Secure Network Coding Against the Contamination and Eavesdropping Adversaries

Yejun Zhou, Hui Li and Jianfeng Ma
Key lab of CNIS Ministry of Education
Xidian University
Xi’an, Shanxi 710071, China
Email: {yjzhou, lihui, jfma}@mail.xidian.edu.cn

Abstract—In this paper, we propose an algorithm that targets contamination and eavesdropping adversaries. We consider the case when the number of independent packets available to the eavesdropper is less than the multicast capacity of the network. By means of our algorithm every node can verify the integrity of the received packets easily and an eavesdropper is unable to get any “meaningful information” about the source. We call it “practical security” if an eavesdropper is unable to get any meaningful information about the source. We show that, by giving up a small amount of overall capacity, our algorithm achieves achieves the practically secure condition at a probability of one, which is much higher than that of Bhattad and Narayanan’s [1]. Furthermore, the communication overhead of our algorithm are negligible compared with previous works, since the transmission of the hash values and the code coefficients are both avoided.

I. INTRODUCTION

The concept of network coding was first introduced by Ahlswede et al. [2]. They showed that multicast rates could be increased by allowing for network coding instead of just routing. Shortly afterwards, Li, Yeung and Cai [3] showed that it is sufficient for the encoding functions at the interior nodes to be linear. Ho et al. [4] and [5] proposed a random coding scheme in which the message on outgoing edges of a node are chosen to be a random linear combination of the message on its incoming edges.

In reality, network transmission may suffer from two kinds of adversaries: contamination and eavesdropping. Network coding has been studied to conquer these two kinds of adversaries. Contamination and eavesdropping adversaries. We consider the problem of network coding in the presence of Byzantine attacker. Gkantsidis et al. [7] also considered the related problem. Jaggi et al. [8] designed a resilient network coding algorithm which is information-theoretically secure and rate-optimal for different adversarial strengths. Homomorphic hashing function was first proposed in [9], which allows nodes to check blocks on-the-fly in a system where content is encoded at the source using rateless codes. However, the total size of the hash values of their scheme is proportional to the number of blocks, which could be very large and the cryptographic hash function is computationally expensive. Li et al. [10] employed a batch content distribution verification scheme, which reduced the computational cost of each node to cache and scan all the received packets when computing a new packet. The cryptographic hash function of their scheme is computationally inexpensive compared with which in [9]. Unfortunately, their scheme deviate from the classical network coding scheme, which is bandwidth consumed and delay could be induced at the sinks. On the other hand, although batching can decrease the computation time, batching block verification has the risk of letting some malicious packets propagate since packets are exchanged without being checked. Thus, standard batching techniques do not work well with network coding. Zhao et al. [11] presented a signature scheme with low computation, but their scheme required long start-up latency. Finally, all the works presented above have to distribute the coefficients which is bandwidth consumed.

Cai and Yeung [12] considered the problem of using network coding to achieve perfect information security against an eavesdropper who can eavesdrop on a limited number of network links, and presented the construction of a secure linear network code for this purpose. A similar problem was considered in [13] featuring a random coding approach in which only the input vector is modified.

Bhattad and Narayanan [1] first defined a model for security that is more suitable for practical applications. In this paper, we also consider this type of model, which is not information theoretically secure, but is secure enough for the application. An interesting observation made in [14] was that for a computation limited eavesdropper with the use of one way function it is possible to transmit at a high rate without the eavesdropper getting any meaningful information about the source. A more general threat posed by intermediate nodes was considered in [15].

In this paper, we consider these two kinds of adversaries at the same time, that is, the adversary can contaminate the transmission on a subset of channels, and at the same time eavesdrop on another subset of channels with cardinality less than or equal to the multicast capacity of the network. Ngai and Yang [16] studied the similar problem and constructed a secure error-correcting network codes.

The main contribution of this paper is to propose an algorithm, which can not only verify the integrity of the received packets easily but also achieve the practically secure condition at a probability of one. In our scheme, we use the public parameters as the “intended” hash values. The original packets are padded so that they are hashed to the public parameters. In this way the transmission of the hash values
is avoided. The code coefficients in our scheme are generated in a pseudo-random number generator in each node, so the distribution of the coefficients is also avoided. We show that the communication overhead and the start-up latency are negligible since the transmission of the hash values and the coefficients are both avoided.

This paper is organized as follows. In the next section we give the notations used in this paper. The secure network coding scheme is proposed in section III. In Section IV, we present the security of our algorithm. Overhead and start-up latency of our scheme are discussed in Section V. Finally, this paper is concluded in section VI.

II. NETWORK MODEL AND NOTIONS

In this paper, we assume that all the messages and coefficients are generated in $\mathbb{F}_p$, where $p$ is a large enough prime number. We shall use small letters $x, y$ etc. to denote vectors whose dimensions will be clear from the context. The matrices are denoted by the capital letters such as $X, \tilde{X}$ etc. The transpose operator of vectors and matrices will be denoted by “$^T$" thus $x^T$ will stand for column vectors.

A. Network Model

We represent a network by a directed graph $G = (V; E)$, where $V$ is the set of vertices (nodes) and $E$ is the set of edges (channels). We assume an order on $V$ which is consistent with the associated partial order on $G$. A network code is said to be linear if the message on any outgoing edge of any node is a linear combination of the messages on the incoming edges of the node.

In this paper, we assume that the source node sends information $X$ of the following form:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{pmatrix} = \left( \begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_m \end{array} \right)$$

We call $x_i, i = 1, \ldots, m$ a packet.

Therefore, for a linear code the message on edge $e_j \in E$ can be written as $F_{e_j}X$ where $F_{e_j}$ is a length $m$ vector over $\mathbb{F}_p$(we call it global encoding kernel in this paper) on edge $e_j \in E$.

B. Threat Model

There is a source, Alice, and a destination, Bob, who communicate over a wired or wireless network. There is also an eavesdropper Calvin, hidden somewhere in the network. He aims to eavesdrop on the transfer of information from Alice to Bob and injects his own. A malicious node can generate corrupted packets and then distribute them to other nodes, which in turn use them to (unintentionally) create new encoded packets that are also corrupted. A wiretap network is specified by a collection $\mathcal{A}$ of sets of edges $\mathcal{A} = \{A_1, A_2, \ldots, A|\mathcal{A}|\}$, $A_i \in E$. Calvin selects a particular set $A_i \in \mathcal{A}$ and listens to all messages transmitted on edges in $A_i$ to get some information. We assume that the set doesn’t change with time. When we are specified a linear code and a wiretap network we use $A_i$ to represent a matrix whose rows contain all linearly independent global encoding kernel corresponding to edge $e_j \in A_i$. In this case, the messages available to Calvin is $A_iX$. The number of rows in $A_i$ is represented by $k_i$. We define $k$ as max$_i k_i$.

C. Notions

1) : The network capacity is the time-average of the maximum number of packets that can be delivered from Alice to Bob, assuming no adversarial interference, i.e., the max-flow. To simplify notion, in this paper, we assume the max-flow from Alice to Bob is $m$.

