### NETWORK CODING FOR WIRELESS SECURITY

M. E. EKPENYONG AND B. L. IBANGA

ABSTRACT. In this paper, a Random Linear Network Coding (RLNC) scheme which takes advantage of redundant network capacity for improved success probability and robustness of wireless communication networks is proposed. We provide sufficient literature and illustrate vital potential benefits of this scheme over the normal Hierarchical Routing Scheme (HRS). Our model derivation yields a new bound on the required field size with useful transition probabilities. Specifically, the system overheads of both schemes are compared through extensive computer simulations in an ideal network setting. Simulation results reveal that optimum field size for effective throughput and overhead management are achievable with the RLNC scheme, and the delivery ratio can be controlled by varying the number of subnets (k) in both schemes.

Keywords and phrases: Network security, random linear coding, system overhead, transmission capacity.
2010 Mathematical Subject Classification: 68U20, 81T80

#### 1. INTRODUCTION

Wireless networks have evolved over the years to offer high speed connectivity and voice communication. They have provided unique opportunities in education and research. They also offer flexibility, by enabling access to network services regardless of physical location. Even as they offer these benefits, they are also susceptible to various attacks such as eaves dropping, data injection and node compromise. These attacks contribute to breaching the security, integrity, confidentiality and authenticity of data on the network. Naturally, the transmission of a network node may be affected by two factors namely, the traffic volume and the random access mechanism [1]. The effect of traffic volume determines the opportunities the nodes have during packet transmission, and will impact on the coding process. This effect becomes more prominent as traffic volume increases. The link-level random access mechanism also affects the number of buffered packets (under heavy traffic), where

Received by the editors February 16 2013; Revised: February 10, 2014; Accepted: February 18, 2014

the network code has equal channel access opportunities as their source nodes, and packets may accumulate quickly in its buffer, resulting in a high encoding number [1].

This contribution therefore deploys network coding as a means to improving communication security in wireless networks. Network coding broadly refers to coding at the intermediate nodes of a network (including routed data) between the source and destination. It represents a recent development in network security, where the communication nodes are permitted to perform algebraic operations within the network, and was first introduced in [2]. It is an attractive technology and has been shown to potentially improve the throughput of wireless networks [3]. Although its workings under practical settings remain unclear, a major concern of any network coding system is the protection against malicious nodes [4], and with the emergence of several coding architectures, it has become pertinent to address these performance issues [5], as well as adopt a more robust approach to tackling security within the network.

Theoretically, network information management can be quantified by the log of behavioural counts (codes) used in the network. Allowing general network coding solutions provide fundamental limits on the information required. The specifics of Scalable Information Operation (SIO) - a network information management solution include scalable coding techniques for network coding, compression, channel coding, multimedia data transmission, encryption, caching, network security, and intrusion detection. These operations are capable of mitigating some of the security challenges in network environments.

Under the classical network paradigm (in traditional networks) where intermediate nodes are only permitted to store and forward packets, information security is usually viewed as an independent feature with little or no relation to other communication tasks. Since intermediate nodes receive exact copies of the sent packets, data confidentiality is commonly ensured by cryptographic means at higher layers of the protocol stack. With network coding, the intermediate nodes are allowed to admit information from various data flows, thus, providing an intrinsic level of data security with privacy enhancement.

The Random Linear Network Coding (RNLC) - a distributed scheme for implementing network coding protocols is presented in this paper. Nodes in this scheme draw several coefficients at random

244

and use these coefficients to form linear combinations of incoming packets. The resulting packet is finally sent with the global encoding vector, which records the cumulative effect of the linear transformations suffered by the original packet while migrating from source to the destination. The global encoding vector enables the receivers to decode the sent packets by means of Gaussian elimination.

