A Combination Scheme against Pollution Attacks in Network Coding

Wu Chi, Huang Cheng, Huang Xiaotao*
Network & Computer Center
Huazhong University of Science & Technology
Wuhan, P.R.china
{wuchi, huangxt}@hust.edu.cn*

***Abstract*—The network coding usually suffers from pollution attacks. Two schemes against pollution attacks are discussed in this paper, based on linear space signature and information theory, respectively. Pollution attacks detection algorithm of information theory based scheme is improved. The time consumptions of two schemes under different conditions of file size and data segmentation size are compared, and then a new scheme combining linear space signature and with information theory is provide to fight pollution attacks.**

Keywords- network security; network coding; pollution attacks; attack detection; linear space signature; information theoretic security; secure network coding

I. INTRODUCTION

The network coding[1, 2], which is an information technology integrating coding and routing, can increase single transmission information of received packets by information fusion based on traditional store-forward method. Thus, the whole network performance is improved. The network node codes the received messages, improving the net-work transmission but with some safety risk. The network coding is subject to pollution attacks [3], and any revised node coding by fusion will pollute data and then transmit to next node. If next node does not detect the polluted data and directly codes the data, the new pollution will rise. The subsequent spread will paralyze the whole network. Thus, it is necessary to detect the data pollution in the network coding.

There are several solutions to pollution attacks in academic circles. Krohn, etc. [4] and Gkantsidis etc. [5] provided a homomorphic hash function scheme to detect data pollution in intermediate nodes. However, this scheme is figured by high calculation cost and needs a safety channel, which makes it hard to conduct this scheme in practical net-work. Charls, etc. [6] developed a new scheme with homomorphic and digital signature with no need of safety channel with independent matching calculation, but his scheme on elliptic curve calculation costs so much time. Zhao, etc. [7] presented a linear space signature method to detect effectively whether the received data vectors exist in the original vector space but with high time consumption. Levente etc. [8] established an information theory scheme to detect pollution attacks by existing data redundancy without any abstract addition into data vectors. This method is suitable for distributed storage model, but the pollution can be detected and revised only when the coding vector number that the nodes receive is over the minimum coding vector number. However, not all the nodes in practical network match this requirement, which limits its application.

This paper presents the linear space signature scheme and information theory method to fight the possibly existing pollution attacks, and the detection algorithm based on information theory is improved. Furthermore, the time consumptions of two schemes under different conditions of file size and data segmentation are compared. According to the analysis results and practical network environment, a scheme combining linear space signature and information theory is designed with calculation algorithm. This scheme can fight pollution attacks effectively, and then decreases the time consumption.

II. NETWORK CODING AND POLLUTION ATTACKS

A. Single-source multicast network coding model

The education circles in China have been fully aware of the importance that discovering students' innovative potential plays in nurturing innovative talents. They realize that the earlier students develop an interest in research and the earlier engineering education is conducted, the better. Therefore, engineering education for undergraduates has become a trend with the growing demand for talents. Many research universities not only organize students to take part in various extra curriculum scientific activities, but also manage students to practice in multinational corporations and state-owned large enterprises. Compared with undergraduates engineering education abroad, we are fairly late in organizing undergraduates to carry out scientific research and engineering education. Despite the rapid development in recent years, there is still a lack of operation model with scale and maturity; especially there is a blank in students' initiative systemic research and engineering education. In view of that, it is significant to further explore bold innovation and strengthen the nurturing model for innovative talents by probing into the theories and practice of engineering education.

Assume the single-source multicast network coding model includes one source node, several intermediate nodes and receiving nodes, as shown in Figure 1. Here, the roles of source node, intermediate nodes and receiving nodes, which can be generally found in complex network coding environment.