2) : Practical security: Consider a set of messages $M$. Let $U$ be subset of the set containing the multicast information $X$. We say that $M$ has no information about $U$ if $I(U; M) = 0$. We say that $M$ has no meaningful information about $U$ if $I(X; M) = 0$, $\forall x_i \in U$. In this paper we concentrate on two special cases and generalize the results towards the end. We say that Calvin has no information about the source if $I(X; M) = 0$ where $M$ is the set of messages that Calvin chooses to observe. The security condition considered by Cai and Yeung [12] falls in this category. We will use Shannon security to refer to this security requirement. The second case we consider is when Calvin gets no meaningful information about the source i.e.$I(X; M) = 0, \forall x_i$ for messages $M$ observed by Calvin. We call this type of security as practical security.

It is noted that if Alice transmits a linear transformation of $X$, $PX$, instead of $X$ then the message transmitted on edge $e_j$ would be $F_{e_j}PX$($P$ is a $m \times m$ matrix which is unknown to Calvin). In this case, although Calvin has some information about the source he is unable to get any meaningful information.

As shown in Fig.1, let us assume that Calvin can listen to any one edge of this network. The multicast capacity for this network is $2$. $x_1$ and $x_2$ are the messages of Alice. In Fig.1 (a), w is a uniform random sequence independent of the messages. This is an example of the coding scheme constructed by Cai.
and Yeung [12]. Obviously, the maximum multicast capacity supported is 1 when this system has to be Shannon secure. When the security condition is relaxed to practical security, as shown in Fig.1 (b), the max-flow can be achieved.

III. SECURE NETWORK CODING

A. The Homomorphic Hash Function

We first choose the hash parameters \( q, g \). Let \( o(x) \) denote the order of \( x \) in the field \( \mathbb{F}_q \). Here we choose \( o(g) = p \) in \( \mathbb{F}_q \) (where \( \mathbb{F}_q \) is a subfield of \( \mathbb{F}_p \)). Furthermore we randomly select \( n + 2 \) numbers \( u_0, u_1, \ldots, u_n, u_{n+1} \) from \( \mathbb{F}_p \). Next, we compute

\[
g_i = g^{u_i} \pmod{q}
\]

for all \( 0 \leq i \leq n + 1 \). The public parameter of the hash function is \( p, q, g, 0, g_1, \ldots, g_{n+1} \). Whereas \( u_0, u_1, \ldots, u_{n+1} \) and \( g \) should be kept secret.

Formally, we define \( DL(G, p, q) \) to be the computational problem: Given \( g, p \) and \( q \), where \( o(g) = p \) in \( \mathbb{F}_p \), find \( x \) such that \( y = g^{x} \pmod{q} \). Hence, we have

**Lemma 1:** Given \( g, q, g_1, \ldots, g_{n+1} \), and the public parameters \( p, q \), it is computationally infeasible for a node to find \( u_0, u_1, \ldots, u_{n+1} \), such that \( g_i = g^{u_i} \pmod{q} \) if \( DL(G, p, q) \) is hard.

Assume that each message is of the form:

\[
x = (x_0, x_1, \ldots, x_n, r)
\]

where \( x_i, r \in \mathbb{F}_p \) for \( 0 \leq i \leq n \), and the hash of \( x \) is computed as

\[
\mathcal{H}(x) = \prod_{i=0}^{n} g_i^{x_i} g_{i+1} \pmod{q}
\]

Based on this construction, we have

\[
\mathcal{H}(x) = g^{\left(\sum_{i=0}^{n} u_i x_i + u_{n+1} r\right)} \pmod{p} \pmod{q}
\]

For any two messages \( x = (x_0, x_1, \ldots, x_n, r_1) \) and \( y = (y_0, y_1, \ldots, y_n, r_2) \), we define the addition of \( x \) and \( y \) as

\[
x + y = (z_0, z_1, \ldots, z_n, r)
\]

where \( r = (r_1 + r_2) \pmod{p} \) and \( z_i = (x_i + y_i) \pmod{p} \) for \( 0 \leq i \leq n \).

Hence, this hash function has the following homomorphic property

\[
\mathcal{H}(x) \cdot \mathcal{H}(y) = \mathcal{H}(x + y)
\]

The security of \( \mathcal{H} \) is defined in terms of the difficulty in finding collisions. It can be shown that the hash function is indeed collision-free if \( DL(G, p, q) \) is hard. In particular we have:

**Lemma 2:** The hash function \( \mathcal{H} \) is collision-free (namely it is computationally infeasible to find two different messages \( x_1 \) and \( x_2 \) such that \( \mathcal{H}(x_1) = \mathcal{H}(x_2) \)) if \( DL(G, p, q) \) is hard.

It can be proved that the hash function is indeed collision-free, using an argument in [17] (proof of Theorem 3.4).

B. Alice’s Encoder and Bob’s Decoder

**Alice’s encoder:** Alice encodes \( X \) in the following steps.

She first chooses \( m \) parity symbols \( r_i, d \), for \( d \in \{1, \ldots, m\} \) uniformly at random from the field \( \mathbb{F}_p \), and then generates a Vandermonde matrix \( P \) as follows

\[
P = \begin{pmatrix}
r_1 & r_2 & \cdots & r_m \\
r_1 & r_2 & \cdots & r_m \\
\vdots & \vdots & \ddots & \vdots \\
r_1 & r_2 & \cdots & r_m
\end{pmatrix}
\]

In the second step, Alice per-multiplies the source message \( X \) with \( P \)

\[
X’ = PX = \begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1n} \\
x_{21} & x_{22} & \cdots & x_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
x_{m1} & x_{m2} & \cdots & x_{mn}
\end{pmatrix}
\]

In the third step, Alice adds \( r_1, r_2, \ldots, r_m \) to \( X’ \) and gets \( X'' \) as follows

\[
X'' = \begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1n} & r_1 \\
x_{21} & x_{22} & \cdots & x_{2n} & r_2 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
x_{m1} & x_{m2} & \cdots & x_{mn} & r_m
\end{pmatrix}
\]

Alice uses \( g_i, 1 \leq i \leq m \) (it is possible since any practical network coding system would make \( m \ll n \)) as the “intended” hash values of \( x_1, x_2, \ldots, x_m \). A list of padding \( x_{10}, x_{20}, \ldots, x_{m0} \) can be computed. We add the padding to every packet and get \( \hat{X} \) as shown in Fig.1.

When a message packet \( x_i'' \) is padded with \( x_{i0} \) to form the new packet \( \hat{x}_i \), then \( \mathcal{H}(\hat{x}_i) = g_i, 1 \leq i \leq m \).

**Bob’s decoder:** Bob first decodes \( \hat{X} \) and gets \( r_1, r_2, \ldots, r_m \). The Vandermonde matrix \( P \) can be computed from \( r_1, r_2, \ldots, r_m \). Bob then per-multiplies the associated matrix \( PX \) with \( P^{-1} \) and gets the original packet \( X \).

C. The Basic Verification Scheme

As shown in Fig.1, it is noted that the message can only be padded using the secret key that is known only by Alice. Next, Alice chooses a seed \( c \) and feed it to a pseudo-random generator \( G \). Instead of choosing the coefficients, the source uses the random numbers \( c_1, c_2, \ldots, c_m \) generated by \( G \) as the “intended” coefficients. Since the coefficients can be computed from the public function \( G \), there would be no need to distribute the coefficients, and it suffices if all the nodes know \( c \).

Our proposed scheme consists of two algorithms, namely the encoding algorithm and the verification algorithm.

**Encoding Algorithm:** The encoder performs the following steps

1) Choose a random seed \( c \).
2) Generate pseudo-random numbers \( c_1, c_2, \ldots, c_m \) from \( G \) with \( c \).