#### 2. REVIEW OF RELATED WORKS

The application of network coding to improve network performance and security in wireless systems has been extensively addressed in literature. Network coding has been shown to improve information flow rates in multicast sessions [2]. The main idea is to allow the nodes to perform coding operations instead of simple replication and forwarding (useful to alleviate competitions among flows). Network coding is performed within segments to which a random linear code is applied. The data are grouped into segments containing a number of blocks that defines the segment size. In a more practical setting, the random linear network coding, first proposed in [6], has been shown to be feasible where a node transmits simultaneously on each of its outgoing links, a random linear combination of incoming messages. For instance, the Avalanche, which uses randomized network coding for content distribution to reduce delay in download, is introduced in [7]. Another advantage of randomized network coding is its ability to increase data blocks diversity and improve resilience in block losses. When network coding is applied, a node can transmit any coded block since the node and coded block contribute equally to the delivery of data blocks. Wu [8] argued that network coding could adapt to network dynamics such as packet losses and link failures. Although network coding offers high throughput and better erasure protection capability, it does not guarantee the recovery of transmitted source data from received packets. This is because some packets could be lost, as such, the global coefficient matrix (GCM) associated with random linear network coding may not have the full rank in the RLNC decoding process at the destination node [9].

The potential benefits of network coding have been studied in various settings. The service curve model for performance analysis of wireless network coding has been analysed in [5]. In conducting the analysis, the authors presented a flow oriented definition of coding opportunity that supports per flow performance analysis, and derived formulations for computing the delay and backlog using opportunistic coding. They observed that network coding improves performance only under heavy traffic load but there is no performance difference under light traffic load.

In [10], the case of network coding as a unifying design paradigm for wireless networks is explored and the issues of throughput, reliability, mobility and management are addressed. They discussed the practical challenges facing the integration of such designs into a network stack, highlighting the benefits of network coding and new challenges in network coding, which include broadcast and coding. Zhu, Yue and Wang [3] have proposed a novel network coding scheme called C & M to deal with both interflow and intra-flow traffic. The first implementation of C-COPE (interflow) and M-MORE (intra-flow) network coding was carried out in [11-12]. They proposed a multiple transmission scheme to increase the throughput of wireless networks with low-link delivery. C & M attempts to combine the advantages of both network coding approaches (COPE & MORE) to develop a more efficient network coding scheme. Results of the research show that the design fulfilled the goal of obtaining high throughput and reliability, simultaneously on the network.

In [13], a class of erasure codes, showing that data could be efficiently diffused by "pre-routing" in only packets per node to randomly selected storage nodes, is introduced. The motivation behind this is to reduce latency and unreliability between query time and the time that the desired data is made available to the data collector. In providing a solution to this, each node was made to operate autonomously without any central point control. The decentralized erasure codes have a unique property that can be created in a distributed manner, without gathering all the input data in one centralised location. Decentralised erasure codes can also be seen as linear network codes [14]. The results of the decentralised erasure codes were found to introduce an increased reliable distributed storage. Yomo and Popovski [15] have addressed a scheduling problem for wireless network coding by introducing opportunistic scheduling for wireless network coding which selects a set of nodes whose packets and data rates are coded for the broadcast transmission. They analysed the average capacity of the scheduling and discussed the impact of different parameters on the average capacity, as well as the practical factors which can additionally affect the best scheduling strategy for wireless network coding. Their research results show that the strategy to combine packets from a fixed number of nodes does not yield the best performance and the opportunistic scheduling which can change the scheduling level adaptively according to the instantaneous link, can maximise the average capacity.

In [4] the use of network coding for security has been studied. The authors overviewed the theoretical limits and application areas of network coding-based security in the areas of robustness to Byzantine attackers and of distributed signature scheme for download, and previewed a network coding scheme which is not based on elliptic curves but designed specifically for random linear coding systems. In [16], a set of security mechanism designed for delaysensitive applications that harvest the robustness of network coding with manageable complexity and without compromising security is proposed. Schemes which explore the algebraic structure of network coding to achieve the dual goal of hierarchical fidelity levels and efficient security are considered. But, an analysis of the performance in terms of overhead of the proposed scheme outsmarts the other schemes.

This paper therefore seeks to maximise the power of network coding for effective security and enhancement of the overall throughput in network coded architectures, through secure practical network coding. A simulation approach is adopted to enable us study the behaviour of two network architectures and the system overhead employed as a measure to compare their performance.

## 3. SYSTEM MODEL

## **3.1 NETWORK MODEL ASSUMPTIONS**

In this section, the model assumptions guiding the proposed design are stated. The assumptions involve the nodes, activities and network environment, and are defined as follows:

Nodes in the network: (i) Nodes are not stand-alone (i.e., all nodes are connected to the network and participate in the network activities), (ii) No malicious nodes are present in the network model. *Network activities:* (i) All nodes implemented using RLNC can be a source or a sink node, and have an equal chance to be either, but not both (source and sink) for the same session, (ii) There maybe multiple instances of sessions with identical source and sink nodes, (iii) Paths are selected on a lowest hop count, with rudimentary link state selection criteria.