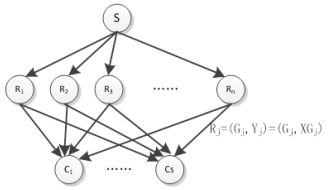

Fig. 1. Single-source multicast network coding model

978-0-7695-4694-0/12 $26.00 © 2012 IEEE
DOI 10.1109/ICIS.2012.5

Assume X denotes the file to be sent from source node, and divide X evenly into n data blocks. Let each block has m elements, then $X_i \in \mathbb{F}_p^m$, where p is prime number. Thus, the file can be expressed by following matrix:

$$X_{m\times n}=\begin{pmatrix} x_{1,1} & x_{1,2}\cdots x_{1,n} \\ x_{2,1} & x_{2,2}\cdots x_{2,n} \\ \vdots & \vdots \quad\quad \vdots \\ x_{m,1} & x_{m,2}\cdots x_{m,n} \end{pmatrix}=(X_1\ X_2\cdots X_n) \tag{1}$$

Before the source data is sent, the extended data block is

$$\left(\overline{X_i}\right)^T=(\overbrace{0,\cdots,0}^{i},\overbrace{1,0,\cdots 0}^{n},X_i^T)\in\mathbb{F}_p^{n+m} \tag{2}$$

For each extended data block to be sent, the system selects the coding vector at random as

$$G_i=(g_{1,i},g_{2,i}\cdots g_{n,i})^T, i=1,2,\cdots \tag{3}$$

Code X_i, and then obtain the following expression

$$Y_i=\overline{X}G_i=\sum_{j=1}^{n} g_{j,i}\overline{X_j} \tag{4}$$

Send the expression above to the next nodes.

Any node except source node in network codes immediately the received k data blocks $Y_1,Y_2\cdots Y_k\in\mathbb{F}_p^{n+m}$ from k paths by $C=\sum_{j=1}^{k} g_{i,j}Y_j, g_{i,j}\in F_p$, and then C is sent to next nodes.

For nodes in non-error network environment, if n linearly independent coding vectors is obtained as

$$\left(Y_1\ Y_2\cdots Y_n\right)=\begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,n} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,n} \\ \vdots & \vdots & & \vdots \\ y_{n+1,1} & y_{n+1,2} & \cdots & y_{n+1,n} \\ y_{n+2,1} & y_{n+2,2} & \cdots & y_{n+2,n} \\ \vdots & \vdots & & \vdots \\ y_{n+m,1} & y_{n+m,2} & \cdots & y_{n+m,n} \end{pmatrix}=\begin{pmatrix} G_{n\times n} \\ \overline{Y}_{m\times n} \end{pmatrix} \tag{5}$$

then the original data matrix $X_{m\times n}$ can be decoded by $\overline{Y}_{m\times n}=X_{m\times n}G_{n\times n}$ in terms of $X_{m\times n}=\left(X_1\ X_2\cdots X_n\right)$ and $\overline{Y}_{m\times n}=X_{m\times n}G_{n\times n}$.

B. Pollution attacks model

To damage the common network communication, the positive attack generally found in the network is to falsify, counterfeit or reset the message vectors transmitted in the network. The attacker creates illegal coding vectors in the intermediate nodes and then sends these vectors to next nodes. In the next nodes, these illegal vectors are coded with other coding vectors. Thus, more nodes receive polluted vectors, and finally the destination will possibly receive false message.

Assume the destination nodes receive k coding vectors in the condition of pollution attacks, the received data vectors can be expressed as $R_{1\cdots k}^{*}=(G_{1\cdots k}^{*},Y_{1\cdots k}^{*})$, where $G_{1\cdots k}^{*}=(G_1^{*},G_2^{*},\cdots,G_k^{*})$, $G_{1\cdots k}^{*}=G_{1\cdots k}+\Delta G_{1\cdots k}$, $Y_{1\cdots k}^{*}=(Y_1^{*},Y_2^{*},\cdots,Y_k^{*})$, $Y_{1\cdots k}^{*}=Y_{1\cdots k}+\Delta Y_{1\cdots k}$, Here the decoded data are $X^{*}=Y_{1\cdots k}^{*}(G_{1\cdots k}^{*})^{-1}$, If without pollution attacks, here $X=Y_{1\cdots k}(G_{1\cdots k})^{-1}$.