3) For each \( 1 \leq i \leq m \), choose \( g_i = g^{u_i} \pmod{q} \) as the “intended” hash values.

4) For each \( 1 \leq i \leq m \), compute \( x_{i0} = \{ u_i - \sum_{j=1}^{n} x_{ij} u_j \} u_0^{-1} \pmod{p} \).

5) Let \( \hat{X} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)^T \), where \( \hat{x}_i = (x_{i0}, x_{i1}, \ldots, x_{in}, r_i) \) for all \( 1 \leq i \leq m \).

6) Output \( x, c \) and the public parameter \( g_0, g_1, \ldots, g_{n+1}, p, q, c \). Where \( x \) is the linear combination \( x = \sum_{i=1}^{m} c_i \hat{x}_i \).

**Verification Algorithm**: During verification, each node is given a packet \( x \) and public information \( t \). In the case where this packet is not tampered with, \( x \) is the linear combination \( x = \sum_{i=1}^{m} c_i \hat{x}_i \), and \( t \) represents public parameters \( g_0, g_1, \ldots, g_{n+1}, p, q, c \).

Each node can verify the integrity of the packet as follows:

1) From \( c \), compute \( c_1, c_2, \ldots, c_m \).

2) Compute the hash value \( \mathcal{H}_1 = \mathcal{H}(x) \).

3) Compute the hash value \( \mathcal{H}_2 = \prod_{i=1}^{m} h_i^{c_i} \pmod{p} \), where \( h_i = g_i \) for \( 1 \leq i \leq m \).

4) Verify that \( \mathcal{H}_1 = \mathcal{H}_2 \).

In our scheme, every node selects and distributes random values to all its following nodes instead of the transmission of the coefficients. The coefficients are generated from a shared pseudo-random number generator in each node and the global encoding kernel can be calculated recursively in any upstream to downstream order.

In practice, the need for distributing random values can be further eliminated by using a public random function. For example, it can be the SHA-1 hash of the original file identifier, creation date, publisher, and other data that are public and should be known to all the receivers before the download session begins.

Our verification scheme enables the nodes to check the integrity of packets without the requirement for a secure channel. Also, the computation involved in the hash values generation and verification processes is very simple.

**IV. Security of our algorithm**

A. Security against the contamination adversaries

It can be shown that the basic verification scheme is indeed secure if \( DL[g,p,q] \) is hard, using an argument similar to that in [10] (proof of Theorem 3).

**Theorem 1**: It is computationally infeasible to find \( \hat{X} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)^T \), and \( c = (c_1, c_2, \ldots, c_m) \) such that for \( x = \sum_{i=1}^{m} c_i \hat{x}_i \), we have \( y \neq x \) and \( \mathcal{H}(x) = \mathcal{H}(y) \), namely the basic scheme is secure if \( DL[g,p,q] \) is hard.

**Proof**: We prove this theorem by showing that if there is a polynomial time algorithm \( A \) that finds \( \hat{X} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)^T \) and \( c = (c_1, c_2, \ldots, c_m) \) such that for \( x = \sum_{i=1}^{m} c_i \hat{x}_i \) we have \( y \neq x \) and \( \mathcal{H}(x) = \mathcal{H}(y) \) with probability \( p \) that is not negligible, we can use it to construct a polynomial time algorithm \( B \) that finds a collision \( x \) and \( y \) in \( \mathcal{H} \) with the same probability \( p \) which is not negligible.

However, if \( DL[g, p, q] \) is hard, Lemma 2 show that the hash function \( \mathcal{H} \) is collision free, and thus \( p \) should be negligible, which is a contradiction. Therefore, the basic schemes are secure if the discrete logarithm \( DL[g, p, q] \) is hard.

B. Security against the eavesdropping adversaries

**Theorem 2**: Given a network that the number of independent message available to Calvin is less than the multicast capacity i.e. \( k = \max_k k_i < m \). The algorithm in section-B achieves the practically secure condition at a probability of one when random code is used.

**Proof**: In our algorithm, Alice transmits \( \hat{X} \) instead of \( X \), so the message available to Calvin is \( A_i \hat{X} \). As long as Calvin doesn’t get \( r_1, r_2, \ldots, r_m \) from \( A_i \hat{X} \), he can’t get the global encoding kernel about \( X \), and still can’t get any meaningful information about \( X \) without the global encoding kernel about \( X \). So by taking linear combinations of the observed packets \( A_i \hat{X} \), Calvin shouldn’t be able to recover \( r_1, r_2, \ldots, r_m \) which implies

\[
b_i A_i \hat{X} \neq I \hat{X} (y b_i, i)
\]

where \( b_i \) is a \( k_i \times m \) matrix in \( \mathbb{F}_p \) and \( I \) is a \( m \times m \) identity matrix.

Since the number of independent messages available to the eavesdropper is less than the multicast capacity of the network, the condition (9) can always be satisfied. Moreover, Calvin can’t get any packet of \( X \) by only getting the value \( r_1, r_2, \ldots, r_m \) which implies

\[
b_i A_i P \neq I_{m,n} (y b_i, n, i)
\]

Multiplying both sides by \( P^{-1} \), we have

\[
b_i A_i \neq I_{m,n} P^{-1} (y b_i, n, i)
\]

Where \( I_{m,n} \) is the \( n^{th} \) row of an \( m \times m \) identity matrix and \( b_i \) is a \( k_i \times m \) matrix in \( \mathbb{F}_p \). The above condition is satisfied if each row of \( P^{-1} \) is not in the row space of each \( A_i \).
Theorem 3: In a network that supports a multicast capacity of \(m\), if at most \(k\) \((k < m)\) edges can be tapped simultaneously then the multicast capacity under practical security requirements is \(m - \frac{2m}{n}\) (Here, the asymptotically negligible term \(\frac{2m}{n}\) corresponds to the overhead due to the redundancy Alice appends to \(X\)).

Proof: The network supports a multicast capacity of \(m\) so a linear code can be found to multicast \(m\) packets \([3]\). From theorem 2 a transformation at the source can be applied to make it practical security.

V. DISCUSSION

In this section, we examine the overhead and the start-up latency induced by our scheme. For fair comparisons, we choose \(n=410\), \(|p| = 320\) and \(|q| = 1024\) in the follow discussion. The size of every packet is 16KB. What’s more, we assume that the original file is divided into \(m=10\) packets.

The comparisons are shown in Fig.3.

A. Communication Overhead

The communication overhead is caused by two parts of parameters. The first part refers to the amount of data we need to distribute to each node for the security of our scheme. The second part is the code coefficients. The actual communication overhead largely depends on the parameters chosen for the actual implementations.

In the scheme proposed by Krohn [9], the parameters chosen for the homomorphic hash function would generate a hash value of size 1024 bits per packet. The total size of coefficients is 3200 bits per packet. Hence, the total size of the “first-order”hash values and the coefficients would be 3.22% of the original data. For a file of size 1 GB, their method would require hash values of size 8 MB. To distribute these hash values, the authors in Krohn [9] proposed to recursively apply the same scheme on the 8 MB hash values, which would generate more “second”or higher order of hash values. The size of the high order hash values constitutes 0.01% of the size of the original data. Hence the total overhead is 3.23%.

