde matrix can be deemed to be negligible; then, generating the remainder of the matrix takes $O(n^2)$ multiplications in $\mathbb{F}_q \backslash \{0\}$. The source then generates the coded vectors by multiplying the matrix and the plaintext, which takes $O(n \log^2 n)$ operations. At the sink, in the worst case scenario in which all packets are encoded, the complexity is

$O(n^2)$ multiplication operations. In applications that use network coding, the proposed scheme does not add to the overall complexity of the system, because Gaussian elimination is already required for the retrieval of the stored data.

B. Comparison with Competing Coding Techniques

The work in [5] establishes the necessary conditions to achieve the defined secrecy criterion, given the topology and the used network code. However, the techniques to find a matrix that satisfies such requirements are arguably too high in terms of computational complexity. The scheme presented in [3] is independent of both the topology and the network code used, but it suffers from similar drawbacks as [5] in terms of finding such a matrix for an arbitrary number of guesses. As seen above, our approach has some practical advantages over the existing ones.

The work in [15] derives bounds for the probability of decoding an individual symbol in a network where Random Linear Network Coding (RLNC, i.e., random mixings of packets at the intermediate nodes of the network) is used. It is shown that RLNC increases the security for a threat model in which the intermediate nodes comply with the protocol however may try to decode as much as possible. Although randomness does not seem sufficient to provide confidentiality with probability one against partial decoding, our coding scheme accomplishes this goal while keeping low requirements on the amount of resources needed. Finally, our model is the general case of the physical access attack performed on the node used to bootstrap the network considered in [16], from which the adversary obtains $n-1$ coded blocks containing n original blocks in $\mathbb{F}_2$. Thus, in our framework, the compromised central node is an untrusted network storing $n-k$ coded symbols, where the security parameter is $k=1$.

VI. Conclusion

We proposed an encoding scheme for achieving confidentiality based on the structure of the Vandermonde matrix. The scheme relies on part of the original information to protect the other part and vice versa. We showed that the proposed approach offers low computational complexity and is easily applicable to distributed storage scenarios with two untrusted networks. Specifically, our theoretical results prove that any privacy attack based on k blocks (available in one network) requires the eavesdropper to guess the remaining $n-k$ blocks (stored in the other network). This is true even if the eavesdropper is interested in acquiring only one information symbol.

In addition, the scheme allows us to share the data with any number of valid users. As in other network coding schemes, if one or more users have access to only one of the untrusted networks (storing k blocks) but already possess $n-k$ blocks that are linearly independent of the k stored blocks, the presented distributed storage 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 some file in an untrusted network 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. The structure of the Vandermonde matrix assures that any k distinct blocks available locally are linearly independent of the remaining $n-k$ blocks that are stored in the untrusted network.

Our ongoing work targets the extension of our distributed storage scheme from two to multiple untrusted networks, as well as the adoption of a stronger threat model involving Byzantine attackers.

Acknowledgment

We are grateful to Dr. Danilo Silva (State University of Campinas, Brazil) and Rui A. Costa (Universidade do Porto, Portugal) for helpful and valuable discussions.

References

[1] A. Dimakis, V. Prabhakaran, and K. Ramchandran, "Decentralized erasure codes for distributed networked storage," *IEEE/ACM Transactions on Networking (TON)*, vol. 14, no. SI, p. 2816, 2006.

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

[3] D. Silva and F. R. Kschischang, "Universal weakly secure network coding," in *Networking and Information Theory, 2009. ITW 2009. IEEE Information Theory Workshop on*, June 2009, pp. 281–285.

[4] M. Luby, "LT codes," in *Proceedings of the 43rd Symposium on Foundations of Computer Science*. IEEE Computer Society, 2002, p. 271.

[5] K. Bhattad and K. Narayanan, "Weakly secure network coding," *Proc. of the First Workshop on Network Coding, Theory, and Applications (NetCod)*, April 2005.

[6] L. H. Ozarow and A. D. Wyner, "Wire-tap channel II," *AT&T Bell Labs.Tech. J.*, pp. 2135–2157, Dec. 1984.

[7] A. Subramanian and S. McLaughlin, "MDS codes on the erasure-erasure wiretap channel," *Arxiv preprint arXiv:0902.3286*, 2009.

[8] N. Cai and R. Yeung, "Secure network coding," in *Proceedings of the IEEE International Symposium on Information Theory*, Lausanne, Switzerland, July 2002.

[9] S. Y. E. Rouayheb, E. Soljanin, and A. Sprintson, "Secure network coding for wiretap networks of type ii," *CoRR*, vol. abs/0907.3493, 2009.

[10] A. Parakh and S. Kak, "A Distributed Data Storage Scheme for Sensor Networks," in *Security and Privacy in Mobile Information and Communication Systems: First International ICST Conference, MobiSec 2009, Turin, Italy, June 3-5, 2009, Revised Selected Papers*. Springer, 2009, p. 14.

[11] ——, "Recursive Secret Sharing for Distributed Storage and Information Hiding," *Arxiv preprint arXiv:1001.3331*, 2010.

[12] S. Pawar, S. El Rouayheb, and K. Ramchandran, "On Secure Distributed Data Storage Under Repair Dynamics," *Arxiv preprint arXiv:1003.0488*, 2010.

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

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

[15] L. Lima, M. Médard, and J. Barros, "Random Linear Network Coding: A Free Cipher?" in *Proc. of the IEEE International Symposium on Information Theory (ISIT)*, June 2007.

[16] P. F. Oliveira and J. Barros, "A network coding approach to secret key distribution," *IEEE Transactions on Information Forensics and Security*, vol. 3, pp. 414–423, September 2008.