If the attacker only modifies one element in some one column of $Y_{1\cdots k}$, all the elements in this column of original data will be modified. However, if the attacker only modifies one element in $G_{1\cdots k}$, $(G_{1\cdots k}^{*})^{-1}$ will differ completely from $(G_{1\cdots k})^{-1}$. Thus, the whole original data vectors will be revised. If the attacker modifies the $Y_{1\cdots k}$ and $(G_{1\cdots k}^{*})^{-1}$ at the same time, the effects above will overlap. Actually, if the attacker modifies adequate coding data packets by some strategy, the destination nodes will receive the modified data which differ totally from original data.

III. DETECTION OF POLLUTION ATTACKS

It is very necessary to detect the pollution attacks because the network subject to pollution attacks. This paper introduces briefly the linear space signature scheme and information theory scheme.

A. Linear space signature scheme

For random linear network coding, the data blocks transmitted in the network include the global coding in existing path and data combined linearly by original data. If the extended data block is taken as the base of n-dimension space, any coding vector in the network is a linear combination on this base in terms of $Y_i=\overline{X}G_i=\sum_{j=1}^{n} g_{i,j}\overline{X_j}$ .Let $V=span\{\overline{X_1},\overline{X_2},\cdots,\overline{X_n}\}$, the received coded data can be detected by $Y_i\overset{?}{\in}V$. If $Y_i\in V$, the coded data are not polluted, otherwise the data are polluted.

The parameters [9] for linear space signature scheme are listed in Table 1.

Table 1. Parameters for linear space signature scheme

Name	Description
p	Big prime number, $2^{159}<p<2^{160}$
q	Big prime number, $2^{l-1}<q<2^{l}$, $512\le l\le 1024$ and $p\mid q-1$
g	Generator of subgroup G with p orders on $\mathbb{F}_q$, here $g=h^{q-1/p} \bmod q$
K_{pr}	Private key, $K_{pr}=\{k_i\}_{i=1,\cdots,n+m}$ is the pseudo random series on $\mathbb{F}_p^{*}$
K_{pu}	Public key, $K_{pu}=\{\beta_i=g^{k_i} \bmod q\}_{i=1,\cdots,n+m}$

The general principle of linear space signature scheme is described as follows:

1. Setting stage: the parameters of source nodes listed in Table 1 are determined, and (p,q,K_{pu}) are sent to all the nodes in the network in public.
2. Signature stage: a vector $u=(u_1,u_2,\cdots,u_{n+m})$ is selected randomly from the vertical subspace $V^{\perp}$ of source node vector space V. Then the signature vector $S=\left(\frac{u_1}{k_1},\frac{u_2}{k_2},\cdots,\frac{u_{n+m}}{k_{n+m}}\right)\in \mathbb{F}_p^{n+m}$ of V is calculated, and send it to other nodes safely by traditional method.
3. Check stage: the intermediate nodes and destination nodes receive (p,q,K_{pu}), S and coding vector of $n+m$ dimensions. Calculate $d=\prod_{i=1}^{n+m}\beta_i^{s_i y_{ji} \bmod p} \bmod q$, $Y_j\in V$ for $d=1$, where Y_j is legal vector, and $Y_j=(y_{j,1},y_{j,2},\cdots,y_{j,n+m})$ then store Y_j in the coding queue or coding buffering queue. If $d\neq 1$, the data is thought to be polluted. Thus, Y_j is illegal vector, and then discard it.

B. Information theory scheme

For the information theory scheme in the Reference [8], the destination node receives at most k linearly independent coding vectors. If $k=n$, the destination node can decode the data X^*. If $X^*=X$, the received vectors are not polluted, otherwise some unknown vectors are polluted in the n coding vectors. If $k>n$, the destination node may receive the additional unpolluted coding vector $(G_{n+1}Y_{n+1})$. Decoding the former n vectors can obtain X^*. If $Y_{n+1}=X^*G_{k+1}$, we think $X^*=X$. Namely, the former $n+1$ vectors are not polluted, or may be all polluted but with little probability (assume t of k received vectors are polluted, the probability of the former $n+1$ vectors is $P\approx\left(\frac{t}{k}\right)^{n+1}$).