In the scheme proposed by Zhao [11], if the file is divided into 10 packets, each packet is a vector in \(\mathbb{F}_q\). The size of each packet is also about 16KB. The size of each augmented vector (with coding vectors in the front) is about 16.4KB, and thus, the overhead of each packet is 2.43%. On the other hand, after the initial setup, the scheme of [11] has to publish 3200 bits for the new signature vector for the security of their scheme. Thus the total overhead of their scheme is 4.86%. In conclusion, although they proposed a simple signature scheme, the communication overhead of their scheme is very high.

The scheme in [10] required padding of three values and they should also distribute the coefficients. The overhead caused by the coefficients themselves is 2.44%. Therefore the communication overhead of their scheme is higher than us, although they use the technical of batching verification. In our scheme, the coefficients are generated from a pseudo-random number generator in each node so the distribution of them is avoided.

The communication overhead of our scheme is only caused by padding we add in every packet. Each packet distributed only incurs 0.48% overhead which is negligible compared with previous works. Formally, Let the file size be \(S\) and each one of which is a vector in \(\mathbb{F}_p\). The size of each vector is \(B = n\log(p)\) and we have \(S = mn\log(p)\). The size of each augmented vector (with the padding in the front and the back) is \(B_a = (n+2)\log(p)\), and thus, the overhead of the packet is \(\frac{2}{n}\) times the file size. Note that the communication overhead of our scheme is asymptotically negligible.

B. Start-Up Latency

At the beginning of a content distribution session, the source and all the nodes participating in the distribution have to agree on the set of parameters used for the coding and verification. The public parameters in our scheme are \(p,q,g_0,g_1,\ldots,g_{n+1}\) and the total size of the public parameters is approximately 16.3 KB. With these parameters it would be sufficient for any node to perform verification. Assuming that the bandwidth between a node and the source (or any other node from which these parameters are distributed) is 1 Mbps, it would take less than 0.127 seconds before the node is ready to perform verification. The start-up latency in our scheme is fixed once the parameters for the hash function and the block size are chosen, and is independent of the size of the content to be distributed. The start-up latency of [10] is 0.127 seconds, more or less the same with us. For the scheme in [9], the size of all the public parameters is the same as the size of the data in a packet, which is 16 KB. It takes 0.125 seconds to be transmitted on the same link. However, when the node needs to receive 8 MB hash values of a 1 GB file as in the example given in [9], it would require 64 seconds, with the same 1 Mbps link. The start-up latency is proportional with the size of the file. The public parameters of [11] consist of two parts: the public parameters and the signature vectors. The size of their public parameters is 32.8KB and it takes 0.25 seconds to be transmitted on the same link. Further more they have to publish new signature vectors for the security of their scheme in every setup.

VI. CONCLUSION

In this paper, we investigate the security issues that arise from using network coding and propose a secure algorithm. By means of our algorithm every node can verify the integrity of the received packets easily and an eavesdropper is unable to get any meaningful information about the source. We show
that when we give up a small amount of overall capacity, the practically secure condition can be achieved at a probability of 1, which is much higher than that of [1]. We also propose a new paradigm where the public parameters are selected as the “intended” hash values and the code coefficients are generated in a pseudo-random number generator in every node. In this way the distribution of the hash values and the coefficients are avoided. We have shown that the communication overhead of our algorithm is $\frac{2}{n}$, which is negligible compared with previous works and the start-up latency is transitory.

ACKNOWLEDGMENT

The work of Yejun Zhou, Hui Li and Jianfeng Ma was supported by National Natural Science Foundation of China grant 60772136, National Natural Science Foundation of China grant 60633020, 863 Hi-Tech Research and Development Program of China grant 2007AA01Z435, and 863 Hi-Tech Research and Development Program of China grant 2007AA01Z429.

REFERENCES

[1] K. Bhattach and K.R. Narayanan, “Weakly secure network coding,” In Proc. of the First Workshop on Network Coding, Theory, and Applications (NetCod), Riva del Garda, Italy, 2005.
[2] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, “Network information flow,” IEEE Trans. Inf. Theory, vol. 46(4), pp. 1204-1216, 2000.
[3] S.-Y. R. Li, R. W. Yeung, and N. Cai, “Linear network coding,” IEEE Trans. on Information Theory, vol. IT-49, pp. 371-381, 2003.
[4] T. Ho, R. Koetter, M. Médard, D. R. Karger, and M. Effros, “The benefits of coding over routing in a randomized setting,” in International Symposium on Information Theory (ISIT), 2003.
[5] T. Ho, M. Médard, J. Shi, M. Effros and D. R. Karger, “On randomized network coding,” In proc. 41st Annual Allerton Conference on Communication Control and Computing, Oct. 2003.
[6] T. C. Ho, B. Leong, R. Koetter, M. Médard, M. Effros, and D. R. Karger, “Byzantine modification detection in multicast networks using randomized network coding,” in International Symposium on Information Theory (ISIT), Chicago, USA, June 2004.
[7] C. Gkantsidis, J. Miller, P. Rodriguez, “Comprehensive View of a Live Network Coding P2P System,” Proceedings of the 6th ACM SIGCOMM conference on Internet measurement Oct. 2006
[8] S. Jaggi, M. Langberg, S. Katti, T. Ho, D. Katabi, and M. Médard, “Resilient network coding in the presence of Byzantine adversaries,” accepted to IEEE INFOCOM ‘07, Anchorage, Alaska, May 2007.
[9] M. N. Krohn, M. J. Freedman, and D. Mazieres, “On-the-fly verification of rateless era-secure codes for efficient content distribution,” IEEE Symp. Security and Privacy, Oak-labd, CA, pp. 226-240, May 2004.
[10] Qiming Li, Dah-Ming Chiu, John C.S. Lui, “On the Practical and Security Issues of Batch Content Distribution Via Network Coding,” 14th IEEE International Conference (ICNP ‘06) pp. 158 - 167, Nov 2006.
[11] Fang Zhao, Ton Kalker, M. Médard, and Keesook J. Han, “Signatures for content distribution with network coding,” ISIT2007, Nice, France, June 24 - June 29, 2007
[12] N. Cai and R. W. Yeung, “Secure network coding,” International Symposium on Information Theory (ISIT) Lausanne, Switzerland, June 30 - July 5, 2002.
[13] J. Feldman, T. Malkin, C. Stein, R. A. Servedio “On the capacity of secure network coding,” In Proc. 42nd Annual Allerton Conference on Communication, Control, and Computing, Sep. 2004.
[14] K. Jain, “Security based on network topology against the wiretapping attack,” IEEE Wireless Communications, pp. 68-71, Feb. 2004.
[15] L. Lima, M. Médard, J Barros, “Random linear network coding: A free cipher?” In Proc. of IEEE International Symposium on Information Theory (ISIT), Nice, France, June 24-29 2007.
[16] Ngai. Chi Kin, Yang. Shenghao, “Deterministic Secure Error-Correcting (SEC) Network Codes,” IEEE Information Theory Workshop (ITW), pp. 96 - 101, Sept2-6 2007.

