Orthogonal Vector based Network Coding against Pollution Attacks in n-layer Combination Networks

Bin Dai, Shijun Zhang*, Yipeng Qu, Jun Yang, Furong Wang

Department of Electronic and Information

Huazhong University of Science and Technology, Wuhan, Hubei, China

* Corresponding author, E-mail: zhangsj@hust.edu.cn

***Abstract*—It is proved that combination networks can achieve unbounded network coding gain measured by the ratio of network throughput with network coding to that without network coding in peer-to-peer file sharing system. On the other side, network coding allows the polluted packet generated by a malicious node to corrupt the packets transmission of the entire network. One polluted packet will prevent the other nodes from correct decoding. To settle the problem, we introduce a polluted packets detection method based on multiple vectors orthogonal to the universal set of vectors for the transmitted network coding packets. The main idea of this method is to calculate numbers of orthogonal vectors and send them to different relay nodes and receiver nodes which use them to judge whether the newly received packets are in the subspace based on the source packets. This is a distributed algorithm, without the requirement of an extra secure channel. Through the analysis and simulation in n-layer combination networks case, the computational complexity of the detection algorithm is O(n), n is the packet length, and the detection rate can be greater than 90%.**

Keywords: combination network, network coding, pollution attack, orthogonal vector

I. INTRODUCTION

Network coding was introduced by Ahlswede, et al of Chinese University of Hong Kong in 2000[1]. Then Li et al proved that using linear network coding can achieve the max-flow min-cut capacity of a general multicast network[2]. In 2003, Medard, et al gave the algebraic approach to linear network coding[3]. In the same year, random linear network coding (RLNC) was introduced by Ho et al, and its decoding rate can be higher than 99% when using RLNC.

Network coding theory shows that the multicast rate can be increased if coding is allowed in the network nodes. Follow the conclusion, Ngai and Yeung studied a class of networks called combination networks[4] and gave the definition of combination networks and the network coding gain of combination networks. Figure 1 show an example of C_3^3 combination network.

Based on the theory of combination network coding, Yang, et al proposed a deterministic linear coding[5] over combination networks to improve the peer-to-peer file sharing system performance in terms of throughput, reliability and link stress.

CHINACOM 2010, August 25-27, Beijing, China

DOI 10.4108/chinacom.2010.89

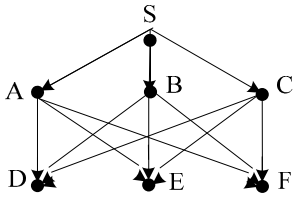

Fig. 1. a C_3^3 combination network

As well as network coding can improve the system performance, it also brings some new challenges. The security aspect of network coding is a tough one and catches the attention of many researchers[6-8]. Network coding has the nature of data mixture which makes it has both advantage and disadvantage to network security. For a wire-tap node, it needs to tap more links to make out the decoded packets. But when a malicious node injects polluted packets to the network, the mixture of data by network coding would make more and more polluted packets, so that the receiver nodes cannot decode.

In traditional network, hash and sign method is used to prevent the malicious nodes pollute data. The main idea of the method is that the source node calculates a signature for each packet it sends and the relay nodes and receiver nodes compare the signature and packet to judge whether it is a correct packet that is sent by the source node. But in the network using network coding, this method is not suitable, since all of receiving data are not the data sent by the source originally, and the source cannot calculate the signature for each possible outcome. So the network coding system needs some other methods to detect the pollution messages.

Krohn, et al proposed a method of on-the-fly verification of rateless erasure codes[9]. In this method, source node calculates the hash value for each packet in the network and forwards them to the other nodes. The distribution of hash value needs secure channels from the source to each destination node which uses the hash value to calculate and compare with its own signature. Gkantsidis gave a homomorphic hashing algorithm to ensure the security of peer-to-peer file sharing system to avoid pollution attacks[10]. In this method the hash signatures and the secure channels are also necessary, but the amount of them will not increase as the packet increases any more. Li, et al introduced another improved algorithm for the detection of pollution attack

in network coding[11]. Source node only calculates a part of hash signatures and delivers them along with the packets. This method reduces the amount of hash signatures, but may lead to the destination nodes can't decode. Charles, et al gave a network coding signature algorithm build on top of expensive Weil paring operations homomorphic hashings[12]. It has large computational complexity and time consuming. [13][14] gave methods to do detection of the polluted packets in network coding systems, but they detect only at the destination nodes, not to the intermediate nodes, so they aren't helpful to improve the security of the whole network. [15][16] gave other algorithms to detect and filter the pollution packets. But all of them need hash signature of network coding.