Network environment: The network environment is closed.

# **3.2 RLNC OVERHEAD**

Network overhead is an important concept because it reveals the

methodology employed by various protocols when routing information between two locations, and the costs involved. We formulate the network overhead for the proposed model as follows: Let the data, of size B awaiting transfer from a source to one or more sinks be segmented into generations of size,  $q \bullet m$ . A generation is sometimes referred to as a source block or a batch. Each generation constitutes g symbols of size m, where q represents the generation size. The original symbols of q, of length m in one generation are arranged in a matrix  $M = [m_1; m_2; ..., m_q]$ , where  $m_i$  is a column vector. In an application, the block of data can be a file or part of a media stream, and is divided into  $\left[\frac{B}{g \bullet m}\right]$  pieces, called symbols. Generation number zero (0) constitutes the first g symbols or the first  $g \bullet m$  bytes of data and there are  $d\left[\frac{B}{g \bullet m}\right]$  of such generations. To encode a new symbol x from a generation at the source, Mis multiplied by a randomly generated coding vector q of length  $q, x = M \times q$ . In this way we can construct q + r coded symbols and coding vectors, where r is any number of redundant symbols, as the code is *rate-less*. When a coded symbol is transmitted on the network, it is accompanied by its coding vector and together they form a coded packet. In order for a sink to successfully decode a generation, it must receive g linearly independent symbols and coding vectors from the said generation. All received symbols are placed in the matrix  $\hat{X} = [\hat{x}_1; \hat{x}_2; ...; \hat{x}_q]$ , and all coding vectors are placed in the matrix  $\hat{G} = [\hat{g}_1; \hat{g}_2; ...; \hat{g}_g]$  , where  $\hat{G}$  is the decoding matrix. The original data M can then be decoded as  $\hat{M} = \hat{X} \times \hat{G}^{-1}$ . To spread the computational load, the decoding is performed using Gaussian elimination. In practice, if approximately any q symbols from a generation are received, the original data in that generation can be decoded. This is a weaker condition, compared to when no coding is applied, where exactly all q unique original symbols must be collected. Any node can receive, g' symbols, where g' = [2, q] is the number of received linearly independent symbols from a generation, and is equivalent to the rank of G. To recode a symbol, these matrices are multiplied with a randomly generated vector hof length q'. Thus,  $\bar{q} = \hat{G} \times \hat{h}$ , and  $\bar{x} = \hat{X} \times \hat{h}$ . In this way we can construct  $r_0$  randomly generated recoding vectors and  $r_0$  recoded symbols. So, r' > q' is possible, and it is not permissible for a node to create more than q' independent symbols. Note that h is only used locally and that there is no need to distinguish between coded and recoded symbols. In practice this implies that nodes that have received more than one symbol can recombine those symbols into recoded symbols, similar to the way coded symbols are constructed at the source. Recall that g is the generation and  $g \bullet m$  is the generation size which represents the number of symbols over which encoding is performed. We define the maximum number of symbols that can be combined into a coded symbol and data is decoded per generation level. Thus, at least g symbols must be received before decoding is possible. Hence the size of a generation  $g \bullet m$  dictates the decoding delay which is the minimum amount of data that must be received before decoding is possible. Suppose q is the field size, the probability of generating g symbols is given as [17, 18]:

$$P(g) = 1 - \prod_{g'=0}^{g-1} \left( 1 - \frac{1}{q^g - g'} \right)$$
(1)

Thus as g' approaches g, it becomes increasingly difficult to receive useful symbols, because the coding vector must be non-zero in at least one of the g - g' corresponding scalars. This yields the following transition probabilities:

$$P_{g' \to g} = \frac{1}{q^g - g'} \to P_{g' \to g+1} = 1 - \frac{1}{q^g - g'} \tag{2}$$