[17] M. Bellare, O. Goldreich, and S. Goldwasser, “Incremental cryptography: The case of hashing and signing,” CRYPTO, 1994.
Secure Network Coding Against the Contamination and Eavesdropping Adversaries

Yejun Zhou, Hui Li and Jianfeng Ma
Key lab of CNIS Ministry of Education
Xidian University
Xi’an, Shanxi 710071, China
Email: {yjzhou, lihui, jfma}@mail.xidian.edu.cn

Abstract—In this paper, we propose an algorithm that targets contamination and eavesdropping adversaries. We consider the case when the number of independent packets available to the eavesdropper is less than the multicast capacity of the network. By means of our algorithm every node can verify the integrity of the received packets easily and an eavesdropper is unable to get any “meaningful information” about the source. We call it “practical security” if an eavesdropper is unable to get any meaningful information about the source. We show that, by giving up a small amount of overall capacity, our algorithm achieves the practically secure condition at a probability of one, which is much higher than that of Bhattacharya and Narayan’s [1]. Furthermore, the communication overhead of our algorithm is negligible compared with previous works, since the transmission of the hash values and the code coefficients are both avoided.

I. INTRODUCTION

The concept of network coding was first introduced by Ahlswede et al. [2]. They showed that multicast rates could be increased by allowing for network coding instead of just routing. Shortly afterwards, Li, Yeung and Cai [3] showed that it is sufficient for the encoding functions at the interior nodes to be linear. Ho et al. [4] and [5] proposed a random coding scheme in which the message on outgoing edges of a node are chosen to be a random linear combination of the message on its incoming edges.

In reality, network transmission may suffer from two kinds of adversaries: contamination and eavesdropping. Network coding has been studied to con-quer these two kinds of adversaries. Ho et al. [6] considered the problem of network coding in the presence of Byzantine attacker. Gkantsidis et al. [7] also considered the related problem. Jaggi et al. [8] designed a resilient network coding algorithm which is information-theoretically secure and rate-optimal for different adversarial strengths. Homomorphic hashing function was first proposed in [9], which allows nodes to check blocks on-the-fly in a system where content is encoded at the source using rateless codes. However, the total size of the hash values of their scheme is proportional to the number of blocks, which could be very large and the cryptographic hash function is computationally expensive. Li et al. [10] employed a batch content distribution verification scheme, which reduced the computational cost of each node to cache and scan all the received packets when computing a new packet. The cryptographic hash function of their scheme is computationally inexpensive compared with which in [9]. Unfortunately, their scheme deviate from the classical network coding scheme, which is bandwidth consumed and delay could be induced at the sinks. On the other hand, although batching can decrease the computation time, batching block verification has the risk of letting some malicious packets propagate since packets are exchanged without being checked. Thus, standard batching techniques do not work well with network coding. Zhao et al. [11] presented a signature scheme with low computation, but their scheme required long start-up latency. Finally, all the works presented above have to distribute the coefficients which is bandwidth consumed.

Cai and Yeung [12] considered the problem of using network coding to achieve perfect information security against an eavesdropper who can eavesdrop on a limited number of network links, and presented the construction of a secure linear network code for this purpose. A similar problem was considered in [13] featuring a random coding approach in which only the input vector is modified.

Bhattacharya and Narayan [14] first defined a model for security that is more suitable for practical applications. In this paper, we also consider this type of model, which is not information theoretically secure, but is secure enough for the application. An interesting observation made in [14] was that for a computation limited eavesdropper with the use of one way function it is possible to transmit at a high rate without the eavesdropper getting any meaningful information about the source. A more general threat posed by intermediate nodes was considered in [15].

In this paper, we consider these two kinds of adversaries at the same time, that is, the adversary can contaminate the transmission on a subset of channels, and at the same time eavesdrop on another subset of channels with cardinality less than or equal to m. Ngai and Yang [16] studied the similar problem and constructed a secure error-correcting network codes.

The main contribution of this paper is to propose an algorithm, which can not only verify the integrity of the received packets easily but also achieve the practically secure condition at a probability of one. In our scheme, we use the public parameters as the “intended”hash values. The original packets are padded so that they are hashed to the public parameters. In this way the transmission of the hash values
is avoided. The code coefficients in our scheme are generated in a pseudo-random number generator in each node, so the distribution of the coefficients is also avoided. We show that the communication overhead and the start-up latency are negligible since the transmission of the hash values and the coefficients are both avoided.

This paper is organized as follows. In the next section we give the notations used in this paper. The secure network coding scheme is proposed in section III. In Section IV we present the security of our algorithm. Overhead and start-up latency of our algorithm are discussed in Section V. Finally, this paper is concluded in section VI.

II. NETWORK MODEL AND NOTIONS

In this paper, we assume that all the messages and coefficients are generated in \( \mathbb{F}_p \), where \( p \) is a large enough prime number. We shall use small letters \( x, y \) etc. to denote vectors whose dimensions will be clear from the context. The matrices are denoted by the capital letters such as \( X, X' \) etc. The transpose operator of vectors and matrices will be denoted by “\(^T\)” thus \( x^T \) will stand for column vectors.

A. Network Model

We represent a network by a directed graph \( G = (V; E) \), where \( V \) is the set of vertices (nodes) and \( E \) is the set of edges (channels). We assume an order on \( V \) which is consistent with the associated partial order on \( G \). A network code is said to be linear if the message on any outgoing edge of any node is a linear combination of the messages on the incoming edges of the node.

In this paper, we assume that the source node sends information \( X \) of the following form:

\[
X = \begin{bmatrix}
  x_{11} & x_{12} & \cdots & x_{1n} \\
  x_{21} & x_{22} & \cdots & x_{2n} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{m1} & x_{m2} & \cdots & x_{mn}
\end{bmatrix} \begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_m
\end{bmatrix} = \begin{bmatrix}
  x_{1} \\
  x_{2} \\
  \vdots \\
  x_{m}
\end{bmatrix}
\]

We call \( x_i, i = 1, \ldots, m \) a packet.

Therefore, for a linear code the message on edge \( e_j \in E \) can be written as \( F_{e_j}X \) where \( F_{e_j} \) is a length \( m \) vector over \( \mathbb{F}_p \) (we call it global encoding kernel in this paper) on edge \( e_j \in E \).

B. Threat Model

There is a source, Alice, and a destination, Bob, who communicate over a wired or wireless network. There is also an eavesdropper Calvin, hidden somewhere in the network. He aims to eavesdrop on the transfer of information from Alice to Bob and injects his own. A malicious node can generate corrupted packets and then distribute them to other nodes, which in turn use them to (unintentionally) create new encoded packets that are also corrupted. A wiretap network is specified by a collection \( A \) of sets of edges \( A = \{ A_1, A_2, \ldots, A|A| \} \), \( A_i \in E \). Calvin selects a particular set \( A_i \in A \) and listens to all messages transmitted on edges in \( A_i \) to get some information. We assume that the set doesn’t change with time. When we are specified a linear code and a wiretap network we use \( A_i \) to represent a matrix whose rows contain all linearly independent global encoding kernel corresponding to edge \( e_j \in A_i \). In this case, the messages available to Calvin is \( A_iX \). The number of rows in \( A_i \) is represented by \( k_i \). We define \( k \) as \( \max_i k_i \).

C. Notions

1) \( I(X; M) = 0 \) : The network capacity is the time-average of the maximum number of packets that can be delivered from Alice to Bob, assuming no adversarial interference, i.e., the max-flow. To simplify notion, in this paper, we assume the max-flow from Alice to Bob is \( m \).