Referring to the orthogonal signature scheme proposed in [17], we introduce a polluted packets detection algorithm based on orthogonal vectors which does not demand any signature scheme for the polluted packets detection with the lower computational complexities and adopts random sample method to reduce the transmission overhead, it's more suitable for the practical network coding system. In section 2 we describe the main principle of the algorithm. Section 3 analyzes its performance and cost. Section 4 gives the simulation results of the orthogonal vectors based detection algorithm.

II. POLLUTION DETECTION ALGORITHM DESCRIPTION

1. System Setup

A transmission network can be represent by a directed graph $G=(V,E)$. All the work is under the assumption that the network is solvable without the occurrence of the malicious nodes. Here are some other assumptions as follows,

1. The target network is an n-layer combination network.
2. The source node has r out-degree.
3. Most of the other nodes in the network have in-degree m and out-degree n, include the malicious node hidden in the network. m and n are equivalent.
4. All nodes know the number of direct downstream nodes.

When the source node sends h packets with N symbols, the message matrix in network coding can be presented as follows,

$$\begin{bmatrix} x_1 \\ \vdots \\ x_h \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 & x_{1,1} & x_{1,2} & \cdots & x_{1,N} \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & x_{h,1} & x_{h,2} & \cdots & x_{h,N} \end{bmatrix} \quad (1)$$

In the matrix $x_{i,j}$ are all symbols in $\mathbb{F}_q$. Message transmitted in any edge e is the linear combination of x, that is $y(e)=\sum_{i=1}^{h} g_i(e)x_i$. $g_i(e)$ is the global encoding vector of network coding. So the vectors $y(e)$ are in $\mathbb{F}_q$.

2. Design of Orthogonal Vector

Since vectors $x_1, x_2, ..., x_h$ are linear independent vectors in $\mathbb{F}_q^{h+N}$, they can span a subspace X. The vectors $y(e)$ are all linear combination of $x_1, x_2, ..., x_h$, so they are all in the subspace X. Each of the vectors in subspace X can be seen as a legal vector, which means the vector is not fabricated. So we can judge whether a vector is a legal by checking whether it is in the subspace X. To do so, orthogonal vectors can be used.

According to the vectors $x_1, x_2, ..., x_h$, we can find a vector $v=(v_1, v_2, ..., v_{h+N}) \in \mathbb{F}_q^{h+N}$ satisfied the condition that $x_i \cdot v = 0, \quad i=1,2,...,h$. It means that the vector v is orthogonal to all x, and v is orthogonal to the subspace X. The source node distributes the vector v to the downstream nodes. Then the downstream nodes check the vector y of each received packet whether it is orthogonal to v, that is to compute whether $y \cdot v$ is equal to 0. If the result is not zero, the packet is a polluted one.

If a malicious node want to get a packet that is not in subspace X but can pass the check without knowing the vector v, on the assumption that it randomly choose a vector, and this vector has a probability $1/q$ to pass the check. In RLNC, q is always chosen to be 2^8 or 2^{16}. The probability $1/q$ is very small and negligible. But if the malicious node knows v, the detection method will fail, so multiple vectors are used. Each node gets different detection vectors and uses them to check.

Since the transmission of multiple vectors will greatly increase the transportation overhead, we use k vectors of length $h+k$ to detect $h+k$ symbols in subspace X. This $h+k$ symbol is randomly chosen by source node. Because the malicious node polluted each of the symbols with an equal probability, the detection rate have no different whether long or short orthogonal vectors are used.

3. Detection Algorithm of n-layer Combination Network

As malicious nodes hidden in the network have the same ability to get vectors from the source. The other nodes need different vectors so that they can do detections that the malicious node cannot generate packages that can pass the check.

Assume the source node separately compute k vectors $v_1, v_2, ..., v_k$, each of them contains $h+k$ symbols. Then the source node hash and sign them and send them to its downstream nodes separately and as fair as possible. Source node makes sure that each of the downstream nodes receives at least one vector and at most $\lceil k/r \rceil$ vectors, here r is the number of downstream nodes from the source node.

