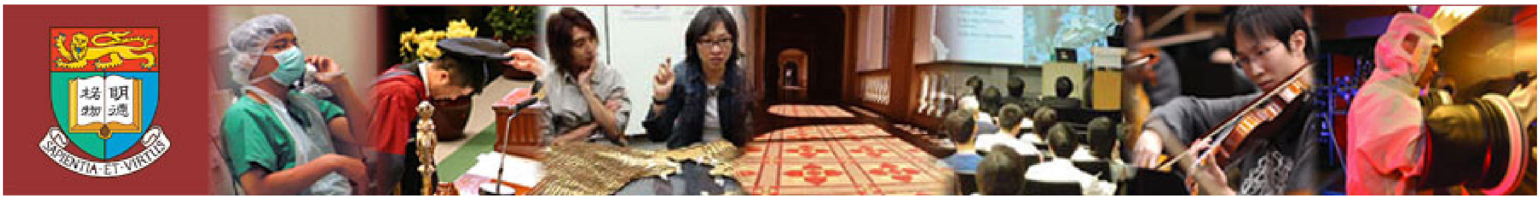


This paper was presented as part of the main technical program at IEEE INFOCOM 2011

Anonymous Communication with Network Coding against Traffic Analysis Attack

Jin Wang[1,2], Jianping Wang[1], Chuan Wu[3], Kejie Lu[4], Naijie Gu[2]
[1] Department of Computer Science, City University of Hong Kong
[2] Department of Computer Science, University of Science and Technology of China
[3] Department of Computer Science, University of Hong Kong
[4] Department of Electrical and Computer Engineering, University of Puerto Rico at Mayagüez

Abstract—**Flow untraceability is one critical requirement for anonymous communication with network coding, which prevents malicious attackers with wiretapping and traffic analysis abilities from relating the senders to the receivers, using linear dependency of the received packets. There have recently been proposals advocating encryptions on the Global Encoding Vectors (GEV) of network coding to thwart such attacks [1], [2]. Nevertheless, there has been no exploration of the capability of networking coding itself, to constitute more efficient and effective algorithms which guarantee anonymity. In this paper, we design a novel, simple, and effective linear network coding mechanism (*ALNCode*) to achieve flow untraceability in a communication network with multiple unicast flows. With solid theoretical analysis, we first show that linear network coding (LNC) *can* be applied to thwart traffic analysis attacks without the need of encrypting GEVs. Our key idea is to *mix* multiple flows at their intersection nodes by generating downstream GEVs from the common basis of upstream GEVs belonging to multiple flows, in order to hide the correlation of upstream and downstream GEVs in each flow. We then design a deterministic LNC scheme to implement our idea, by which the downstream GEVs produced are guaranteed to obfuscate their correlation with the corresponding upstream GEVs. We also give extensive theoretical analysis on the intersection probability of GEV bases and the influential factors to the effectiveness of our scheme, as well as the algorithm complexity to support its efficiency.**

I. Introduction

Following its success in maximizing network throughput [3]–[5], *network coding* has recently been shown to provide information security in a content distribution network, against active entropy and Byzantine modification attacks, as well as passive wiretapping attacks. To battle active attacks where malicious attackers can alter message content, secure network coding schemes have been proposed to guarantee *integrity* of the transmitted data in the network, based on the error correction and detection capabilities of network coding [6]–[8]. In face of passive wiretappers which may eavesdrop on the links to acquire transmitted information, network coding can naturally and effectively provide *confidentiality* of the messages as long as the attacker cannot obtain sufficient numbers of linearly independent coded messages [9]–[11].

In this paper, we consider how to use network coding to achieve *anonymity*, another key aspect of secure data communication, by which the identities of the sender and the receiver of a unicast flow, as well as the path traversed, are desirably hidden from wiretappers with traffic analysis abilities.

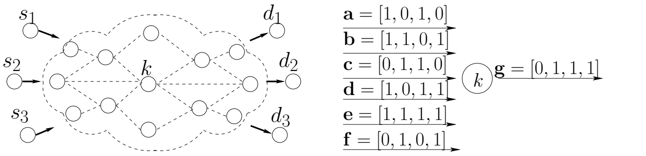
Flow untraceability has become increasingly important in today's Internet with more and more users resuming both roles of a content supplier and a content consumer. In online social networks such as Facebook and Twitter, friends exchanging private messages or contents may not want others to know; In a peer-to-peer (P2P) file sharing system, anonymous P2P communication is desired to share copyrighted files without revealing one's identity and risking litigation, as well as to prevent tracking or data mining activities from spammers.

To provide flow untraceability against traffic analysis attacks, traditional approaches without network coding choose from the following approaches [12]–[16]: (1) hide the content correlation among messages with encryption, (2) hide the size correlation of messages by padding with random symbols, (3) hide the time correlation among messages of the same flow by mixing the order of message transmissions from different flows at intermediate nodes, as well as (4) protect the routing information using secure routing protocols [15]–[17].

Linear network coding (LNC) has been promising to achieve the same objectives with much better efficiency: it naturally conceals content correlation of messages in the same flow, instead of using computationally expensive encryption and decryption for each message at each intermediate node; coded messages have an equal size and are buffered at intermediate nodes to generate new coded messages, naturally preventing correlating message sizes and arrival time patterns.

In LNC, each coded message in the network corresponds to a *Global Encoding Vector* (GEV), which consists of the coding coefficients it is produced with, with respect to the set of original messages. Given the encoding mechanisms of LNC, linear dependency among GEVs of coded messages may reveal information of the flow path, if the wiretapper analyzes the correlation between coded messages in the upstream and those in the downstream. Therefore, if a secure anonymous routing protocol is in place to hide the routing information, the key challenge of applying LNC for anonymous communication is to hide the correlation among GEVs.

A natural solution is to use encryption. Fan *et al.* [1] and Zhang *et al.* [2] propose to share a secret key between the source and the destination, apply an encryption function which allows intermediate nodes to produce new encrypted GEVs



(a) Network model (b) GEV generation at intermediate node k

Fig. 1. *ALNCode* design: an example.

without knowing the secret key, and then let the destination decrypt received GEVs with the pre-shared secret key. To implement such an encryption/decryption scheme, a Trust Authority (TA) or Key Distribution Center (KDC) is needed to distribute the secrete key to the source and the destination before the flow starts. In networks with large numbers of unicast flows, this approach may not scale well.

Can anonymous communication be provided in a more scalable, efficient, but effective fashion without intensive encryptions at all? This paper provides a "yes" answer, by exploring the power of LNC itself. We design a novel, simple, and effective Anonymous Linear Network Coding mechanism, referred to as *ALNCode*, which thwarts against traffic analysis attacks in a communication network with multiple unicast flows, without encrypting GEVs and sharing any secrete keys among sources and destinations. Our key idea is to *mix* multiple flows at their intersection nodes by generating downstream GEVs from the common basis of upstream GEVs belonging to multiple flows, in order to hide the correlation of upstream and downstream GEVs of each flow.

An example in Fig. 1 illustrates the idea: In a network with 3 unicast flows from source s_i to destination $d_i, i \in \{1,2,3\}$, respectively, each flow may go through multiple paths in the cloud and paths of different flows intersect at common intermediate nodes. LNC is performed at each node among messages of the same flow over finite field $\mathbb{F}_2$. Each flow has 4 original messages. Assume an intermediate node k receives 6 coded messages from 3 flows: two coded messages with GEVs $\mathbf{a} = [1,0,1,0]$ and $\mathbf{b} = [1,1,0,1]$ from flow $s_1 \to d_1$, two coded messages with GEVs $\mathbf{c} = [0,1,1,0]$ and $\mathbf{d} = [1,0,1,1]$ from flow $s_2 \to d_2$, and the rest two with GEVs $\mathbf{e} = [1,1,1,1]$ and $\mathbf{f} = [0,1,0,1]$ from flow $s_3 \to d_3$. k can generate a new coded message for flow $s_1 \to d_1$ by xor-ing the two received messages with GEVs $\mathbf{a}$ and $\mathbf{b}$, and derive the new GEV as $\mathbf{g} = \mathbf{a} + \mathbf{b} = [0,1,1,1]$. This new GEV is *obfuscated*, because $\mathbf{g} = \mathbf{a}+\mathbf{b} = \mathbf{a}+\mathbf{c}+\mathbf{d} = \mathbf{b}+\mathbf{e}+\mathbf{f} = \mathbf{c}+\mathbf{d}+\mathbf{e}+\mathbf{f}$, *i.e.*, $\mathbf{g}$ is not only correlated with $\{\mathbf{a},\mathbf{b}\}$, but also $\{\mathbf{a},\mathbf{c},\mathbf{d}\}$, $\{\mathbf{b},\mathbf{e},\mathbf{f}\}$, and $\{\mathbf{c},\mathbf{d},\mathbf{e},\mathbf{f}\}$. Therefore, to any traffic analysis attacker that tries to correlate the upstream and downstream GEVs, it would not be able to tell which messages belong to the same flow.