The pollution detection by information theory is presented as follows:

(a) The destination node receives n coding vectors;
(b) Calculate X^* in terms of n linear independent vectors, $X^*=Y^*_{1\cdots n}\left(G^*_{1\cdots n}\right)^{-1}$;
(c) The destination node receives continually the n+1th coding vector $R^*_{n+1}=(G^*_{n+1}\ Y^*_{n+1})$;
(d) Calculate $Y^*_{n+1}\overset{?}{=}X^*G^*_{n+1}$ in terms of n+1th coding vector and X^*. If matched, there are no pollution attacks, otherwise the data are polluted.

The scheme above may result in false positive. If the former n coding vectors are not polluted and the n+1th coding vector is just polluted, the step (d) will not match, with false positive probability of $P_{fpos}=\frac{t}{k-n}$. This problem can be solved by revising step (d). Calculate $Y^*_{n+1}\overset{?}{=}X^*G^*_{n+1}$ in terms of n+1th coding vector and X^*. If matched, there are no pollution attacks, and the detection is completed. If not matched, the n+2th coding vector is received for continual detection until kth vector. If one or more expression is matched, there are no pollution attacks, otherwise the attacks occur. This solution can reduce the false positive probability greatly. The false positive will not occur if there is one or more unpolluted data block in vector set of $(n+1,n+2\cdots,k)$.

The steps above can detect the pollution attacks, but they can not determine the polluted coding vectors and get the number of polluted vectors. The Reference [8] provides three algorithms to recover the original data. The improved pollution attack detection algorithm is presented as follows:

```
  // attack_detction algorithm
  // if pollution exists, return
TURE; if not, return False
  Download  R*_{1...n}
  X* = decode(R*_{1...n})
  For  i =n+1  to  i=k
       Download  R*_i
  If    Y*_i = X*G*_i
  Return  FALSE
  End if
  End for
  Return TURE
```

IV. COMPARISON ANALYSIS

The linear space signature scheme conducts much exponential calculation for detection of each vector, and the signature scheme needs additional message redundancy. It is hard to match these requirements for node with weak calculation and high transmission speed. The information theory scheme can reduce time consumption by detecting $n+1$ vectors at a time without any additional redundancy. However, the information theory scheme need receive redundant vectors, namely $k>n$, and then the application node is limited.

The arithmetic complexities of linear space signature scheme and information theory scheme under the condition of $k>n+1$ and $2^{159}<p<2^{160}$ are compared as follows.

Let T_e denote the time consumption of modular exponentiation calculation, T_m the time consumption of modular multiplication calculation, T_1 the time consumption of linear space signature scheme, T_2 the time consumption of information theory scheme.

The pollution calculation by linear space signature scheme on n coding vectors is nd, where

$d = \prod_{i=1}^{n+m} \beta_i^{s_i y_{ji} \bmod p} \bmod q$. Calculation on d involves the modular exponentiation and modular multiplication, and the modular exponentiation comprises of a series of modular multiplication.

$$T_1 = n\ (n+m)(T_e + T_m) \tag{5}$$

The time consumption of modular exponentiation calculation depends on the algorithm. The square-multiply algorithm is used to calculate the modular exponentiation in Reference [10], with T_e about $80T_m$, because the exponent and base are both 160-bit binary number.

The information theory scheme need solve $(G^*_{1\cdots n})^{-1}$, $Y^*_{1\cdots n}\left(G^*_{1\cdots n}\right)^{-1}$ and $X^* G^*_{n+1}$ generally by matrix modular multiplication, here $T_2 = (n^3 + mn^2 + mn)T_m$ without consideration of nodes being polluted.

According to the well-developed BitTorrent protocol [11], the data block is generally 256KB, 512KB or 1024KB in size. The data block sizes for three various file sizes are compared in Table 2.