For the relay nodes and receiver nodes in the network, they receive detection vectors from upstream nodes. Assume the node get m different vectors randomly. Then the node sends the n vectors to its downstream nodes separately. It also sends the vectors as fair as possible.

Based on the assumption made and the vector distribution algorithm described above, the number of vectors each nodes (except the direct downstream nodes of source nodes) got will be almost the same, including the malicious nodes.

For two arbitrary nodes A and B separately receive s and t vectors, the probability that node B gets at least one vectors that unknown to A can be calculate through the following formula:

$$p = \begin{cases} 1 & t > s \\ 1 - \dfrac{C_s^t}{C_k^t} & t \le s \end{cases} \tag{2}$$

In formula 2, k is the total number of different vectors. Suppose node A is a malicious pollution attacker, node B is A's downstream nodes and will receive polluted packets from A. The vectors B received from A are known to A, so it is reasonable to assume $t = s - 1$. For fixed s and t, the change of k will lead to the change of probabilities that B gets vectors unknown to A.

The probability that B will get any vectors unknown to A is shown in figure 2, and this probability is also the detection probability.

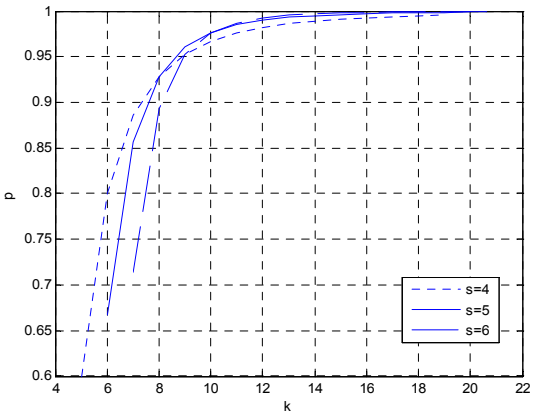

Fig. 2. Relationship between detection probability and number of vectors

In figure 2, the probability p directly shows the detection algorithm's ability of correctly discovering the polluted packets. Figure 2 shows that if a detection probability needs to be greater than 90%, k should be chosen to be no less than (8, 8, 9) while s is equal to (4, 5, 6) and the detection probability will goes to near 1 as k increase.

According to the assumption we have made that in-degree m and out-degree n of the network are equivalent, we construct a network topology shown as figure 3 to illustrate the detection algorithm. Figure 3 presents a 5-layer C_3^3 combination network. In the combination network, S is the source node, J, K and L are receiver nodes, and the other nodes are relay nodes of both good nodes and malicious nodes. And source S have direct downstream nodes of A, B and C. These three nodes have in-degree of 1 and other forwarding nodes' in-degree are all 3. And there out-degree are all 3. So here in-degree and out-degree of nodes are equal. According to the formula 2, if the detection probability wants to be higher than 90%, the number of vectors should be greater than 6. So source node S computes 6 different vectors numbered from 1 to 6 and sends them to node A, B and C. Each of nodes gets two different vectors. In figure 3, the set of vector number the nodes receive has been represented above the nodes. All nodes follows the vector fairly distribution algorithm.

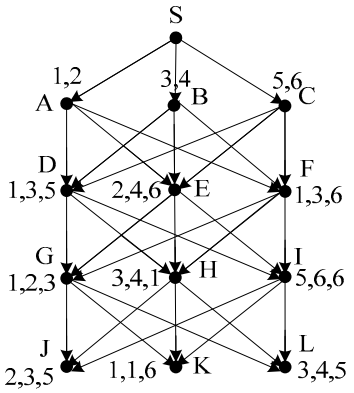

Fig. 3. Orthogonal vectors based detection algorithm in 5-layer C_3^3 combination network.

It can be seen in figure 3, some nodes receive the same vector from different upstream nodes, such as node I and K. In some case the relay node has only two different vectors but it has three downstream nodes. So it sends the same vector to two different downstream nodes. In the process, each node uses the vector they received to detect whether there are polluted packets. In the distributed condition, no matter which node is malicious, its downstream nodes have the ability to seek out the polluted packets by received orthogonal vectors.

III. Performance Analysis of Algorithm

1. Detection performance

In orthogonal vectors based polluted packets detection method, if each node has known vectors to each of the upstream nodes, the detection probability would be $1 - 1/q$ because of the inherent characteristics of orthogonal vectors. Since q is quite large, the probability is near to 1.