To the best of our knowledge, no previous work has aimed to preserve flow untraceability with such obfuscated network codes across multiple flows to hide GEV correlations. The main contributions of the paper are summarized as follows:

- ▷ We propose *ALNcode*, a light-weighted LNC mechanism for flow untraceability in networks with multiple unicast flows. We present solid theoretical analysis to support the validity of our idea.
- ▷ We design a deterministic LNC scheme to implement *ALNcode*, by which the downstream GEVs produced are guaranteed to obfuscate their correlation with the corresponding upstream GEVs, under mild conditions.
- ▷ We give extensive theoretical analysis on the intersection probability of the GEV bases and the influential factors to the effectiveness of our scheme, as well as the algorithm complexity to support its efficiency. We also discuss how our scheme can be practically applied to provide full-fledged flow untraceability, in terms of the anonymity of the source, the destination, as well as the paths of packets.

The rest of the paper is organized as follows. We formally present the models of network coding and wiretapping attacks, as well as our anonymous communication objectives in Sec. II. We present the key idea of *ALNcode* and the detailed deterministic LNC design in Sec. III. Sec. IV gives our extensive theoretical analysis on the effectiveness of our design and Sec. V discusses how our mechanism can efficiently throttle traffic analysis attacks. We discuss related work in Sec. VI and conclude the paper in Sec. VII.

II. Anonymous Communication Model with LNC

In this section, we present the network and LNC model, the traffic analysis attacks we consider, as well as the goals of anonymous communication.

A. Network and Linear Network Coding Model

We consider a communication network with multiple unicast flows between multiple pairs of *source* and *destination* nodes. Each flow has a unique flow number and may go through multiple paths; the paths of different flows may intersect at common intermediate nodes (*e.g.*, Fig. 1(a)). Linear network coding (LNC) [18] is applied in the transmission of each unicast flow: a source node partitions the data flow into *messages* of the fixed size H, and every h consecutive messages in the flow form a *generation*. LNC is performed among messages in the same generation of a flow.

Source encoding: Given original messages $\{\mathbf{m}_1, \cdots, \mathbf{m}_h\}$ in generation j of flow i, the source node selects h linearly independent GEVs, $\{\mathbf{v}_1, \cdots, \mathbf{v}_h\}$, over finite field $\mathbb{F}_q^h$, and generates h coded messages $\{\mathbf{m}'_1, \cdots, \mathbf{m}'_h\}$ using these GEVs. The h coded messages are generated as follows, shown together with the GEVs:

$$\left[\begin{array}{c|c} \mathbf{v}_n & \mathbf{m}'_n \end{array}\right] = \left[\begin{array}{c|c} \mathbf{v}_n & \sum_{l=1}^{h} v_{n,l}\mathbf{m}_n \end{array}\right], \quad (1)$$

where $1 \leq n \leq h$ and $v_{n,l}$ is the l_{th} element of vector $\mathbf{v}_n$.

Intermediate node encoding: Each intermediate node buffers coded messages received for a generation of a flow for T time slots, and produces new coded messages for this generation from the buffered messages. Suppose the node has received r coded messages $\{\mathbf{m}'_1, \cdots, \mathbf{m}'_r\}$ for generation j of flow i during time T, corresponding to r GEVs $\{\mathbf{v}'_1, \cdots, \mathbf{v}'_r\}$. To generate a new coded message, it produces a local encoding vector $\mathbf{c} = [c_1, \cdots, c_r]$ from finite field $\mathbb{F}_q^r$, and then generates the new coded message $\mathbf{m}''$ together with a new GEV $\mathbf{v}''$ as:

$$\left[\ \mathbf{v}'' \mid \mathbf{m}''\ \right] = \left[\ \sum_{l=1}^{r} c_l \mathbf{v}'_l \;\middle|\; \sum_{l=1}^{r} c_l \mathbf{m}'_l\ \right]. \tag{2}$$

Destination decoding: After receiving h coded messages $\{\mathbf{m}''_1, \cdots, \mathbf{m}''_h\}$ from generation j of flow i with linearly independent GEVs $\{\mathbf{v}''_1, \cdots, \mathbf{v}''_h\}$, a destination node recovers the original messages $\{\mathbf{m}_1, \cdots, \mathbf{m}_h\}$ by inverting the matrix composed by the GEVs:

$$\begin{bmatrix} \mathbf{m}_1 \\ \vdots \\ \mathbf{m}_h \end{bmatrix} = \begin{bmatrix} \mathbf{v}''_1 \\ \vdots \\ \mathbf{v}''_h \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{m}''_1 \\ \vdots \\ \mathbf{m}''_h \end{bmatrix}. \tag{3}$$

In a practical communication network, each coded message to be delivered is tagged with its routing information, flow number, generation number, and the GEV, which together is referred to as a *data packet*. We assume all data packets in the network have an equal size. We also assume that a secure, anonymous routing protocol [15]–[17] is in place (similar to the assumptions made in [1] and [2]). With such a secure, anonymous routing protocol, the route of each flow from the source to the destination is decided locally at each intermediate node, who knows only the previous-hop and next-hop nodes along the path. The routing information, flow and generation numbers attached to each coded message is protected by encryption with public/private keys generated locally by each intermediate node and exchanged only among neighbors. On the other hand, GEVs and message contents are not encrypted.

B. The Attack Model

We consider passive wiretapping attackers from outside of the network with traffic analysis abilities. For such an outside attacker, we assume it can observe all the packets along all the links in the network and analyze them, attempting to identify sources, destinations, and paths of the flows [1], [12]–[14]. For each attacker, routing, flow, and generation information in each data packet sniffed is hidden (by the secure, anonymous routing protocol), but GEVs and coded messages are open.

C. The Anonymous Communication Goals

The flow untraceability objectives we aim to achieve in this paper include:

- ▷ *Flow path anonymity.* The attacker cannot deduce the flow paths of each flow, *i.e.*, if an attacker observes an upstream packet and a downstream packet at a node, it cannot distinguish whether they are in the same flow or not.
- ▷ *Source and destination anonymity.* The attacker cannot determine which node each flow originates from or terminates at, *i.e.*, it is not able to tell which nodes in the network (sources) are communicating with which other nodes (destinations).

We summarize important notations in the paper for ease of reference in Table I.

TABLE I

Symbol	**Definition**
$\mathbb{F}_q$	a finite field of size q, where q is a prime number
h	the number of messages in each generation of a flow
H	the size of each message
$\mathbf{V}_{i,j,k}$	the set of GEVs received by node k from generation j of flow i in the past T time slots
$\widetilde{\mathbf{V}}_{i,j,k}$	the set of GEVs received by node k from flows other than i in the past T time slots
f_1	the number of GEVs received by node k from generation j of flow i in the past T time slots
f_2	the number of GEVs received by node k from flows other than $i$in the past T time slots
F	the total number of GEVs received by node k from all the flows and generations in the past T time slots
$L(\cdot)$	linear span of a set of vectors. For a matrix $\mathbf{B}$, $L(\mathbf{B})$ is the row vector space of $\mathbf{B}$.
Symbol $^{\mathrm{T}}$	transpose of a matrix or a vector
r_1	$dim(L(\mathbf{V}_{i,j,k}))$
r_2	$dim(L(\widetilde{\mathbf{V}}_{i,j,k}))$
R	$dim(L(\bigcup_{\forall i,j} \mathbf{V}_{i,j,k}))$
$\overline{\mathbf{C}}$	the matrix formed by nonzero vectors in the set of vectors $\mathbf{C}$ as its rows
$\mathbf{N}_{i,j,k}$	the basis of vector space $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$
N	$N = \lvert\mathbf{N}_{i,j,k}\rvert = dim(L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k}))$
$\mathbf{\Theta}_{i,j,k}$	the obfuscated basis of $L(\mathbf{V}_{i,j,k})$ which is the basis of $L(\mathbf{V}_{i,j,k})$ extended from $\mathbf{N}_{i,j,k}$

III. *ALNCode*: Anonymous Linear Network Coding against Traffic Analysis Attacks

We now present our Anonymous Linear Network Coding (*ALNCode*) mechanism to provide flow untraceability. We first present the general idea and then design a detailed deterministic LNC scheme to achieve the objective.

A. The Basic Idea

The key idea in *ALNCode* is to produce new coded messages with *obfuscated* GEVs at intermediate nodes, which are linearly correlated not only with received GEVs from the same flow, but also those from other flows. Fig. 2 gives an illustration: Let $\mathbf{V}_{i,j,k}$ denote the set of GEVs of coded messages received at intermediate node k from generation j of flow i in the past T time slots, and $L(\mathbf{V}_{i,j,k})$ be the linear span of these GEVs. Let $\widetilde{\mathbf{V}}_{i,j,k} = \bigcup_{l \neq i, \forall j} \mathbf{V}_{l,j,k}$ be the set of GEVs received at k from flows other than i before and within the T time slots, and $L(\widetilde{\mathbf{V}}_{i,j,k})$ be its linear span.