2) Practical security: Consider a set of messages \( M \). Let \( U \) be subset of the set containing the multicast information \( X \). We say that \( M \) has no information about \( U \) if \( I(U; M) = 0 \). We say that \( M \) has no meaningful information about \( U \) if \( I(x_i; M) = 0, \forall x_i \in U \). In this paper we concentrate on two special cases and generalize the results towards the end. We say that Calvin has no information about the source if \( I(X; M) = 0 \) where \( M \) is the set of messages that Calvin chooses to observe. The security condition considered by Cai and Yeung [12] falls in this category. We will use Shannon security to refer to this security requirement. The second case we consider is when Calvin gets no meaningful information about the source i.e. \( I(x_i; M) = 0, \forall x_i \) for messages \( M \) observed by Calvin. We call this type of security as practical security.

It is noted that if Alice transmits a linear transformation of \( X \), \( PX \), instead of \( X \) then the message transmitted on edge \( e_j \) would be \( F_{e_j}PX \) (\( P \) is a \( m \times m \) matrix which is unknown to Calvin). In this case, although Calvin has some information about the source he is unable to get any meaningful information.

As shown in Fig.1, let us assume that Calvin can listen to any one edge of this network. The multicast capacity for this network is \( 2, x_1 \), and \( x_2 \) are the messages of Alice. In Fig.1 (a), \( w \) is a uniform random sequence independent of the messages. This is an example of the coding scheme constructed by Cai.

Fig. 1. Networks
and Yeung [12]. Obviously, the maximum multicast capacity supported is 1 when this system has to be Shannon secure. When the security condition is relaxed to practical security, as shown in Fig.1 (b), the max-flow can be achieved.

III. SECURE NETWORK CODING

A. The Homomorphic Hash Function

We first choose the hash parameters $q, g$. Let $o(x)$ denote the order of $x$ in the field $\mathbb{F}_q$. Here we choose $o(g) = p$ in $\mathbb{F}_q$ (it is a subfield of $\mathbb{F}_q$). Furthermore we randomly select $n+2$ numbers $u_0,u_1,\cdots,u_n,u_{n+1}$ from $\mathbb{F}_p$. Next, we compute $g_i = g^{u_i}$ (mod $q$) for all $0 \leq i \leq n+1$. The public parameter of the hash function is $p,q,g,0,\cdots,g_{n+1}$. Whereas $u_0,u_1,\cdots,u_{n+1}$ and $g$ should be kept secret.

Formally, we define $DL[g, p, q]$ to be the computational problem: Given $g, p$ and $q$, where $o(q) = p$ in $\mathbb{F}_q$, find $x$ such that $y = q^x$ (mod $q$). Hence, we have

Lemma 1: Given $g_0,g_1,\cdots,g_{n+1}$, and the public parameters $p,q$, it is computationally infeasible for a node to find $u_0,u_1,\cdots,u_{n+1}$, such that $g_i = g^{u_i}$ (mod $q$) if $DL[g, p, q]$ is hard.

Assume that each message is of the form: $x = (x_0,x_1,\cdots,x_n, r)$ where $x_i, r \in \mathbb{F}_p$ for $0 \leq i \leq n$, and the hash of $x$ is computed as

$$\mathcal{H}(x) = \prod_{i=0}^{n} g_i^{x_i}g_{i+1}^{-r} \quad (\text{mod } q) \quad (2)$$

Based on this construction, we have

$$\mathcal{H}(xy) = g^{(\sum_{i=0}^{n} u_i x_i + u_{n+1} r)} \quad (\text{mod } q) \quad (3)$$

For any two messages $x = (x_0,x_1,\cdots,x_n, r_1)$ and $y = (y_0,y_1,\cdots,y_n, r_2)$, we define the addition of $x$ and $y$ as

$$x + y = (x_0 + y_0, x_1 + y_1, \cdots, x_n + y_n, r_1 + r_2) \quad (\text{mod } p)$$

where $r = (r_1 + r_2)$ (mod $p$) and $z_i = (x_i + y_i)$ (mod $p$) for $0 \leq i \leq n$

Hence, this hash function has the following homomorphic property

$$\mathcal{H}(x)\mathcal{H}(y) = \mathcal{H}(x+y) \quad (5)$$

The security of $\mathcal{H}$ is defined in terms of the difficulty in finding collisions. It can be shown that the hash function is indeed collision free if $DL[g, p, q]$ is hard. In particular we have:

Lemma 2: The hash function $\mathcal{H}$ is collision-free (namely it is computationally infeasible to find two different messages $x_1$ and $x_2$ such that $\mathcal{H}(x_1) = \mathcal{H}(x_2)$ if $DL[g, p, q]$ is hard.

It can be proved that the hash function is indeed collision-free, using an argument in [17] (proof of Theorem 3.4).

B. Alice’s Encoder and Bob’s Decoder

Alice’s encoder: Alice encodes $X$ in the following steps. She first chooses $m$ parity symbols $r_d$, for $d \in \{1,\cdots,m\}$ uniformly at random from the field $\mathbb{F}_p$, and then generates a Vandermonde matrix $P$ as follows

$$P = \begin{pmatrix} r_1^1 & r_2^1 & \cdots & r_m^1 \\ r_1^2 & r_2^2 & \cdots & r_m^2 \\ \vdots & \vdots & \ddots & \vdots \\ r_1^n & r_2^n & \cdots & r_m^n \end{pmatrix} \quad (6)$$

In the second step, Alice per-multiplies the source message $X$ with $P$

$$X’ = PX = \begin{pmatrix} x_{11}’ & x_{12}’ & \cdots & x_{1n}’ \\ x_{21}’ & x_{22}’ & \cdots & x_{2n}’ \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}’ & x_{m2}’ & \cdots & x_{mn}’ \end{pmatrix} = \begin{pmatrix} x_1’ \\ x_2’ \\ \vdots \\ x_m’ \end{pmatrix} \quad (7)$$

In the third step, Alice adds $x_1, x_2, \cdots, x_m$ to $X’$ and gets $X”$ as follows

$$X” = \begin{pmatrix} x_{11}'' & x_{12}'' & \cdots & x_{1n}'' & r_1'' \\ x_{21}'' & x_{22}'' & \cdots & x_{2n}'' & r_2'' \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1}'' & x_{m2}'' & \cdots & x_{mn}'' & r_m'' \end{pmatrix} = \begin{pmatrix} x_1'' \\ x_2'' \\ \vdots \\ x_m'' \end{pmatrix} \quad (8)$$

Alice uses $g_i, 1 \leq i \leq m$ (it is possible since any practical network coding would make $m \ll n$) as the “intended”hash values of $x_1, x_2, \cdots, x_m$. A list of padding $x_{10}, x_{20}, \cdots, x_{m0}$ can be computed. We add the padding to every packet and get $\hat{X}$ as show in Fig 1.

When a message packet $x_i”$ is padded with $x_{i0}$ to form the new packet $\hat{x}_i$, then $\mathcal{H}(\hat{x}_i) = g_i, 1 \leq i \leq m$.

Bob’s decoder: Bob first decodes $\hat{X}$ and gets $r_1, r_2, \cdots, r_m$. The Vandermonde matrix $P$ can be computed from $r_1, r_2, \cdots, r_m$. Bob then per-multiplies the associated matrix $PX$ with $P^{-1}$ and gets the original packet $X$.