Table 2. Comparison of data block sizes between three various files sizes

Parameter		1MB	10MB	100MB
256KB	n	4	40	400
	m	13103	13067	12707
512KB	n	2	20	200
	m	26212	26194	26014
1024KB	n	1	10	100
	m	52427	52418	52328

The time consumptions for data blocks with different sizes on detecting n coding vectors are shown in Figure 2.

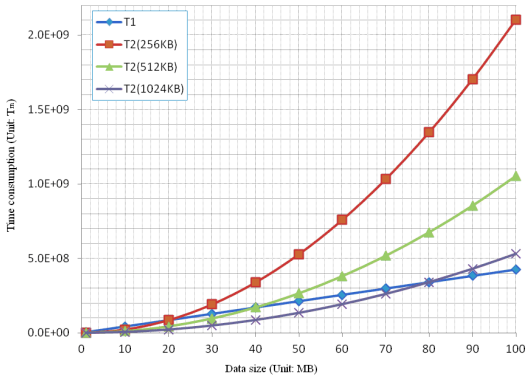

Fig. 2. Time consumption

As shown in figure 2, T_1 increases with data size, and T_1 is not subject to data block size; T_2 increases rapidly with data size, and T_2 is subject to data block size; the larger the data block size is, the smaller the n and T_2 are. If n is less than 80, the information theory scheme take less time to detect pollution attacks than linear space signature scheme. The smaller n is, the more obvious this advantage is. If n is over 80, the linear space signature scheme take less time to detect pollution attacks than information theory scheme. The larger n is, the more obvious this advantage is.

V. COMBINATION SCHEME

The received coding vectors must be detected one by one for the safety of all the nodes in the actual network if a few vectors are just polluted, which will cost much time. The information theory scheme provides an approach to conduct the batch detection, which can reduce time consumption in special condition. To detect the pollution attacks effectively, this paper presents a new scheme combining linear space signature method and information theory method without any redundancy addition into the existing model. This new scheme can solve the problems about limited application and possible false positive due to the vector number requirement by information theory method, and at the same time reduce the time consumption that the linear space signature method results in.

For any node except the source node, the new scheme judges whether the number of received coding vectors is larger than the minimum for decoding and whether the number of data blocks are less than minimum threshold. If the requirements above are matched, the information theory scheme is then used for pollution detection; if not matched, the linear space signature scheme is used for detection.

The general steps for new scheme are described as follows:

(a) Receive $R^*_{1\cdots k}$ from previous nodes;

(b) If k>n and n is less than threshold, the information theory scheme is used for pollution detection, and return $R^*_{1\cdots n}$ if no pollution exists, and then goes into step (d); otherwise, goes into step (c);

(c) The linear space signature scheme is used to detect pollution, remove polluted vectors, and finally return vector set $R'_{1\cdots j}$ without pollution;

(d) The pollution detection is completed.

The algorithm of this new scheme is listed as follows:

```
  //  n is the number of data
segmentations
  Get  n
  Int  j=0
  Set  R'_{1...j} ∈ F_p^{n+m}
  Set  Threshold=X // set threshold,
where x is determined by network
environment
```

```
Set  Flag=TURE
Download   R*1…k
BEGIN
If  (k>n&&n< Threshold)
     {
      // detected by information theory method
          For  i =n+1  to  k
{
If  attack_detction (R*1…n, R*i )  = no attack
Return  R*1…n
               }
          Flag=FALSE
     }
   Else if(k<=n && n>=Threshold || Flag==FALSE)
  {
   // detected by linear space signature method
          For  i =1  to  k
     {
         j=1;
If  Linear Space Signature_detection( R*i )=no attack
          { R'j  =  R*i ;
              j++;
            }
       }
Return  R'1…j
  }
   END
```

Assume all the nodes in the network match k>n, the linear space signature scheme, information theory scheme and combination scheme are compared under the conditions of different data size, network environment, with data block size of 1024KB, without threshold being set. The results are shown in Figure 3, where T_1, T_2 and T_3 denote the time consumptions of three schemes, respectively.