Suppose $\mathbf{N}_{i,j,k} = \{\mathbf{n}_1, \cdots, \mathbf{n}_{|\mathbf{N}_{i,j,k}|}\}$ denotes the basis of vector space $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$, $\mathbf{N}_{i,j,k}$ can be extended to the basis of vector space $L(\mathbf{V}_{i,j,k})$ (with methods described in Sec. III-B), *i.e.*, letting $r_1 = dim(L(\mathbf{V}_{i,j,k}))$, there exist $r_1 - |\mathbf{N}_{i,j,k}|$ vectors, $\{\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_{r_1 - |\mathbf{N}_{i,j,k}|}\}$, in $L(\mathbf{V}_{i,j,k})$, such that $\mathbf{\Theta}_{i,j,k} = \{\mathbf{n}_1, \cdots, \mathbf{n}_{|\mathbf{N}_{i,j,k}|}, \boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_{r_1 - |\mathbf{N}_{i,j,k}|}\}$ forms the basis of $L(\mathbf{V}_{i,j,k})$. We hereinafter refer to $\mathbf{\Theta}_{i,j,k}$ as the *obfuscated basis* of $L(\mathbf{V}_{i,j,k})$.

To produce a new coded message for generation j of flow i, we seek to generate a new GEV $\mathbf{v}_{i,j}$ that is linear combination of $\mathbf{a}$ and $\mathbf{b}$, where vector $\mathbf{a}$ is generated from $L(\mathbf{V}_{i,j,k})$ and vector $\mathbf{b}$ is generated from $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$. In this way, $\mathbf{v}_{i,j}$ *could* have linear correlation with both GEVs in $\mathbf{V}_{i,j,k}$ and $\widetilde{\mathbf{V}}_{i,j,k}$. If an attacker attempts to trace back the source of the coded packet with GEV $\mathbf{v}_{i,j}$, both the GEVs in $\mathbf{V}_{i,j,k}$ and

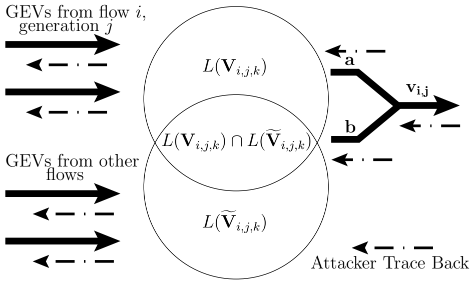

GEVs from flow i, generation j
GEVs from other flows
$L(\mathbf{V}_{i,j,k})$
$L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$
$L(\widetilde{\mathbf{V}}_{i,j,k})$
a
b
$\mathbf{v_{i,j}}$
Attacker Trace Back

Fig. 2. Obfuscated GEV production for generation j of flow i at intermediate node k.

$\mathbf{V}_{i,j,k}$ could be correlated with $\mathbf{v}_{i,j}$, and the attacker would fail to identify which flow the packet actually belongs to.

How do we generate $\mathbf{v}_{i,j}$ as the linear combination of vectors in $L(\mathbf{V}_{i,j,k})$ and $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$, respectively? Let $\overline{\mathbf{\Theta}}_{i,j,k}$ be the matrix formed by vectors in $\mathbf{\Theta}_{i,j,k}$ as its rows, and $\boldsymbol{\rho} = \{\rho_1, \cdots, \rho_{r_1}\}$ be a vector in $\mathbb{F}_q^{r_1}$. Set $\mathbf{v}_{i,j} = \boldsymbol{\rho}\overline{\mathbf{\Theta}}_{i,j,k} = \sum_{l=1}^{|\mathbf{N}_{i,j,k}|} \rho_l \mathbf{n}_l + \sum_{l=|\mathbf{N}_{i,j,k}|+1}^{r_1} \rho_l \boldsymbol{\delta}_{l-|\mathbf{N}_{i,j,k}|}$. Then $\mathbf{v}_{i,j}$ is linear combination of vector $\mathbf{b} = \sum_{l=1}^{|\mathbf{N}_{i,j,k}|} \rho_l \mathbf{n}_l$ from $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$ and vector $\mathbf{a} = \sum\limits_{l=|\mathbf{N}_{i,j,k}|+1}^{r_1} \rho_l \boldsymbol{\delta}_{l-|\mathbf{N}_{i,j,k}|}$ from $L(\mathbf{V}_{i,j,k})$.

Above we have provided one method to generate the GEV $\mathbf{v}_{i,j}$ which is potentially an obfuscated GEV, but not guaranteed. We next prove the sufficient and necessary condition that an obfuscated new GEV does exist, and the sufficient condition under which $\mathbf{v}_{i,j}$ produced above *is* an obfuscated GEV.

Theorem 1: At intermediate node k, an obfuscated GEV $\mathbf{u}_{i,j}$ exists for generation j of flow i, iff $dim(L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})) \neq 0$.

Proof: We first prove the necessity of the condition. If $\mathbf{u}_{i,j}$ is a GEV generated by node k for generation j of flow i, we have $\mathbf{u}_{i,j} \in L(\mathbf{V}_{i,j,k})$. If $\mathbf{u}_{i,j}$ is an obfuscated GEV, $\mathbf{u}_{i,j}$ is linearly correlated with both the GEVs in $L(\mathbf{V}_{i,j,k})$ and those in $L(\widetilde{\mathbf{V}}_{i,j,k})$, *i.e.*, there exist $\mathbf{a} \in L(\mathbf{V}_{i,j,k})$ and $\mathbf{b} \in L(\widetilde{\mathbf{V}}_{i,j,k})$ where $\mathbf{b} \neq \mathbf{0}$, such that $\mathbf{u}_{i,j} = \mathbf{a}+\mathbf{b}$. Therefore, $\mathbf{b} = \mathbf{u}_{i,j} - \mathbf{a} \in L(\mathbf{V}_{i,j,k})$. Since we also have $\mathbf{b} \in L(\widetilde{\mathbf{V}}_{i,j,k})$, $\mathbf{b} \in L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$. Thus, there exists this non-zero vector $\mathbf{b}$ in $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$, *i.e.*, $dim(L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})) \neq 0$.

We next prove the sufficiency of the condition. If $dim(L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})) \neq 0$, there exists a non-zero vector $\mathbf{b} \in L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$, and we can just set $\mathbf{u}_{i,j} = \mathbf{b}$, which is linearly correlated with both the GEVs in $L(\mathbf{V}_{i,j,k})$ and those in $L(\widetilde{\mathbf{V}}_{i,j,k})$). Thus the theorem is proved. ■

Theorem 2: When $dim(L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})) \neq 0$, a GEV $\mathbf{v}_{i,j}$ generated by node k for generation j of flow i using the above mentioned method, is an obfuscated GEV, if not all the first $|\mathbf{N}_{i,j,k}|$ elements of $\boldsymbol{\rho}$ are zero.

Proof: Using the above mentioned method, we know

$$\mathbf{v}_{i,j} = \sum_{l=1}^{|\mathbf{N}_{i,j,k}|} \rho_l \mathbf{n}_l + \sum_{l=|\mathbf{N}_{i,j,k}|+1}^{r_1} \rho_l \boldsymbol{\delta}_{l-|\mathbf{N}_{i,j,k}|}.$$

Since each $\mathbf{n}_l, 1 \leq l \leq |\mathbf{N}_{i,j,k}|$, is a vector in $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$, $\sum_{l=1}^{|\mathbf{N}_{i,j,k}|} \rho_l \mathbf{n}_l$ is a linear combination of the vectors in $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$. Since not all the first $|\mathbf{N}_{i,j,k}|$ elements of $\boldsymbol{\rho}$ are zero and vectors $\mathbf{n}_1, \ldots, \mathbf{n}_{|\mathbf{N}_{i,j,k}|}$ are linearly independent, $\sum_{l_1=1}^{|\mathbf{N}_{i,j,k}|} \rho_{l_1} \mathbf{n}_{l_1}$ is a non-zero vector. On the other hand, $\sum_{l=|\mathbf{N}_{i,j,k}|+1}^{r_1} \rho_l \boldsymbol{\delta}_{l-|\mathbf{N}_{i,j,k}|}$ is a linear combination of the vectors in $L(\mathbf{V}_{i,j,k})$. Thus, GEV $\mathbf{v}_{i,j}$ has linear correlation with both GEVs in $L(\mathbf{V}_{i,j,k})$ and those in $L(\widetilde{\mathbf{V}}_{i,j,k})$). ■

B. The Deterministic Linear Network Coding Scheme

Based on Theorems 1 and 2, we now design a detailed LNC scheme, by which r_1 new coded messages with linearly independent obfuscated GEVs can be generated at each intermediate node k, after it receives r_1 linearly independent GEVs in the past T time slots from generation j of flow i, as long as $dim(L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})) \neq 0$ is satisfied.

The basic method proposed in the previous section shows how to generate obfuscated GEVs: We obtain $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$ from the received GEVs at node k from generation j of flow i and other flows. We derive $\mathbf{N}_{i,j,k}$, *i.e.*, its basis, and extend it to $\mathbf{\Theta}_{i,j,k}$, *i.e.*, the basis of vector space $L(\mathbf{V}_{i,j,k})$. We then select vectors $\boldsymbol{\rho} \in \mathbb{F}_q^{r_1}$ with the first $|\mathbf{N}_{i,j,k}|$ elements not all zero and generate GEVs $\mathbf{v}_{i,j} = \boldsymbol{\rho}\overline{\mathbf{\Theta}}_{i,j,k}$. We know these GEVs are obfuscated GEVs (Theorem 2), but we need to select different $\boldsymbol{\rho}$'s, such that r_1 linearly independent obfuscated GEVs can be produced. We next detail these procedures, as well as how local encoding vectors are formed at k to generate these new GEVs from received GEVs.