Due to vector distribution algorithm, not each of the node will have new vectors to the upstream nodes, and the probability of that is p less than 1. So some of the nodes will fail in detection of the malicious nodes and be

polluted. But their downstream nodes will do detections too. Suppose all detection probabilities are p, the probability of polluted packages will still in the network after passing n nodes is $(1-p)^n$. This process is shown in figure 4.

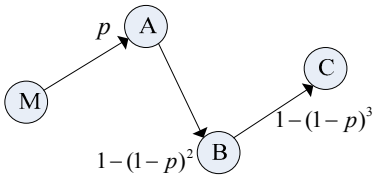

Fig. 4. Detection probabilities after fails

2. Costs analysis

(a) Transmission overhead

The main transmission overhead is the transmission of the detection vectors. If there are k vectors of $h+k$ symbols to be transmitted, the transmission overhead of vectors is $k\cdot(h+k)/h\cdot(h+N)$ times of the total packages of data per generation.

In general practical network coding systems, the number of packets transmitted per generation h and symbols N is much larger than k, so the proportion of vectors overhead in the total packages is much smaller than k/h, that is a quite low proportion.

(b) Computational overhead

When a new package arrives, the relay nodes conduct the package pollution detection by point multiplication operations. For each package check, the relay node only needs to execute m times point multiplication operations. In other words, the computation complexity of pollution detection for a node is $m(h+k)$ times multiplication operations and $m(h+k-1)$ times addition operations. To the nodes in a peer-to-peer file sharing system, the computational overhead is negligible.

(c) Start delay

Before the algorithm start, the source node should calculate orthogonal vectors and then distribute them to the relay nodes. The computation progress can be done before transmission, so it will not generate any start delay.

IV. Simulation Results

The simulation of polluted packets detection by orthogonal vectors is implemented using C language. We simulate the vector distributed progress and calculate the detection probabilities with the varying number of orthogonal vectors. The simulation network topology is a 6-layer C_4^4 combination network topology, satisfying the assumption we made in section 2, shown as figure 5.

In 6-layer C_4^4 combination network topology, there are one source node, four layers relay nodes and one layer receiver nodes, each layer contains four relay nodes. There is no link between the nodes at the same layer, but each pair of nodes at neighbor layer has a link between them.

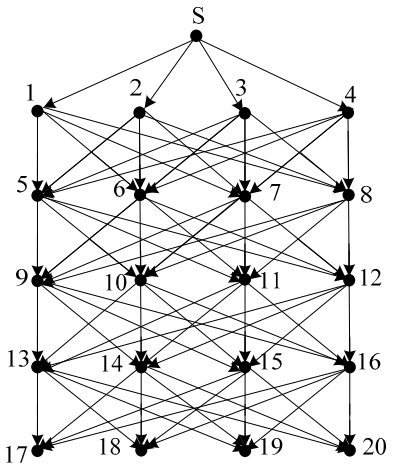

Fig. 5. A 6-layer C_4^4 combination network topology

Through the simulation experiments in 6-layer C_4^4 combination network, the simulation results are shown as figure 6.

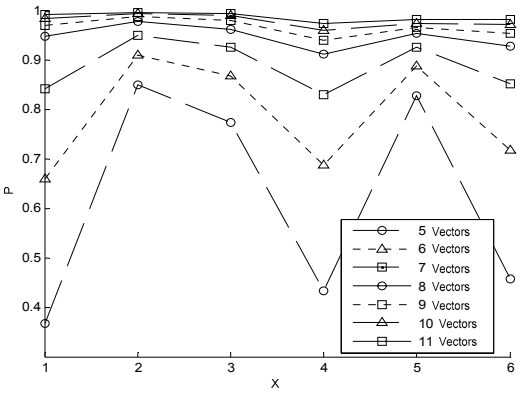

Fig. 6. Detection probabilities among different layer nodes in 6-layer C_4^4 combination network

Figure 6 shows the detection probabilities among different layer nodes with the varying vector number. Y axis means the average detection probabilities p, X axis means detection probabilities among different layer nodes, $x=1$ means detection probabilities between layer 3 nodes and layer 2 nodes, $x=2$ is layer 4 and layer 3, $x=3$ is layer 5 and layer 4, $x=4$ is layer 4 to layer 2, $x=5$ is layer 5 to layer 3, $x=6$ is layer 5 to layer 2. All simulations run 10^4 times to get the final results.

