PANDA: A protocol-assisted network decoding algorithm

Claudio Greco, Michel Kieffer, Cédric Adjih, Beatrice Pesquet-Popescu

PANDA: A PROTOCOL-ASSISTED NETWORK DECODING ALGORITHM

Claudio Greco[1,2,3], Michel Kieffer[1], Cedric Adjih[2], Beatrice Pesquet-Popescu[3]

[1] *L2S, CNRS-Supélec-Univ Paris-Sud, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France*

[2] *INRIA Rocquencourt, HIPERCOM Project Team, Domaine de Voluceau, BP 105, 78153 Le Chesnay, France*

[3] *Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI 46 rue Barrault, 75634 Paris Cedex 13, France*

ABSTRACT

With random linear network coding, mixed packets contain in their headers information about the coding operations performed on the original packets to allow their recovery by the receiver. This introduces an overhead that could be significant if the packet size is relatively small w.r.t. the size of the generation.

In this paper, we propose to remove the part of the added header related to the network coding coefficients and to consider network decoding as a source separation problem. This problem is addressed using a maximum *a posteriori* estimation technique. It exploits some *a priori* information related to the content of the headers added to the *original* packets by the upper layers of the protocol stack, *before* network coding.

Experiments show that, despite the fact that traditional source separation techniques are completely inadequate to handle this scenario, the proposed approach is able to recover all packets within streams of thousands of generations without a single decoding error.[1]

***Index Terms*—Blind Source Separation; Joint Source-Channel Decoding; Network Coding.**

I. INTRODUCTION

To cope with dynamic network topologies, random linear Network Coding (NC) performs packet mixing with randomly-generated coefficients within some finite field [1]. These coefficients need to be transmitted along with the packets to allow network decoding at the destination [2]. The resulting overhead may become significant, especially when the packets are comparatively small w.r.t. the size of the generation.

The aim of this paper is to show that one may avoid the transmission of NC coefficients, while being able to recover the original packets with a vanishing error probability. If the packets have the same length, *i.e.*, if no padding is needed, this amounts to introducing no overhead (additional symbols might be needed otherwise). For that purpose, we account for the fact that the data packets generated by sources are usually encapsulated by upper layers of the protocol stacks, *before* NC. Headers introduced by these layers contain enough redundancy to perform *protocol-assisted network decoding*, *i.e.*, to infer the NC operations applied on the received packets by observing the combined headers, and to recover the original packets by inverting these operations.

Several attempts have been made to reduce the overhead of a NC transmission without significantly affecting the decoding probability. For instance, Jafari *et al.* [3] have proposed to employ shortened coding vectors to efficiently convey the coding coefficients. They exploit the fact that, in some networks, usually not *all* source packets are combined. This approach can indeed reduce the overhead, but only if the number of packets in each combination is much smaller than the size of the generation. Thomos *et al.* [4] proposed to generate the coding coefficients such that they can be described by using just one symbol per packet. In particular, the coding vectors are generated like rows of a modified Vandermonde matrix. This approach can drastically reduce the overhead at the price of an increased probability of generating identical rows, thus making the encoding matrix singular and non-decodable.

The problem considered in this paper is reminiscent of the reconstruction of a source matrix $\mathbf{X}$ from a set of measurements consisting of unknown linear combinations of the rows of $\mathbf{X}$, which can be seen as a *Blind Source Separation* (BSS) problem. Generally speaking, BSS [5] consists in recovering G source vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_G$, defined over a field (usually $\mathbb{R}$ or $\mathbb{C}$) and organized in a matrix $\mathbf{X}$, from a set of linear combinations $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_G$, (collected in a matrix $\mathbf{Y}$) of the source vectors.

One popular approach to BSS is *Independent Component Analysis* (ICA) [5, 6], which assumes statistical independence of the sources. ICA estimates $\mathbf{X}$ as the set of vectors in the span of $\mathbf{Y}$ that minimizes the joint entropy and the mutual similarity among vectors. Recently, BSS has also been considered in a finite field $\mathbb{F}_q$ by minimizing the marginal entropy of the estimates through an exhaustive exploration of the row span of $\mathbf{Y}$ [7, 8].

The idea of BSS in finite field has been exploited in [9] to recover randomly network-coded packets without including the combination coefficients in the transmitted packets.