1) Derive $\mathbf{\Theta}_{i,j,k}$

Let $\mathbf{\Lambda} = \{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_{r_1}\}$ be the maximum independent set of $\mathbf{V}_{i,j,k}$. $\mathbf{\Lambda}$ is the basis of $L(\mathbf{V}_{i,j,k})$. Let $\overline{\mathbf{\Lambda}}$ be the matrix formed by vectors in $\mathbf{\Lambda}$ as its rows. Let $\mathbf{\Gamma} = \{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{r_2}\}$ be the maximum independent set of $\widetilde{\mathbf{V}}_{i,j,k}$. $\mathbf{\Gamma}$ is the basis of $L(\widetilde{\mathbf{V}}_{i,j,k})$. We compute the basis of $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$ from $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ and extend it to the basis of $L(\mathbf{V}_{i,j,k})$ following the general method [19] to get the basis of the intersection of two vector spaces and extend it to the basis of one vector space:

i) Construct a matrix $\mathbf{A} = \{\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_{r_1}^T, \boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_{r_2}^T\}$ with dimension $h \times (r_1 + r_2)$ and then reduce $\mathbf{A}$ to its row-echelon form $rref(\mathbf{A})$ by Gaussian elimination. Note that if a row of $rref(\mathbf{A})$ is non-zero, the first non-zero element of this row is refereed to as the *pivot* of the row. A *non-pivotal column* refers to a column no elements of which is a pivot.

ii) Let N be the number of non-pivotal columns of $rref(\mathbf{A})$. Then $N = dim(L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k}))$ [19]. We can obtain N linear combinations $\sum\limits_{l=1}^{r_1} a_{n,l}\boldsymbol{\alpha}_l, 1 \leq n \leq N$, where $[a_{n,1}, \cdots, a_{n,r_1}, a_{n,r_1+1}, \cdots, a_{n,h}]^T$ is the n_{th} non-pivotal column of $rref(\mathbf{A})$. These linear combinations form the basis of $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$, *i.e.*, $\mathbf{N}_{i,j,k}$, and $N = |\mathbf{N}_{i,j,k}|$.

iii) To derive $\mathbf{\Theta}_{i,j,k}$, a basis of $L(\mathbf{V}_{i,j,k})$ which contains $\mathbf{N}_{i,j,k}$, we first construct a $h \times (N + r_1)$ dimensional matrix

$$\mathbf{\Phi} = \left[\begin{array}{cccccc} \sum\limits_{l=1}^{r_1} a_{1,l}\boldsymbol{\alpha}_l^T, & \cdots, & \sum\limits_{l=1}^{r_1} a_{N,l}\boldsymbol{\alpha}_l^T, & \boldsymbol{\alpha}_1^T, & \cdots, & \boldsymbol{\alpha}_{r_1}^T \end{array} \right].$$

We know the basis of the column space of $\mathbf{\Phi}$ can be derived as follows: reduce $\mathbf{\Phi}$ to its row-echelon form $rref(\mathbf{\Phi})$, and then those column vectors in $\mathbf{\Phi}$, that correspond to the columns in $rref(\mathbf{\Phi})$ containing pivots, form the basis. The column space of $\mathbf{\Phi}$ is indeed $L(\mathbf{N}_{i,j,k} \cup \mathbf{V}_{i,j,k}) = L(\mathbf{V}_{i,j,k})$, and thus we have derived a basis of $L(\mathbf{V}_{i,j,k})$. In addition, since the set of vectors in $\mathbf{N}_{i,j,k}$ are linearly independent, all the column vectors in the $\mathbf{N}_{i,j,k}$ part of $\mathbf{\Phi}$ correspond to columns in $rref(\mathbf{\Phi})$ containing pivots. Thus, the basis of $L(\mathbf{V}_{i,j,k})$ derived above is composed of all the vectors in $\mathbf{N}_{i,j,k}$, as well as $r_1 - N$ other GEVs in $\mathbf{V}_{i,j,k}$, which we denote as $\{\boldsymbol{\alpha}_{L_1}, \cdots, \boldsymbol{\alpha}_{L_{r_1-N}}\}$. $\mathbf{\Theta}_{i,j,k}$, the basis of $L(\mathbf{V}_{i,j,k})$ which contains $\mathbf{N}_{i,j,k}$, is thus derived as

$$\{\sum_{l=1}^{r_1} a_{1,l}\boldsymbol{\alpha}_l, \cdots, \sum_{l=1}^{r_1} a_{N,l}\boldsymbol{\alpha}_l, \boldsymbol{\alpha}_{L_1}, \cdots, \boldsymbol{\alpha}_{L_{r_1-N}}\}. \tag{4}$$

2) Generate r_1 linearly independent obfuscated GEVs

The vectors in $\mathbf{\Theta}_{i,j,k}$ form the basis of $L(\mathbf{V}_{i,j,k})$ and the first N vectors in $\mathbf{\Theta}_{i,j,k}$ are the basis of $L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})$. In general, to produce r_1 linearly independent obfuscated GEVs, we left-multiply $\mathbf{\Theta}_{i,j,k}$ by a nonsingular matrix composed by r_1 linearly independent vectors from $\mathbb{F}_q^{r_1}$, the first N elements of each of which are not all zeros. We can select a nonsingular lower triangular matrix $\mathbf{C}_1$ as follows:

$$\mathbf{C}_1 = \begin{bmatrix} c_{1,1} & & & \mathbf{0} \\ c_{2,1} & c_{2,2} & & \\ \vdots & \vdots & \ddots & \\ c_{r_1,1} & c_{r_1,2} & \cdots & c_{r_1,r_1} \end{bmatrix},$$

where each $c_{i',j'}, 1 \le j' \le i' \le r_1$ is randomly selected from $\{1 \cdots q-1\}$. Since $dim(L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})) \neq 0$ and the leading element of each row of $\mathbf{C}_1$ is non-zero, each row vector of $\mathbf{C}_1\overline{\mathbf{\Theta}}_{i,j,k}$ is an obfuscated GEV; since matrix $\mathbf{C}_1$ has full rank, these r_1 obfuscated GEVs are linearly independent.

3) Construct local encoding vectors

Recall the intermediate node encoding model described in Sec. II-A: after receiving coded messages corresponding to r_1 linearly independent GEVs $\mathbf{\Lambda} = \{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_{r_1}\}$, node k selects r_1 coding coefficients from $\mathbb{F}_q^{r_1}$. Then according to this local encoding vector, it does a linear combination of the r_1 received coded messages to produce a coded message, and also does a linear combination of the received GEVs to produce the new GEV for the message. We can derive the local encoding vectors to produce the r_1 linearly independent obfuscated GEVs, *i.e.*, the rows of $\mathbf{C}_1\overline{\mathbf{\Theta}}_{i,j,k}$, as follows.

Let $\mathbf{\Omega}$ denote the $r_1 \times r_1$ dimensional local encoding matrix, whose rows are the local encoding vectors. It should satisfy $\mathbf{\Omega}\overline{\mathbf{\Lambda}} = \mathbf{C}_1\overline{\mathbf{\Theta}}_{i,j,k}$. Since the matrix $\overline{\mathbf{\Theta}}_{i,j,k}$ is formed by the obfuscate basis as its rows, it can be represented as $\overline{\mathbf{\Theta}}_{i,j,k} = \mathbf{C}_2\overline{\mathbf{\Lambda}}$, where $\mathbf{C}_2$ is a $r_1 \times r_1$ dimensional matrix as follows:

$$\mathbf{C}_2 = \left[\begin{array}{ccc} a_{1,1} & \cdots & a_{1,r_1} \\ \vdots & \vdots & \vdots \\ a_{N,1} & \cdots & a_{N,r_1} \\ \multicolumn{3}{c}{\mathbf{I}_{r_1,L_1}} \\ \multicolumn{3}{c}{\vdots} \\ \multicolumn{3}{c}{\mathbf{I}_{r_1,L_{r_1-N}}} \end{array}\right] \quad \text{(according to Eq. (4))},$$

where $\mathbf{I}_{r_1,L_n}$ is the L_nth row of a $r_1 \times r_1$ identity matrix.