From the analysis of simulation results, it can be seen that when vector number is 5, the detection probabilities are very low. For examples, if vector number is 5, the detection probabilities between layer 3 and layer 2 nodes are no more than 40%, and the probabilities between layer 4 and layer 5 to layer 2 nodes are no more than 50%, so if the malicious node is

in layer 2, the pollution detection would fail with a high probability.

As the number of vectors increases, the detection probabilities increase. When there are 8 different vectors in the network topology, the entire detection rate is higher than 90% and most of them are about 95%. The detection probabilities are approach to 1 as the vector number increases. Both of security and transmission overhead are under consideration, the number of vectors should be a moderate value.

V. Conclusions and Future Work

In this paper, we have proposed an orthogonal vector based polluted packets detection method used in n-layer combination networks. We have also analyzed the cost and security aspects of the algorithm. Then simulation analysis is done to prove detection performance of the scheme. Through the analysis and simulation results, we find that the orthogonal vectors based method can reach a high detection probability to detect polluted packets with a low extra cost.

We have only discussed pollution detection algorithm under the combination network topology limited by assumptions. In the future work, we will do more research on how to implement the detection algorithm to adapt more network topology.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No.60803005, No.60872010 and No.60972016, and the Important National Science & Technology Specific Projects 2009ZX03004-004 and 2010ZX03003-003.

References

[1] G.J.Simmons, “And Introduction to Shared Secret and/or R. Ahlswede, et al, "Network information flow", IEEE Transactions on Information Theory, vol.46, no.5,pp. 1204-1216,Jul. 2000

[2] S. Li, R. Yeung and N. Cai, "Linear Network coding", IEEE Transactions on Information Theory vol. IT-49 pp.371-381,2003

[3] R.Kotter and M.Medard, "An algebraic approach to network coding", IEEE/ACM Trans. Networking, vol. 11, no.5, pp. 1973-1982, June 2005.

[4] C. K. Ngai, and R. W. Yeung, “Network Coding Gain of Combination Networks,” In Proc. IEEE Info. Theory Workshop, pp. 283-287, San Antonio, USA, Oct. 2004.

[5] Min Yang and Yuanyuan Yang, “Peer-to-peer File Sharing Based on Network Coding”, In Proc. The 28th ICDCS, pp. 168-175, 17-20 June 2008, Beijing, China.

[6] N. Cai and R. Yeung, "Secure Network Coding", ISIT 2002, Palais de Beaulieu, Lausnne, Switzerland, July 2002

[7] N. Cai and R. Yeung, "A Security Condition for Multi-Source Linear Network Coding", ISIT 2007, Nice, France, June 2007

[8] J. Tan and M. Medard, "Secure Network Coding with a Cost Criterion", IEEE Symp. on Ad Hoc and Wireless Networks 2006

[9] M. Krohn, M. Freedman and D. Mazieres, "On-the-fly verification of rateless erasure codes for efficient content distribution", IEEE Symp. on Security and Privacy 2004

[10] C. Gkantsidis and P. Rodriguez, "Cooperative Security for Network Coding File Distribution", IEEE INFOCOM 2005, Miami, FL, March 2005

[11] Qiming Li, Dah-Ming Chiu and John C.S. Lui, "On the Practical and Security Issues of Batch Content Distribution Via Network Coding", IEEE Conf. on ICNP 2006

[12] D. Charles, K. Jian and K. Lauter, "Signature for Network Coding", Technique Report MSR-TR-2005-159, Microsoft, 2005

[13] T. Ho, et al, "Byzantine Modification Detection in Multicast Networks Using Randomized Network Coding", ISIT 2004

[14] S. Jaggi, et al, "Resilient Network Coding in the Presence of Byzantine Adversaries", IEEE INFOCOM 2007

[15] Zhen Yu, et al, "An Efficient Signature-based Scheme for Securing Network Coding against Pollution Attacks", IEEE INFOCOM 2008

[16] S. Jaggi,et al, "Resilient network coding in the presence of byzantine adversaries", IEEE INFOCOM 2007

[17] Fan Zhao, et al, "Signatures for Content Distribution with Network Coding", ISIT 2007, Nice, France, June 2007