C. The Basic Verification Scheme

As show in Fig.1, It is noted that the message can only be padded using the secret key that is known only by Alice. Next, Alice chooses a seed $c$ and feed it to a pseudo-random generator $G$. Instead of choosing the coefficients, the source uses the random numbers $c_1, c_2, \cdots, c_m$ generated by $G$ as the “intended”coefficients. Since the coefficients can be computed from the public function $G$, there would be no need to distribute the coefficients, and it suffices if all the nodes know $c$.

Our proposed scheme consists of two algorithms, namely the encoding algorithm and the verification algorithm.

Encoding Algorithm: The encoder performs the following steps

1) Choose a random seed $c$. 

...
2) Generate pseudo-random numbers \(c_1, c_2, \cdots, c_m\) from \(G\) with \(c\).

3) For each \(1 \leq i \leq m\), choose \(g_i = g^{u_i} \pmod q\) as the “intended” hash values.

4) For each \(1 \leq i \leq m\), compute \(x_{i0} = \{u_i - \sum_{j=1}^{n} x_{ij} u_j - u_{n+1} r_j\} u_0^{-1} \pmod p\).

5) Let \(\hat{X} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_m)^T\), where \(\hat{x}_i = (x_{i0}, x_{i1}, \cdots, x_{in}, r_i)\) for all \(1 \leq i \leq m\).

6) Output \(x, c\) and the public parameter \(g_0, g_1, \cdots, g_{n+1}, p, q\). Where \(x\) is the linear combination \(x = \sum_{i=1}^{m} c_i \hat{x}_i\).

Verification Algorithm: During verification, each node is given a packet \(x\) and public information \(t\). In the case where this packet is not tampered with, \(x\) is the linear combination \(x = \sum_{i=1}^{m} c_i \hat{x}_i\), and \(t\) represents public parameters \(g_0, g_1, \cdots, g_{n+1}, p, q\) and \(c\).

Each node can verify the integrity of the packet as follows:

1) From \(c\), compute \(c_1, c_2, \cdots, c_m\).

2) Compute the hash value \(H_1 = H(x)\).

3) Compute the hash value \(H_2 = \prod_{i=1}^{m} h_i^{c_i} \pmod q\), \((h_i = g_i, \text{ for } 1 \leq i \leq m)\).

4) Verify that \(H_1 = H_2\).

In our scheme, every node selects and distributes random values to all its following nodes instead of the transmission of the coefficients. The coefficients are generated from a shared pseudo-random number generator in each node and the global encoding kernel can be calculated recursively in any upstream to downstream order.

In practice, the need for distributing random values can be further eliminated by using a public random function. For example, it can be the SHA-1 hash of the original file identifier, creation date, publisher, and other data that are public and should be known to all the receivers before the download session begins.

The verification scheme enables the nodes to check the integrity of packets without the requirement for a secure channel. Also, the computation involved in the hash values generation and verification processes is very simple.

IV. SECURITY OF OUR ALGORITHM

A. Security against the contamination adversaries

It can be shown that the basic verification scheme is indeed secure if \(DL[g, p, q]\) is hard, using an argument similar to that in [10] (proof of Theorem 3).

Theorem 1: It is computationally infeasible to find \(\hat{X} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_m)^T\), \(y\) and \(c = (c_1, c_2, \cdots, c_m)\) such that for \(x = \sum_{i=1}^{m} c_i \hat{x}_i\), we have \(y \neq x\) and \(H(x) = H(y)\), namely the basic scheme is secure if \(DL[g, p, q]\) is hard.

Proof: We prove this theorem by showing that if there is a polynomial time algorithm \(A\) that finds \(\hat{X} = (\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_m)^T\), \(y\) and \(c = (c_1, c_2, \cdots, c_m)\) such that for \(x = \sum_{i=1}^{m} c_i \hat{x}_i\), we have \(y \neq x\) and \(H(x) = H(y)\) with probability \(p\) that is not negligible, we can use it to construct a polynomial time algorithm \(B\) that find a collision \(x\) and \(y\) in \(H\) with the same probability \(p\) which is not negligible.

However, if \(DL[g, p, q]\) is hard, Lemma 2 shows that the hash function \(H\) is collision free, and thus \(p\) should be negligible, which is a contradiction. Therefore, the basic schemes are secure if the discrete logarithm \(DL[g, p, q]\) is hard.

B. Security against the eavesdropping adversaries

Theorem 2: Given a network that the number of independent message available to Calvin is less than the multicast capacity \(i.e. k = \max_i k_i < m\). The algorithm in section-B achieves the practically secure condition at a probability of one when random code is used.

Proof: In our algorithm, Alice transmits \(\hat{X}\) instead of \(X\), so the message available to Calvin is \(A, \hat{X}\). As long as Calvin doesn’t get \(r_1, r_2, \cdots, r_m\) from \(A, X\), he can’t get the global encoding kernel about \(X\), and still can’t get any meaningful information about \(X\) without the global encoding kernel about \(X\). So by taking linear combinations of the observed packets \(A, \hat{X}\) Calvin shouldn’t be able to recover \(r_1, r_2, \cdots, r_m\) which implies

\[
b_i A_i \hat{X} \neq I \hat{X} (\forall b_i, i)\]

(9)

where \(b_i\) is a \(k_i \times m\) matrix in \(F_{p^m}\) and \(I\) is an \(m \times m\) identity matrix.

Since the number of independent messages available to the eavesdropper is less than the multicast capacity of the network, the condition (9) can always be satisfied.

Moreover, Calvin can’t get any packet of \(X\) by only getting the value \(r_1, r_2, \cdots, r_m\) which implies

\[
b_i A_i P \neq I_{m,n}(\forall b_i, n, i)\]

(10)

Multiplying both sides by \(P^{-1}\), we have

\[
b_i A_i \neq I_{m,n} P^{-1}(\forall b_i, n, i)\]

(11)

Where \(I_{m,n}\) is the \(n^{th}\) row of an \(m \times m\) identity matrix and \(b_i\) is a \(k_i \times m\) matrix in \(F_{p^m}\). The above condition is satisfied if each row of \(P^{-1}\) is not in the row space of each \(A_i\).
Theorem 3: In a network that supports a multicast capacity of \( m \), if at most \( k \) \((k < m)\) edges can be tapped simultaneously then the multicast capacity under practical security requirements is \( m - \frac{2m}{n} \) (Here, the asymptotically negligible term \( \frac{2m}{n} \) corresponds to the overhead due to the redundancy Alice appends to \( X \)).

Proof: The network supports a multicast capacity of \( m \) so a linear code can be found to multicast \( m \) packets [3]. From theorem 2 a transformation at the source can be applied to make it practical security.

V. DISCUSSION

In this section, we examine the overhead and the start-up latency induced by our scheme. For fair comparisons, we choose \( n=410, |p| = 320 \) and \( |q| = 1024 \) in the following discussion. The size of every packet is 16KB. What’s more, we assume that the original file is divided into \( n=10 \) packets.

The comparisons are shown in Fig.3

| index | Communication Overhead | Start-Up Latency |
|-------|------------------------|-----------------|
| ours  | 0.48%                  | 0.127s          |
| [9]   | 3.23%                  | 0.125s          |
| [10]  | >2.44%                 | 0.127s          |
| [11]  | 4.86%                  | 0.25s           |

Fig. 3. Comparisons of Communication Overhead and Start-Up Latency

A. Communication Overhead

The communication overhead is caused by two parts of parameters. The first part refers to the amount of data we need to distribute to each node for the security of our algorithm. The second part is the code coefficients. The actual communication overhead largely depends on the parameters chosen for the actual implementations.