Algorithm 1 Local Encoding Matrix Computing

```
Find the maximum independent set of $\mathbf{V}_{i,j,k}$ and $\widetilde{\mathbf{V}}_{i,j,k}$ by Gaussian elimination which are $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_{r_1}\}$ and $\{\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{r_2}\}$ respectively.
Construct a matrix $\mathbf{A} = \{\boldsymbol{\alpha}_1^T, \cdots, \boldsymbol{\alpha}_{r_1}^T, \boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_{r_2}^T\}$ with dimension $h \times (r_1 + r_2)$.
Compute row-echelon form matrix $rref(\mathbf{A})$ by Gaussian elimination. Let the number of non-pivotal columns of $rref(\mathbf{A})$ be $N$.
for $n$ from 1 to $N$ do
    $\boldsymbol{\theta}_n = \sum_{l=1}^{r_1} a_{n,l}\boldsymbol{\alpha}_l$ where $[a_{n,1}, \cdots, a_{n,h}]^T$ is the $n_{th}$ non-pivotal column of $rref(\mathbf{A})$. The $n_{th}$ row of $\mathbf{C}_2$ is set to $[a_{n,1}, \cdots, a_{n,r_1}]$.
end for
$\mathbf{N}_{i,j,k} = \bigcup_{n=1}^{N} \boldsymbol{\theta}_n$.
Find $r_1 - N$ vectors in $\mathbf{V}_{i,j,k}$, $\{\boldsymbol{\alpha}_{L_1}, \cdots, \boldsymbol{\alpha}_{L_{r_1-N}}\}$, such that GEVs in set $\{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_N, \boldsymbol{\alpha}_{L_1}, \cdots, \boldsymbol{\alpha}_{L_{r_1-N}}\}$ are linearly independent.
for $n$ from 1 to $r_1 - N$ do
    $\boldsymbol{\theta}_{N+n} = \boldsymbol{\alpha}_{L_n}$. The $(N+n)_{th}$ row of $\mathbf{C}_2$ is set to $\mathbf{I}_{r_1,L_n}$.
end for
The obfuscated basis of $L(\mathbf{V}_{i,j,k})$ is $\mathbf{\Theta}_{i,j,k} = \{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_{r_1}\}$.
for $l$ from 1 to $r_1$ do
    select $l$ numbers from $\{1, \cdots, q-1\}$ as the first $l$ elements of the $l_{th}$ row of matrix $\mathbf{C}_1$. The remaining $r_1 - l$ elements are set to 0.
end for
return local encoding matrix $\mathbf{\Omega} = \mathbf{C}_1\mathbf{C}_2$.
```

Since $\mathbf{\Omega}\overline{\mathbf{\Lambda}} = \mathbf{C}_1\overline{\mathbf{\Theta}}_{i,j,k} = \mathbf{C}_1\mathbf{C}_2\overline{\mathbf{\Lambda}}$, we derive $\mathbf{\Omega} = \mathbf{C}_1\mathbf{C}_2$, *i.e.*, each row of $\mathbf{C}_1\mathbf{C}_2$ is a local encoding vector, which node k should use to generate r_1 independent obfuscated GEVs.

Let $\{\mathbf{m}'_1, \cdots, \mathbf{m}'_{r_1}\}$ denote the r_1 received coded messages corresponding to the r_1 linearly independent GEVs in $\mathbf{\Lambda}$, and $\mathbf{M}'$ be the matrix formed by these coded messages as its rows. The r_1 linearly independent obfuscated new GEVs $\{\mathbf{v}''_1, \cdots, \mathbf{v}''_{r_1}\}$ and the corresponding new coded messages $\{\mathbf{m}''_1, \cdots, \mathbf{m}''_{r_1}\}$ can be calculated as follows:

$$\left[\begin{array}{c|c} \mathbf{v}''_1 & \mathbf{m}''_1 \\ \vdots & \vdots \\ \mathbf{v}''_{r_1} & \mathbf{m}''_{r_1} \end{array}\right] = \mathbf{\Omega} \left[\begin{array}{c|c} \boldsymbol{\alpha}_1 & \mathbf{m}'_1 \\ \vdots & \vdots \\ \boldsymbol{\alpha}_{r_1} & \mathbf{m}'_{r_1} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{\Omega}\overline{\mathbf{\Lambda}} & \mathbf{\Omega}\mathbf{M}' \end{array}\right].$$

Algorithm 1 shows to calculate the local encoding matrix. Let f_1 be the number of GEVs node k received from generation j of flow i in the past T time slots and f_2 be the number of GEVs from other flows. Since the computational complexity of Gaussian elimination applied to a $m \times n$ dimensional matrix is $O(mn\min(m,n))$ and that of matrix multiplication between a $m \times n$ dimensional matrix and a $n \times l$ dimensional matrix is $O(mnl)$, the computational complexity of Algorithm 1 is $O(h^2(f_1 + f_2))$. To further calculate r_1 new coded messages, the total computational complexity is $O(h^2(f_1 + f_2) + r_1^2 H)$. Due to space limit, we omit the details.

At the source and destination. Our previous discussions have been focusing on recoding at intermediate nodes to hide relationship of its incoming and outgoing packets. We next show that with a similar scheme, the source and destination nodes of a flow can also hide themselves, as long as there are other flows going through them.

A source node s of flow i can produce coded messages with any GEVs from $\mathbb{F}_q^h$. If s also receives other flows, it can produce

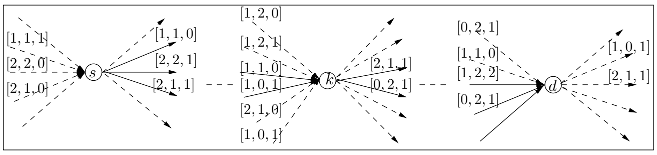


Fig. 3. *ALNCode* at the source node and the destination node.

obfuscated GEVs for generation j of flow i, using vectors from $L(\widetilde{\mathbf{V}}_{i,j,s})$. This is because vectors in $L(\widetilde{\mathbf{V}}_{i,j,s})$ are linearly correlated with the received GEVs from other flows, and they are valid GEVs to produce coded messages at the source for generation j of flow i. At the destination node d of flow i, if some other flows go through it, it can generate obfuscated GEVs for these other flows, exploiting the received GEVs from flow i. In both cases, the anonymity of source and destination nodes are protected, and an attacker can not distinguish whether coded messages originate from or terminate at a node.

Fig. 3 gives an example, where solid directed lines denote packets of generation j of flow i and dotted directed lines denote packets of other flows. Let $h = 3$ and LNC is performed over $\mathbb{F}_3$. At the source of flow i, $\mathbf{a} = [1, 1, 1], \mathbf{b} = [2, 2, 0], \mathbf{c} = [2, 1, 0]$ are incoming GEVs from other flows. The source can generate obfuscated GEVs such as $\mathbf{d} = 2\mathbf{b} = [1, 1, 0], \mathbf{e} = \mathbf{a} + 2\mathbf{b} = [2, 2, 1]$, and $\mathbf{f} = \mathbf{a} + \mathbf{b} + \mathbf{c} = [2, 1, 1]$. At the destination, $\mathbf{o} = [1, 2, 2], \mathbf{p} = [0, 2, 1]$ are incoming GEVs from generation j of flow i and $\mathbf{m} = [0, 2, 1], \mathbf{n} = [1, 1, 0]$ are from generation j' of flow i'. The destination can generate obfuscated GEVs for flow i', such as $\mathbf{g} = \mathbf{m}+\mathbf{n} = \mathbf{o}+2\mathbf{p} = [1, 0, 1], \mathbf{h} = \mathbf{m} + 2\mathbf{n} = 2\mathbf{o} = [2, 1, 1]$.

IV. Effectiveness Analysis of *ALNCode*

Using *ALNCode*, a node k which receives multiple flows can effectively produce new coded messages for each outgoing flow, with obfuscated GEVs against traffic analysis attacks, as long as the condition of $dim(L(\mathbf{V}_{i,j,k}) \cap L(\widetilde{\mathbf{V}}_{i,j,k})) \neq 0$ is satisfied. But does this condition hold with high probability in practice? The answer to this question is crucial to the practical effectiveness of our scheme, which we seek in the following.

In this section, we use L_1 to denote $L(\mathbf{V}_{i,j,k})$ and L_2 to denote $L(\widetilde{\mathbf{V}}_{i,j,k})$, respectively, for simplified references. We assume that the GEVs received by node k are randomly and independently selected from finite field $\mathbb{F}_q^h$. In this case, L_1 is a span space of f_1 vectors randomly selected from $\mathbb{F}_q^h$ and L_2 is a span space of f_2 vectors randomly selected from $\mathbb{F}_q^h$ (Recall that f_1 and f_2 are the numbers of GEVs received at k from generation j of flow i and from other flows, respectively).

A. The Intersection Probability

We prove the lower bound of the probability $dim(L_1 \cap L_2) \neq 0$, for any $f_1 \geq 0, f_2 \geq 0$ (which is referred to as *the intersection probability*), in Theorem 3.

Let $\begin{bmatrix} m \\ r \end{bmatrix}_q$ be the Gaussian binomial coefficient, *i.e.*, $\begin{bmatrix} m \\ r \end{bmatrix}_q = \frac{(q^m-1)(q^{m-1}-1)\cdots(q^{m-r+1}-1)}{(q-1)(q^2-1)\cdots(q^r-1)}, 0 < r \leq m$. We set $\begin{bmatrix} m \\ 0 \end{bmatrix}_q = 1, \forall m > 0$. We first prove two lemmas.