Therefore, the expected amount of overhead for a generation can be obtained by evaluating the probability that the rank is not at capacity after k transmissions (i.e.  $P(g' \neq g)$ ). Initially no symbols are received and therefore the starting probability mass function (pmf), s, is s = [1, 0, ...]. When less than g symbols are received,  $P(g' \neq g)$ , and hence the overhead can be evaluated as:

$$\alpha \ge \sum_{k=g'}^{\infty} P^k(g' \neq g), P^k(P^k \times s)$$
(3)

This can be rewritten as:

$$\alpha(q,g) \geq \sum_{g'=0}^{g-1} \left( \left( 1 - \frac{1}{q^g - g'} \right)^{-1} - 1 \right)$$
$$= \sum_{g'=0}^{g-1} \left( \frac{1}{q^g - g' - 1} \right)^{-1}$$
(4)

The density d of a coding vector h with a generation size g is defined by:

$$d(h) = \frac{\sum_{k=1}^{g} h_k \neq 0}{g} = \sum_{g'=0}^{g-1} \left(\frac{1}{q^g - g' - 1}\right)^{-1}$$
(5)

Initially all columns in the decoding matrix consist of zero vectors and at least g symbols must be received for decoding to take place. Hence, the estimated number of symbols that must be received in addition to g before all columns contain non-zero values can be evaluated as:

$$\beta \ge \sum_{k=g}^{\infty} t^k (j \neq g), t^k (0^k \times s)$$
(6)

From this, the probability that at least one additional packet is required when k symbols have been received can be defined as:

$$\beta(q,g) \ge \sum_{g'=0}^{g-1} \left(1 - (1 - (1 - d)^k)^g\right); 0 > d \le 1 - q^{-1}$$
(7)

The total overhead of a given code is given by the expected number of redundant symbols necessary (i.e.,  $\frac{\alpha+\beta}{g}$ ). The coding vector is represented as a bit array that indicates which scalars are non-zero and the values of these scalars has the format:

$$a_0 a_1 \dots a_g \mid S_x \mid S_y \mid \dots \mid S_z$$

The bit array can be represented by g bits. Each of the scalars takes  $\log_2(q)$  bits, and on the average there are  $g \bullet d$  such scalars for each encoded symbol. If the bit array is compressed with an optimal code, the amount of bits necessary to represent it can be reduced from g to the entropy of the bit vector H(a), which can be computed from d and g. Thus,

$$\gamma = H(a) + \log_2(q).g.d \tag{8}$$

So, the total overhead constitutes the linear dependence and header overhead, divided by the size of a generation  $g \bullet m$  and is given by:

$$T_{ovh} = \frac{(\alpha + \beta).m + (g + \alpha + \beta).\gamma}{g.m}$$
(9)

where,  $\alpha$  is the overhead for a generation when less than g symbols are received,  $\beta$  is the estimated number of symbols that must be received before decoding, m is the symbol size, g is the number of

250

generation,  $\gamma$  is the overhead introduced by the coding vector and is defined by equation (8)

# 3.3 HRS OVERHEAD

The formula for computing the overhead of a typical hierarchical routing structure is given as [19]:

$$O(N) = o_p \mu(AN_p) N^2 / k + \phi + h\eta N \tag{10}$$

where,  $o_p$  is the broadcast optimisation factor,  $\mu$  is the link breakage rate,  $AN_p$  is the active links of a node, N is the number of nodes, k is the number of subnets in the network,  $\phi$  is the cost of locating the routes for communication, h is the packets emitted for proactive node discovery,  $\eta$  is the subnet crossover rate. In the next section, we simulate the performance of both schemes (the RLNC and HRS) and compare their performance. We rely on input parameters under ideal network conditions as a means of validating the performance of the proposed model in practice.

Input parameters	Values
$\alpha$ , overhead for a generation when less than g sym-	0.1 - 1
bols are received	
$\beta$ , estimated number of symbols that must be re-	0.8
ceived before decoding	
m, symbol size	1, 2, 8
g, number of generation	10 - 100
q, field size	1 - 10
d, coding vector density	3
H(a), bit vector entropy	2 - 20
$o_p$ , broadcast optimisation factor	0.7
$\mu$ , link breakage rate	0.5
$AN_p$ , active links of a node	1 - 10
N, number of nodes	10 - 100
K, number of subnets in the network	8, 16, 64
$\phi$ , cost of locating the routes for communication	0.6
h, packets emitted for proactive node discovery	10, 20, 100
$\eta$ is the subnet crossover rate	0.7

Table 1. Input parameters and values for proposed RLNC and HRS models simulation

# 4. MODEL SIMULATION AND RESULTS

## 4.1. SIMULATION INPUT

The input parameters used for simulating the two schemes are given in Table 1. Table 1 represents theoritical estimates obtained from wireless simulation studies [21]. The values are optimal and suitable for ideal network coding environments.