In the scheme proposed by Krohn [9], the parameters chosen for the homomorphic hash function would generate a hash value of size 1024 bits per packet. The total size of coefficients is 3200 bits per packet. Hence, the total size of the “first-order” hash values and the coefficients would be 3.22% of the original data. For a file of size 1 GB, their method would require hash values of size 8 MB. To distribute these hash values, the authors in Krohn [9] proposed to recursively apply the same scheme on the 8 MB hash values, which would generate more “second” or higher order of hash values. The size of the high order hash values constitutes 0.01% of the size of the original data. Hence the total overhead is 3.23%.

In the scheme proposed by Zhao [11], if the file is divided into 10 packets, each packet is a vector in \( \mathbb{F}_q \). The size of each packet is also about 16KB. The size of each augmented vector (with coding vectors in the front) is about 16.4KB, and thus, the overhead of each packet is 2.43%. On the other hand, after the initial setup, the scheme of [11] has to publish 3200 bits for the new signature vector for the security of their scheme. Thus the total overhead of their scheme is 4.86%. In conclusion, although they proposed a simple signature scheme, the communication overhead of their scheme is very high.

The scheme in [10] required padding of three values and they should also distribute the coefficients. The overhead caused by the coefficients themselves is 2.44%. Therefore the communication overhead of their scheme is higher than us, although they use the technical of batching verification. In our scheme, the coefficients are generated from a pseudo-random number generator in each node so the distribution of them is avoided. The communication overhead of our scheme is only caused by padding we add in every packet. Each packet distributed only incurs 0.48% overhead which is negligible compared with previous works.

B. Start-Up Latency

At the beginning of a content distribution session, the source and all the nodes participating in the distribution have to agree on the set of parameters used for the coding and verification. The public parameters in our scheme are \( p,q,g_1, \cdots, g_{n+1} \) and the total size of the public parameters is approximately 16.3 KB. With these parameters it would be sufficient for any node to perform verification. Assuming that the bandwidth between a node and the source (or any other node from which these parameters are distributed) is 1 Mbps, it would take less than 0.127 seconds before the node is ready to perform verification. The start-up latency in our scheme is fixed once the parameters for the hash function and the block size are chosen, and is independent of the size of the content to be distributed. The start-up latency of [10] is 0.127 seconds, more or less the same with us. For the scheme in [9], the size of all the public parameters is the same as the size of the data in a packet, which is 16 KB. It takes 0.125 seconds to be transmitted on the same link. However, when the node needs to receive 8 MB hash values of a 1 GB file as in the example given in [9], it would require 64 seconds, with the same 1 Mbps link. The start-up latency is proportional with the size of the file. The public parameters of [11] consist of two parts: the public parameters and the signature vectors. The size of their public parameters is 32.8KB and it takes 0.25 seconds to be transmitted on the same link. Further more they have to publish new signature vectors for the security of their scheme in every setup.

VI. CONCLUSION

In this paper, we investigate the security issues that arise from using network coding and propose a secure algorithm. By means of our algorithm every node can verify the integrity of the received packets easily and an eavesdropper is unable to get any meaningful information about the source. We show that when we give up a small amount of overall capacity, the practically secure condition can be achieved at a probability of 1, which is much higher than that of [1]. We also propose a new paradigm where the public parameters are selected as the “intended” hash values and the code coefficients are generated in a pseudo-random number generator in every node. In this way the distribution of the hash values and the coefficients are...
avoided. We have shown that the communication overhead of our algorithm is $\frac{2}{n}$, which is negligible compared with previous works and the start-up latency is transitory.

ACKNOWLEDGMENT

The work of Yejun Zhou, Hui Li and Jianfeng Ma was supported by National Natural Science Foundation of China grant 60772136, National Natural Science Foundation of China grant 60633020, 863 Hi-Tech Research and Development Program of China grant 2007AA01Z435, and 863 Hi-Tech Research and Development Program of China grant 2007AA01Z429.

REFERENCES

[1] K. Bhattad and K.R. Narayanan, “Weakly secure network coding,” In Proc. of the First Workshop on Network Coding, Theory, and Applications (NetCod), Riva del Garda, Italy, 2005.
[2] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, “Network information flow,” IEEE Trans. Inf. Theory, vol. 46(4), pp. 1204-1216, 2000.
[3] S.- Y. R. Li, R. W. Yeung, and N. Cai, “Linear network coding,” IEEE Trans. on Information Theory, vol. IT-49, pp. 371-381, 2003.
[4] T. Ho, R. Koetter, M. Médard, D. R. Karger, and M. Effros, “The benefits of coding over routing in a randomized setting,” In International Symposium on Informaion Theory (ISIT), 2003.
[5] T. Ho, M. Médard, J. Shi, M. Effros and D. R. Karger, “On randomized network coding,” In Proc. 41st Annual Allerton Conference on Communication Control and Computing, Oct. 2003.
[6] T. C. Ho, B. Leong, R. Koetter, M. Médard, M. Effros, and D. R. Karger, “Byzantine modification detection in multicast networks using randomized network coding,” In International Symposium on Information Theory, Chicago, USA, June 2004.
[7] C. Gkantsidis, J. Miller, P. Rodriguez, “Comprehensive View of a Live Network Coding P2P System,”
[8] S. Jaggi, M. Langberg, S. Katti, T. Ho, D. Katabi, and M. Médard, “Resilient network coding in the presence of Byzantine adversaries,” accepted to IEEE INFOCOM '07, Anchorage, Alaska, May 2007.
[9] M. N. Krohn, M. J. Freedman, and D. Mazières, “On-the-fly verification of rateless era-sure codes for efficient content distribution,” IEEE Symp. Security and Privacy, Oakland, CA, pp. 226-240, May 2004.
[10] Qiming Li, Dah-Ming Chiu, John C.S. Lui, “On the Practical and Security Issues of Batch Content Distribution Via Network Coding,” 14th IEEE International Conference (ICNP '06) pp. 158 - 167, Nov 2006.
[11] Fang Zhao, Ton Kalker, M. Médard, and Keesook J. Han, “Signatures for content distribution with network coding,” ISIT2007, Nice, France, June 24 - June 29, 2007
[12] N. Cai and R. W. Yeung, “Secure network coding,” International Symposium on Information Theory (ISIT) Lausanne, Switzerland, June 30 - July 5, 2002.
[13] J. Feldman, T. Malkin, C. Stein, R. A. Servedio “On the capacity of secure network coding,” In Proc. 42nd Annual Allerton Conference on Communication, Control, and Computing, Sep. 2004.
[14] K. Jain, “Security based on network topology against the wiretapping attack,” IEEE Wireless Communications, pp. 68-71, Feb. 2004.
[15] L. Lima, M. Médard, J Barros, “Random linear network coding: A free cipher?” In Proc. of IEEE International Symposium on Information Theory (ISIT), Nice, France, June 24-29 2007.
[16] Ngai. Chi Kin, Yang. Shenghao, “Deterministic Secure Error-Correcting (SEC) Network Codes,”IEEE Information Theory Workshop (ITW), pp. 96 - 101, Sept2-6 2007.
[17] M. Bellare, O. Goldreich, and S. Goldwasser, “Incremental cryptography: The case of hashing and signing,” CRYPTO, 1994.