Lemma 1: For any $m \times n$ dimensional matrix $\mathbf{B}$ whose elements are randomly selected from finite field $\mathbb{F}_q$, the probability that $rank(\mathbf{B}) = r, 0 \leq r \leq min(m, n)$ is given by: $p_1(m, n, r, q) = \begin{bmatrix} m \\ r \end{bmatrix}_q \prod_{l=n-r+1}^{n} (q^l - 1)q^{\frac{r(r-1)}{2} - mn}$.

Proof: From Theorem 1.10 in [20], the number of $m \times n$ dimensional matrices with rank $r, 0 \leq r \leq min(m, n)$ is: $\begin{bmatrix} m \\ r \end{bmatrix}_q \prod_{l=n-r+1}^{n} (q^l - 1)q^{\frac{r(r-1)}{2}}$. Since the total number of $m \times n$ dimensional matrices is q^{mn}, the probability that a $m \times n$ dimensional matrix has a rank of r is $\frac{\begin{bmatrix} m \\ r \end{bmatrix}_q \prod_{l=n-r+1}^{n} (q^l-1)q^{\frac{r(r-1)}{2}}}{q^{mn}} = \begin{bmatrix} m \\ r \end{bmatrix}_q \prod_{l=n-r+1}^{n} (q^l - 1)q^{\frac{r(r-1)}{2} - mn}$. ■

From Lemma 1, we have $p_1(m, n, r, q) = p_1(n, m, r, q)$.

Lemma 2: Given a vector space L_1 with $dim(L_1) = r(r \geq 0)$, if L_2 is a vector space spanned by f_2 vectors randomly selected from $\mathbb{F}_q^h$, the probability that $dim(L_1 \cap L_2) \neq 0$ is:

$$p_2(r, f_2, h, q) \geq 1 - \sum_{g=0}^{\min(f_2, h-r)} p_1(f_2, h - r, g, q)q^{(g-f_2)r}.$$

Proof: Given $dim(L_1) = r$, let the basis of L_1 be $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$. Suppose the basis of $\mathbb{F}_q^h$ extended by $\{\mathbf{v}_1, \cdots, \mathbf{v}_r\}$ is $\mathbf{V} = \{\mathbf{v}_1, \cdots, \mathbf{v}_h\}$. Let $\overline{\mathbf{V}}$ be the matrix formed by vectors in $\mathbf{V}$ as its rows. Since matrix $\overline{\mathbf{V}}$ is formed by the basis of $\mathbb{F}_q^h$, for any vector $\mathbf{v}_0$ in $\mathbb{F}_q^h$, there exists a unique vector $\boldsymbol{\alpha}$ in $\mathbb{F}_q^h$, such that $\mathbf{v}_0 = \boldsymbol{\alpha}\overline{\mathbf{V}}$.

A $f_2 \times h$ dimensional matrix $\mathbf{B}$ is formed by f_2 vectors randomly selected in $\mathbb{F}_q^h$ as its rows. We can express $\mathbf{B}$ as

$$\left[\begin{array}{ccc|ccc} \alpha_{1,1} & \cdots & \alpha_{1,r} & \alpha_{1,r+1} & \cdots & \alpha_{1,h} \\ \alpha_{2,1} & \cdots & \alpha_{2,r} & \alpha_{2,r+1} & \cdots & \alpha_{2,h} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha_{f_2,1} & \cdots & \alpha_{f_2,r} & \alpha_{f_2,r+1} & \cdots & \alpha_{f_2,h} \end{array}\right] \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_h \end{bmatrix}$$
$$= [\ \Lambda_1 \mid \Lambda_2\ ]\overline{\mathbf{V}}, \tag{5}$$

where Λ_1 denotes a $f_2 \times r$ dimensional matrix and Λ_2 is a $f_2 \times (h - r)$ dimensional matrix. Since the first r rows of $\overline{\mathbf{V}}$ compose the basis of L_1, the intersection of L_1 and the row space of matrix $\mathbf{B}$ (*i.e.*, L_2) is nonzero, iff there exists a vector $[b_1, \cdots, b_r][\overline{\mathbf{V}}]_1^r$ in the row space of matrix $\mathbf{B}$, where $[b_1, \cdots, b_r] \neq \mathbf{0}$ and $[\overline{\mathbf{V}}]_1^r$ denotes a matrix formed by the set of rows in $\overline{\mathbf{V}}$ with indices from 1 to r. This is equivalent to that $dim(L_1 \cap L_2) \neq 0$, iff the row space of matrix $\begin{bmatrix} \Lambda_1 & \Lambda_2 \end{bmatrix}$ includes a nonzero vector $[b_1, \cdots, b_r, 0, \cdots, 0]$ of length h, based on Eq. (5).

Since the f_2 row vectors of $\mathbf{B}$ are randomly selected from $\mathbb{F}_q^h$, the row vectors of matrix $\begin{bmatrix} \Lambda_1 & \Lambda_2 \end{bmatrix}$ are randomly distributed in $\mathbb{F}_q^h$. Consequently, the probability of $dim(L_1 \cap L_2) \neq 0$ equals to the probability that the row space of matrix $\begin{bmatrix} \Lambda_1 & \Lambda_2 \end{bmatrix}$, with randomly selected elements from

$\mathbb{F}_q^h$, includes a nonzero vector $(b_1, \cdots, b_r, 0, \cdots, 0)$ of length h. The latter further equals to the probability that the row space of the matrix $\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right]$ includes a nonzero vector $[0, \cdots, 0, b_1, \cdots, b_r]$ of length h.

By Gaussian elimination, we can derive the row-echelon form of matrix $\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right]$, $rref(\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right])$, which has $rank(\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right])$ non-zero rows. The column vectors in $\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right]$, those corresponding to columns in $rref(\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right])$ containing pivots, form the basis of the column space of $\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right]$. If the column space of Λ_1 (*i.e.*, $L(\Lambda_1^T)$) is not a subspace of that of Λ_2 (*i.e.*, $L(\Lambda_2^T)$), then the basis of the column space of $\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right]$ must contain at least one column vector from Λ_1. Therefore, there exists an index l where $h - r < l \leq h$, such that the l_{th} column vector in $rref(\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right])$ contains a pivot and the corresponding row vector in $rref(\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right])$ with that pivot takes the form of $[b'_1, \cdots, b'_{l-1}, b'_l, \cdots, b'_h]$ with $b'_n = 0$ for $0 \leq n \leq l - 1$ and $b'_l \neq 0$. Since Gaussian elimination equals to a series of elementary row operations, we derive that the row space of $\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right]$ contains the nonzero vector $[b'_1, \cdots, b'_{l-1}, b'_l, \cdots, b'_h]$ with $b'_n = 0$ for $0 \leq n \leq l - 1$ and $b'_l \neq 0$. Therefore, if $L(\Lambda_1^T) \nsubseteq L(\Lambda_2^T)$, then the row space of matrix $\left[\begin{array}{cc}\Lambda_2 & \Lambda_1\end{array}\right]$ includes a nonzero vector of length h in the form of $[0, \cdots, 0, b_1, \cdots, b_r]$. We thus know that the probability of $dim(L_1 \cap L_2) \neq 0$ is lower bounded by the probability of $L(\Lambda_1^T) \nsubseteq L(\Lambda_2^T)$, *i.e.*,

$$\begin{aligned} p_2(r, f_2, h, q) &\geq P(L(\Lambda_1^T) \nsubseteq L(\Lambda_2^T)) \\ &= 1 - P(L(\Lambda_1^T) \subseteq L(\Lambda_2^T)). \end{aligned} \tag{6}$$

Since the column vectors of Λ_1 and Λ_2 are randomly and independently selected from $\mathbb{F}_q^{f_2}$, we derive $P(L(\Lambda_1^T) \subseteq L(\Lambda_2^T))$:

$$\sum_{g=0}^{\min(f_2, h-r)} P(rank(\Lambda_2^T) = g) P(L(\Lambda_1^T) \subseteq L(\Lambda_2^T) | rank(\Lambda_2^T) = g). \tag{7}$$

If $rank(\Lambda_2^T) = g$, then the number of vectors in $L(\Lambda_2^T)$ is q^g. Therefore, when a vector is randomly selected from $\mathbb{F}_q^{f_2}$, the probability that it is in $L(\Lambda_2^T)$ is q^{g-f_2}. Given a vector space $L(\Lambda_2^T)$, we have $L(\Lambda_1^T) \subseteq L(\Lambda_2^T)$ iff every row vectors in Λ_1^T is in $L(\Lambda_2^T)$. Since the number of row vectors in Λ_1^T is r, the probability that every row vectors in Λ_1^T is in $L(\Lambda_2^T)$ is $q^{(g-f_2)r}$, *i.e.*,

$$P(L(\Lambda_1^T) \subseteq L(\Lambda_2^T) | rank(\Lambda_2^T) = g) = q^{(g-f_2)r}.$$

From Lemma 1, $P(rank(\Lambda_2^T) = g) = p_1(h - r, f_2, g, q) = p_1(f_2, h - r, g, q)$. Therefore, Eq. (7) equals to

$$\sum_{g=0}^{\min(f_2, h-r)} p_1(f_2, h - r, g, q)\left(\frac{q^g}{q^{f_2}}\right)^r.$$

Thus, $p_2(r, f_2, h, q) \geq 1 - \sum_{g=0}^{\min(f_2, h-r)} p_1(f_2, h - r, g, q) q^{(g-f_2)r}$. ■

Now we prove the lower bound of the probability that $dim(L_1 \cap L_2 \neq 0)$ when L_1 is a span space of f_1 vectors randomly selected from $\mathbb{F}_q^h$ and L_2 is a span space of f_2 vectors randomly selected from $\mathbb{F}_q^h$.