## **4.2 DISCUSSION OF RESULTS**

252

In Fig. 1, the effect of field size on network overhead for RLNC scheme is presented. The plots reveal that the symbol size decreases with increase in network overhead. The reason for this trend is because increase in the number of nodes ultimately affects the field size and each node is required to recode the data it receives, thus leading to a corresponding rise in the overhead. However, considerably low overheads can be obtained with small field size to guarantee the throughput of the network. A decrease in symbol size can be likened to a case of untrustworthy networks, where network nodes are frequently compromised to provide false data purposely or unknowingly (e.g., through network line tapping) and the data transmission is subject to channel error. But there exists an optimum field size that could maximise the throughput (bits/channel used) or minimise the symbol error probability.



Fig. 1. A graph of network overhead vs. field size in RLNC

Figs. 2, 3 and 4, relate the network overhead and network capacity (or generations) in RLNC and HRS schemes.



**Fig. 2.** A graph of overhead vs. number of nodes (for k = 8, h = 10)



**Fig. 3.** A graph of overhead vs. number of nodes (for k = 16, h = 10)



**Fig. 4.** A graph of overhead vs. number of nodes (for k = 64, h = 10)

For network coding to be efficient, the required number of transmissions must be reduced by at least the overhead. As observed in Figs. 2 and 3, the RLNC experienced a sharp rise in overhead at the initial stage, compared to HRS, but the former improved with increase in number of nodes. This proves the merits of RLNC over the normal routing and contributes to improving the robustness of the network in practice. But as the number of subnets grows, both schemes stabilise and nearly maintains same delivery ratio (see Fig. 4). Therefore, a corresponding increase in the generation size,  $q \bullet m$ , makes no sense with further increase in the number of subnets (k), since the whole packet payload would be exhausted as the system approaches its full capacity. Therefore, at N = 60, RLNC degrades to HRS, but small generation sizes may still be useful. In Figs. 5 and 6, we observe that as h (the packets emitted for proactive node discovery) increases, the HRS model experienced high network overhead because discovery packets and acknowledgements must be sent and received, resulting in severe network degradation and lowering of the throughput. This approximately doubles the network overheads and increases the matrix size. In practice, network coding should be efficiently managed to offer benefits even in the most constrained scenarios and the performance gains become much more significant for moderately large generations with less stringent delivery requirements.



**Fig. 5.** A graph of overhead vs. number of nodes (for k = 64, h = 20)



**Fig. 6.** A graph of overhead vs. number of nodes (for k = 64, h = 100)

### 5. CONCLUDING REMARKS

Fast and secure wireless network communication has posed a major concern for wireless network users. Although various routing and packet forwarding schemes are implemented on wireless networks, this paper has employed the random linear network coding to provide a packet mixing feature which stiffens the process of traffic monitoring and possible network intrusions. The scheme has also been found to increase the throughput and robustness of the wireless network. A mathematical model was derived based on the assumed properties of the RLNC scheme and validated through extensive computer simulations. The proposed model yields lower overheads than the Hierarchical Routing Structure (HRS), which indicates an improvement in security and the network throughput.

#### **ACKNOWLEDGEMENTS**

We acknowledge the anonymous reviewers for their useful comments.