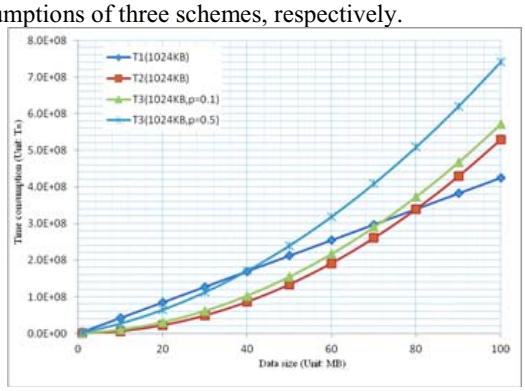

Fig. 3. Time consumptions of three scheme

After the pollution is detected by information theory scheme, the polluted vectors shall be removed by linear space signature scheme. Assume the pollution probability of a node is p, $T_3 = T_2 + pT_1$, the combination scheme is more effective than the linear space signature scheme only when the time consumption T_2 by information theory scheme is (1-p) times than T_1 by linear space signature scheme. As shown in Figure 3, when p=0.5 and n=40, then T_2 =0.5 T_1 , the time consumption T_3 by combination scheme equals to T_1 by linear space signature scheme. Assume the threshold is 40, the time consumption by combination scheme is less than that by linear space signature scheme; otherwise the both are equal. The determination of threshold depends on the algorithm for modular exponentiation calculation and the probability to pollute the nodes. If other algorithms are used to calculate the modular exponentiation, the threshold can be adjusted accordingly. The larger the pollution probability of a node, the smaller the threshold is. In actual network, not all the nodes match $k > n$. The nodes that do not match $k > n$ will be detected by linear space signature scheme. Thus, the combination scheme is suitable for usually actual network.

VI. CONCLUSIONS

This paper improves the pollution detection algorithm provided in Reference [8], which reduces the false positive probability of pollution detection. Based on Reference [7] and [8], a new scheme is provided in this paper to conduct the batch detection, without any redundancy and any effect on the existing scheme. The combination scheme can reduce the time consumption effectively on pollution detection for data of a few blocks.

ACKNOWLEDGMENT

The work is partly supported by the Natural Science Foundation of Hubei Province, China (Grant No. 2011CDB048) and the supported by 'the Fundamental Research Funds for the Central Universities', HUST: 2011QN077

REFERENCES

[1] R Ahlswede, N Cai, S Y-R Li, et al, Network information flow[J],IEEE Trans.Inf. Theory, 2000,46(4):1204-1216.

[2] LI S Y-R, YEUNG R W, CAI N Linear network coding[J], IEEE Trans.Inf.Theory, 2003, 49(2):371-381.

[3] C Fragouli, J-Y L Boudec, J Widmer, Network coding: an instant primer[J], SIGCOMM Comput.Commun.Rev, 2006,36(1):63-68.

[4] M Krohn, M FreedMan , D Mazieres , On-the-fly verification of rateless erasure codes for efficient content distribution[J],IEEE symposium on Security and privacy,Oakland,CA,2004:226-240.

[5] C Gkantsidis, P Rodriguez , Cooperative Security for Network Coding File Distributio IEEE INFOCOM, Barcelona, 2006.

[6] Charles D,Jian K,Lauter K. Signature for Network Coding. Technique Report MSR-TR-2005-159, Microsoft[C], 2005, 7(13):153-158.

[7] Fang Zhao,Ton Kalker , M′ edard M , et al. Signatures for Content Distribution with Network Doding,ISIT2007, Nice, France[A], 2007, 6(13):37-39.

[8] Buttyan L,Czap L,Vajda,I,Detection and Recovery From Pollution Attacks in Coding Based Distributed Storage Schemes[J],IEEE Trans.Dependable and Secure Computing,2011,8(6):824-838.

[9] U.S.Department of Commerce.Digital Signature Standard[S].2000.

[10] Yin Xinchun, Zhang Baohua. A Parallel Window Algorithm for Large Integer Modular Exponentiation[J]. Computer Engineering and Application. 2004, 25(18):50-53

[11] Cohen B.The BitTorrent Protocol Specification[EB/OL].(2004-12-12).

[12] http://www.bittorrent.org/.