Theorem 3: Let L_1 and L_2 be span spaces of f_1 and f_2 vectors randomly selected from $\mathbb{F}_q^h$, respectively, for any $h, f_1, f_2 \geq 0$. The probability $dim(L_1 \cap L_2) \neq 0$ satisfies:

$$\begin{aligned} &P(dim(L_1 \cap L_2) \neq 0) \\ \geq &\sum_{r=0}^{\min(f_1, h)} \left(\left[\begin{array}{c} f_1 \\ r \end{array}\right]_q \prod_{l=h-r+1}^{h} (q^l - 1) q^{\frac{r(r-1)}{2} - f_1 h} \right) \times \\ &\left(1 - \sum_{g=0}^{\min(f_2, h-r)} \left(\left[\begin{array}{c} f_2 \\ g \end{array}\right]_q \prod_{l=h-r-g+1}^{h-r} (q^l - 1) q^{\frac{g(g-1)}{2} - f_2 h + gr} \right) \right). \end{aligned}$$

Proof: From Lemma 1, the probability that $dim(L_1) = r$ is $p_1(f_1, h, r, q)$. From Lemma 2, when $dim(L_1) = r$, the probability that $dim(L_1 \cap L_2) \neq 0$ is $p_2(r, f_2, h, q)$. Therefore,

$$\begin{aligned} P(dim(L_1 \cap L_2) \neq 0) &= \sum_{r=0}^{\min(f_1, h)} p_1(f_1, h, r, q) p_2(r, f_2, h, q) \\ &\geq p_3(f_1, f_2, h, q), \end{aligned}$$

where the lower bound $p_3(f_1, f_2, h, q)$ is defined to be the term in the right-hand side of the inequality in the theorem. ■

B. The Influential Parameters

To provide a better idea of the intersection probability $P(dim(L_1 \cap L_2) \neq 0)$ with its deciding parameters, we show the lower bound derived in Theorem 3 at different values of f_1, f_2, h, and q in Fig. 4. Fig. 4 (a) and (b) show that the probability increases with the increase of f_1 and f_2, respectively. [1] The reason is straightforward: when h and q are fixed, the more GEVs a node receives in the current flow and in other flows, the larger probability the two vector spaces L_1 and L_2 have nonzero intersection.

Fig. 4(c) shows that the probability decreases with the increase of h, while Fig. 4(d) demonstrates different trends with the increase of q in different cases. In particular, for $\forall f_1, f_2 > 0$, we show below with analysis that: when $f_1 + f_2 \leq h$, this lower bound probability decreases with the increase of q and h; when $f_1 + f_2 > h$, it increases with the increase of q and decreases with the increase of h.

If $f_1 + f_2 \leq h$, from Lemma 1, the probability that the $(f_1 + f_2) \times h$ dimensional matrix formed by the received GEVs as its rows has full rank, is higher with larger q and h. When the $(f_1 + f_2) \times h$ dimensional matrix has full rank, the $f_1 + f_2$ GEVs are linearly independent (since $f_1 + f_2 \leq h$), *i.e.*, $dim(L_1 \cap L_2) = 0$. Thus, the intersection probability decreases with the increase of q and h.

If $f_1 + f_2 > h$, we use $\mathbf{V}_1$ to denote $\mathbf{V}_{i,j,k}$ and $\mathbf{V}_2$ to denote $\widetilde{\mathbf{V}}_{i,j,k}$. Let $\overline{\mathbf{V}}_3 = \left[\begin{array}{c} \overline{\mathbf{V}}_1 \\ \overline{\mathbf{V}}_2 \end{array}\right]$. The analysis is shown below [21]:

$$\begin{aligned} &dim(L_1 \cap L_2) = dim(L(\mathbf{V}_1) \cap L(\mathbf{V}_2)) \\ &= dim(L(\mathbf{V}_1)) + dim(L(\mathbf{V}_2)) - dim(L(\mathbf{V}_1 \cup \mathbf{V}_2)) \\ &= rank(\overline{\mathbf{V}}_1) + rank(\overline{\mathbf{V}}_2) - rank(\overline{\mathbf{V}}_3). \end{aligned}$$

[1] Though not shown in the plots, we note that: in the case of Fig. 4(a), the probability approaches 1 when $f_1 > 10$; in the case of Fig. 4(b), the probability goes to zero when $f_2 < 5$ and approaches 1 when $f_2 > 15$.

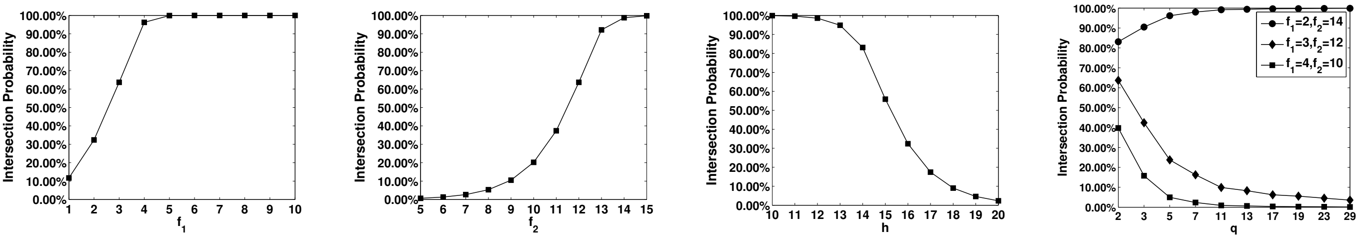


Fig. 4. Lower bound of intersection probability $P(dim(L_1 \cap L_2)) \neq 0$, when (a) $f_2 = 12, q = 2, h = 15$; (b) $f_1 = 3, q = 2, h = 15$; (c) $f_1 = 3, f_2 = 12, q = 2$; (d) $h = 15$.

If the f_1 GEVs received from generation j of flow i are linearly independent, and the f_2 GEVs received from other flows are linearly independent, then $rank(\overline{\mathbf{V}}_1) = \min(f_1, h)$, $rank(\overline{\mathbf{V}}_2) = \min(f_2, h)$, and $rank(\overline{\mathbf{V}}_3) \leq h$. Then

$$\begin{aligned} dim(L_1 \cap L_2) &= rank(\overline{\mathbf{V}}_1) + rank(\overline{\mathbf{V}}_2) - rank(\overline{\mathbf{V}}_3) \\ &\geq \min(f_1, h) + \min(f_2, h) - h \\ &> 0. \text{ (since } f_1 + f_2 > h \text{ and } f_1, f_2 > 0) \end{aligned} \tag{8}$$

Therefore, if the probabilities that $\mathbf{V}_1$ and $\mathbf{V}_2$ have full ranks increase, the probability that $dim(L(\mathbf{V}_1) \cap L(\mathbf{V}_2)) > 0$ increases. From Lemma 1, the former probabilities increase with the increase of q and the decrease of h.

The above results can guide the practical selection of field size (q), the number of messages per generation (h), and the number of messages to buffer before recoding (f_1) in *ALNCode*, given the routes of flows decided by the routing protocol (which determines f_2). In general, an intermediate node may buffer sufficient number of messages with linearly independent GEVs in generation j of flow i to produce different new coded messages. When dividing a flow into generations, a reasonably small h should be chosen to guarantee a good intersection probability, as well as low decoding complexity. The finite field size q can then be set accordingly: if many linearly independent GEVs can be received at each node such that $f_1 + f_2 > h$, a relatively large q can be used, but not too large considering the communication overhead and decoding complexity; if few GEVs can be received, we can simply select $q = 2$ for the best intersection probability.

V. Discussions on Anonymity Against Attacks

We now discuss how the proposed *ALNCode* can practically provide anonymity against traffic analysis attacks.

A traffic analysis attacker may try to identify the source, destination, and the path of a flow by analyzing the correlation among coded packets it observes. In a coded packet, the routing, flow, and generation information are protected by the secure routing protocol, and the attacker can only try to identify the correlation among GEVs along the links.