#### REFERENCES

- J. Le, J. C. S. Lui and D. Chiu, On the Performance Bounds of Practical Wireless Network Coding, IEEE CS, CASS, COMSOC, IES, and SPS 10.1109/TMC.2010.77, 2010.
- [2] R. Alswede, N. Cai, R. S. Li and R. W. Yeung, Network Information flow. IEEE Transactions on Information Theory, 46:1204 -1216, 2000.
- [3] X. Zhu, H. Yue and Y. Wang, C and M: A Network Coding Scheme for Wireless Networks. In Proceedings of 5th International Conference on Information Assurance and Security, Xi'an, 2, 432-436, 2009.
- [4] K. Han, T. Ho, R. Koetter, M. Medard and F. Zhao, On network coding for security. In Proceedings of Military Communications Conference, Orlando, 1- 6, 2007.
- [5] H. Li, Q. Feng, G. Han, and W. Dou, Performance Analysis in Wireless Network Coding: An Approach Based on Service Curve Model. In Proceedings of the 3rd IEEE International Conference on Communication Software and Networks, Xi'an, 233-237, 2011.
- [6] T. Ho, R. Koetter, M. Medard, D. R. Karger and M. Effros, *The Benefits of Coding Over Routing in a Randomised Setting*. In Proceedings of IEEE International Symposium on Information Theory, 1-6, 2003.
- [7] C. Gkantsidis and P. R. Rodriguez, Network Coding for Large Scale Content Distribution. In Proceedings of IEEE INFOCOM, 1-12, 2005.
- [8] Y. Wu, Network Coding for Multicasting. Ph.D. Thesis, Princeton University, USA, 2006.
- H. Wang and C.- C. J. Kuo, Robust Video Multicast with Joint Network Coding and Video Interleaving. Journal of Visual Communication and Image Representation, 21(2), 77-88, 2010.
- [10] C. Fragouli, D. Katabi, A. Markopoulou, M. Medard and H. Rahul, Wireless Network Coding: Opportunities and Challenges, In Proceedings of Military Communications Conference, Orlando, 1-8, 2007.

- [11] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Medard and J. Crowcroft, XORs in the Air: Practical Network Coding. IEEE/ACM Transactions on Networking, 16(3), 497-510, 2008.
- [12] S. Chachulski, M. Jennings, S. Katti and D. Katabi, *Trading Structure for Randomness in Wireless Opportunistic Routing*. In proceedings of the ACM SIG-COMM, New York, 169-180, 2007.
- [13] A. G. Dimakis, V. Prabhakaran and K. Ramchandran, Ubiquitous Access to Distributed Data in Large Scale Sensor Networks Through Decentralized Erasure Codes, In Proceedings of the 4th International Symposium on Information Processing in Sensor Networks, IEEE Press Piscataway, NJ, USA, 2005.
- [14] T. Ho, M. Medard, R. Koetter, D. Karger, M. Effros, J. Shi and B. Leon, A Random Linear Network Coding Approach to Multicast. IEEE Transactions on Information Theory, 52, 4413- 4430, 2006.
- [15] H. Yomo and P. Popovski, Opportunistic Scheduling for Wireless Network Coding. In Proceedings of IEEE International Conference on Communications, Glasgow, 5610-5615, 2007.
- [16] L. Lima, J. Barros, M. Medard and A. Toledo, *Toward Secure Multiresolution Net*work Coding. In Proceedings of IEEE International Telecommunications Week, Volos, Greece, 125-129, 2009.
- [17] J. Heide, M. V. Pedersen, H. P. Fitzek and T. Larsen, Network Coding for Mobile Devices - Systematic Binary Random Rate-less Codes. In IEEE International Conference on Communications, Dresden, Germany, 1-6, 2009.
- [18] J. Heide, M. V. Pedersen, H. P. Fitzek and M. Medard, On Code Parameters and Coding Vector Representation for Practical RLNC. In IEEE International Conference on Communications, Kyoto, 1-5, 2011.
- [19] J. Lpez, J. M. Barcel and J. Garca-Vidal, Analysing the Overhead in Mobile Ad-Hoc Network with A Hierarchical Routing Structure, Report, Spain, P231-P239, 2005.
- [20] N. I. Sarkar, The Impact of Transmission Overheads on IEEE 802.11 Throughput: Analysis and Simulation, Journal of Selected Areas in Telecommunication, 2(3): 49-55, 2011.
- [21] R. Stoian, Random Network Coding for Wireless Ad-hoc Networks, In IEEE International Symposium on Signals, Circuits and Systems, lasi, 1-4, 2009.

### DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF UYO, UYO NIGE-RIA

*E-mail addresses*: mosesekpenyong@(uniuyo.edu.ng, gmail.com)

DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF UYO, UYO NIGERIA

*E-mail address*: icebankis@gmail.com