We show the computational complexity for GEV correlation analysis at each node the attacker eavesdrops on is very high in *ALNCode*. Let the input GEVs at a node be $\{\mathbf{v}_1, \cdots, \mathbf{v}_F\}$ and an output GEV is $\mathbf{v}$. To find the correlation between $\mathbf{v}$ and all the input GEVs, the attacker needs to solve a system of linear equations $\sum_{l=1}^{F} x_l \mathbf{v}_l = \mathbf{v}$, with F variables, $x_1, \ldots, x_F$, and h equations. If $dim(L(\bigcup_{l=1}^{F} \mathbf{v}_l)) = R$, $F - R$ variables are designated as free, *i.e.*, they can take any value from $\mathbb{F}_q$, while the remaining variables are dependent on these free variables. The dimension of the solution set is $F - R$. To solve those equations, the attacker may first use Gaussian elimination to find the $F - R$ free variables and the linear dependence among the R remaining variables and the free variables. The computational complexity is $O(Fh \min(F, h)))$. After values of the $F - R$ free variables are given, the R remaining variables can be calculated within $O(R(F - R))$ time. Since the number of different value combinations of the $F - R$ free variables is q^{F-R}, the computational complexity to find the solution set is $O(q^{F-R}(Fh))$. Therefore, the overall computational complexity to analyze the linear correlation between one output GEV and all input GEVs is $O(Fh \min(F, h)) + q^{F-R}R(F - R))$. Note that $R \leq h$. If the total number of GEVs received by each node from different flows, F, is sufficiently large, the computational complexity grows to infinite.

Even if the attacker obtains the correlation among upstream and down stream GEVs at a node, the computational complexity to trace the flow back (or forth) to the source (or destination) grows exponentially, given that each output GEV in *ALNCode* is correlated with multiple sets of input GEVs with high probability. In a network with a large number of nodes and flows, it is almost computationally impossible for an attacker to correctly identify the source, destination, and paths of a flow. Therefore, *ALNCode* is efficient to defend against traffic analysis attacks and protect the anonymity of the source, the destination, and the paths of each flow.

In addition, confidentiality of the message content can also be protected in *ALNCode* from any outside attacker, since the flow and generation information in each coded packet is hidden by the secure, anonymous routing protocol. Even if an outside attacker can acquire all the packets transmitted in the network, it does not know which flow and generation coded packets belong to, and thus cannot group packets belonging to the same generation of a flow to decode the original messages.

VI. Related Work

Network coding has been widely explored in recent years, to achieve the maximum throughput of a network [3]–[5], as well as to provide information security in a content distribution network against active modification attacks and passive wiretapping attacks [6]–[11], [18]. With respect to defense against wiretapping attacks, the main focus has been on exploring the capability of network coding to provide *confidentiality* of the message content [9]–[11]. Few efforts have been devoted to utilizing network coding on communication anonymity.

The concept of anonymous communication between sources and destinations was first introduced by Chaum *et al.* [12]. Among all attack models against anonymity, traffic analysis attack is a major one. There exist a number of approaches on defending anonymity against traffic analysis attack in traditional networks without network coding, with three representative ones: *the Crowds approach*, *the onion routing approach*, and *the Mix approach*.

Crowds [22] provides a centralized service to randomly select participants of a network into a group (the "crowd"), which includes the source. Each message is routed through the crowd before it is sent to the destination node, such that the attacker cannot tell which node in the crowd is the original source. In the onion routing approach [13], the source establishes a path to the destination through a number of nodes called *onion routers*, and encrypts the routing information and message repeatedly with public keys of the onion routers, in order to prevent any attacker from learning the path information. With the Mix approach [14], [15], instead of forwarding each message as it arrives, an intermediate node, *i.e.*, the *Mix* node, waits for a random period of time and then forwards messages it received in mixed order, so as to hide the time correlation among messages of the same flow. These existing approaches either require centralized services, which is not scalable, or demands encryption of whole messages, which is computationally expensive.

Among the few proposals which utilize LNC for anonymous communication, we have discussed the work by Fan *et al.* [1] and Zhang *et al.* [2] in Sec. I. Both approaches require centralized key distributed services for encryption and decryption of the GEVs at source and destination, which limits the scalability of the system. In contrast, we have shown that *ALNcode* provides flow anonymity with low computational complexity in a fully distributed manner.

VII. Conclusion

This paper explores the power of linear network coding to provide flow untraceability against traffic analysis attackers in networks with multiple unicast flows. An effective LNC mechanism, *ALNCode*, is proposed, that protects anonymity of source, destination, and paths of each flow with a simple but novel idea: nodes in the network mix the flow correlation between downstream and upstream packets by generating GEVs for outgoing coded messages from the common basis of incoming GEVs belonging to multiple flows. A deterministic LNC scheme is discussed in details that implements the idea. Other highlights of our contributions include the solid and extensive theoretical analysis on the existence condition of obfuscated GEVs, the intersection probability of GEV bases, and the complexity of our scheme, as well as abundant discussions on practical settings of influential parameters for the best effectiveness of our scheme.

Acknowledgement

The work is supported in part by State Key Laboratory of Advanced Optical Communication Systems and Networks, P. R. China under grant 2009SH01.

References

[1] Y. Fan, Y. Jiang, H. Zhu, and X. Shen, "An efficient Privacy-Preserving scheme against traffic analysis attacks in network coding," in *Proceedings of IEEE INFOCOM 2009, the 28th Conference on Computer Communications*, Rio De Janeiro, Brazil, 2009, pp. 2213–2221.

[2] P. Zhang, Y. Jiang, C. Lin, Y. Fan, and X. Shen, "P-Coding: secure network coding against eavesdropping attacks," in *Proc. of IEEE INFOCOM 2010, the 29th Conference on Computer Communications*, 2010.

[3] R. Ahlswede, N. Cai, S. Li, and R. Yeung, "Network information flow," *IEEE Transactions on Information Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[4] S. Li, R. Yeung, and N. Cai, "Linear network coding," *IEEE Transactions on Information Theory*, vol. 49, no. 2, pp. 371–381, 2003.

[5] Z. Li, B. Li, and L. C. Lau, "On achieving maximum multicast throughput in undirected networks," *IEEE/ACM Transaction on Networking*, vol. 14, no. SI, pp. 2467–2485, 2006.

[6] N. Cai and R. W. Yeung, "Network coding and error correction," in *Proc. of the 2002 IEEE Inform. Theory Workshop*, Bangalore, India, Oct. 2002, pp. 119–122.

[7] Z. Yu, Y. Wei, B. Ramkumar, and Y. Guan, "An efficient Signature-Based scheme for securing network coding against pollution attacks," in *Proc. of IEEE INFOCOM 2008, the 27th Conference on Computer Communications*, 2008, pp. 1409–1417.

[8] E. Kehdi and B. Li, "Null keys: Limiting malicious attacks via null space properties of network coding," in *Proc. of the 26th Conference on Computer Communications (INFOCOM)*, 2009, pp. 1224–1232.

[9] N. Cai and R. Yeung, "Secure network coding," in *Proc. of IEEE International Symposium on Information Theory 2002*, 2002, p. 323.

[10] 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.

[11] J. Wang, J. Wang, K. Lu, B. Xiao, and N. Gu, "Optimal linear network coding design for secure unicast with multiple streams," in *Proc. of IEEE INFOCOM 2010, the 29th Conference on Computer Communications*, San Diego, CA USA, 2010.

[12] D. Chaum, C. O. T. Acm, R. Rivest, and D. L. Chaum, "Untraceable electronic mail, return addresses, and digital pseudonyms," *Communications of the ACM*, vol. 24, pp. 84—88, 1981.

[13] M. G. Reed, P. F. Syverson, and D. M. Goldschlag, "Anonymous connections and onion routing," *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 482—494, 1998.

[14] G. Danezis, R. Dingledine, D. Hopwood, and N. Mathewson, "Mixminion: Design of a type III anonymous remailer protocol," in *Proc. of the Workshop on Privacy in the Electronic Society*, 2003, pp. 2–15.

[15] M. Rennhard, "Introducing MorphMix: Peer-to-Peer based anonymous internet usage with collusion detection," in *Proc. of the workshop on privacy in the electronic society(WPES2002)*, 2002, pp. 91—102.

[16] T. Isdal, M. Piatek, A. Krishnamurthy, and T. Anderson, "Privacy-preserving p2p data sharing with oneswarm," University of Washington, Technical Report, 2009.

[17] X. Lin, R. Lu, Z. Huafei, P. Ho, X. Shen, and Z. Cao, "ASRPAKE: an anonymous secure routing protocol with authenticated key exchange for wireless ad hoc networks," in *Proc. of IEEE International Conference on Communications 2007*, Glasgow, Scotland, 2007, pp. 1247–1253.

[18] C. Gkantsidis and P. R. Rodriguez, "Cooperative security for network coding file distribution," in *Proc. of the 25th IEEE International Conference on Computer Communications (INFOCOM)*, 2006, pp. 1–13.

[19] K. Yang, "A basis for the intersection of subspaces," *Mathematics Magazine*, vol. 70, no. 4, p. 297, Oct. 1997.

[20] Z. Wan, *Geometry of Classical Groups over Finite Fields*, 2nd ed. Beijing, P.R.China: Science Press, 2002.

[21] K. M. Hoffman and R. Kunze, *Linear Algebra*, 2nd ed. Prentice Hall, Apr. 1971.

[22] M. K. Reiter and A. D. Rubin, "Crowds: anonymity for web transactions," *ACM Transactions on Information and System Security (TISSEC)*, vol. 1, no. 1, pp. 66–92, 1998.