[1] This work has been partly supported by DIM-LSC SWAN.

However, this technique does not rely exclusively on entropy to perform the separation, as purely entropy-based methods strive to deliver good separation properties when the sources have a distribution close to uniform, which is typically the case for flows of compressed data. Thus, [9] includes a form of channel coding of the source vectors to inject an identifiable feature. This allows to reduce the search space considerably by rejecting solutions that do not carry a valid channel code. The main difference with the present work is that we exploit the overhead inherent to the upper layers of the transmission protocol.

This type of redundancy has been shown to be helpful at the receiver to get more reliable packet header recovery, channel decoding, or frame synchronization [10]. This work exploits the same redundancy in the very different context of BSS of randomly network coded data.

The rest of this paper is organized as follows. In Section II, network decoding is formulated as a maximum *a posteriori* estimation problem. Section III describes the way the structure imposed by the protocol layers on the message headers may be exploited. Section IV presents a decoding algorithm that implements the maximization problem efficiently. In Section V, typical examples of header fields, commonly found in communication standards, are identified. For three of these classes we evaluate the reduction of the search space provided by the class. A comparison of our approach with a reference technique is provided in Section VI. Finally, Section VII draws some conclusions and outlines some future work.

II. PROBLEM FORMULATION

Let $\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_G$ be a generation of G encapsulated source messages. Each message consists of some protocol header followed by a compressed payload. The messages are represented by vectors of L symbols in $\mathbb{F}_q$ and organized in a $G \times L$ source matrix $\mathbf{X}$, realization of a random matrix $\boldsymbol{X}$, assumed of rank G. This assumption is reasonable in practice, since usually $L \gg G$ and the data compression makes it unlikely for the set of messages to be linearly dependent. Assume also that the receiver collects $G' \geqslant G$ random linear combinations of the source messages from which a set of G linearly independent mixed packets $\mathbf{y}_1, \mathbf{y}_2 \dots \mathbf{y}_G$ can be selected and organized in a $G \times L$ matrix $\mathbf{Y}$.

Under these hypotheses, it is always possible to describe the network coding operations that mapped $\mathbf{X}$ into $\mathbf{Y}$ with a full-rank $G \times G$ coding matrix $\mathbf{A}$ such that

$$\mathbf{Y} = \mathbf{A}\mathbf{X} \tag{1}$$

Our goal is to estimate the encoding operations described by $\mathbf{A}$ given only the *a priori* information on the structure of the encoding matrix and the source messages, and the received mixed packets $\mathbf{Y}$. This can be formulated as a MAP estimation problem as follows

$$\widehat{\mathbf{A}} \;=\; \arg \max_{\mathbf{A} \in \mathbb{F}_q^{G \times G}} \Pr\{\boldsymbol{A} = \mathbf{A} \mid \boldsymbol{Y} = \mathbf{Y}\}, \tag{2}$$

where $\boldsymbol{A}$ and $\boldsymbol{Y}$ are random matrices representing the network coding operation and the received packets respectively. In the following, the random variables will be omitted when no ambiguity arises, *e.g.*, $\Pr\{\boldsymbol{A}{=}\mathbf{A} \mid \boldsymbol{Y}{=}\mathbf{Y}\}$ will be denoted as $\Pr\{\mathbf{A} \mid \mathbf{Y}\}$. This can be rewritten as

$$\begin{aligned} \Pr\{\mathbf{A} \mid \mathbf{Y}\} &= \sum_{\mathbf{X} \in \mathbb{F}_q^{G \times L}} \Pr\{\mathbf{A},\mathbf{X} \mid \mathbf{Y}\} \\ &= \sum_{\mathbf{X} \in \mathbb{F}_q^{G \times L}} \Pr\{\mathbf{Y} \mid \mathbf{A},\mathbf{X}\} \Pr\{\mathbf{A},\mathbf{X}\} \\ &= \sum_{\mathbf{X} \in \mathbb{F}_q^{G \times L}} \delta[\mathbf{Y}{-}\mathbf{A}\mathbf{X}] \Pr\{\mathbf{A}\} \Pr\{\mathbf{X}\} \end{aligned} \tag{3}$$

Where $\delta[\cdot]$ denotes the Kronecker delta function, defined as

$$\delta[\mathbf{X}] = \begin{cases} 1 & \text{if } \mathbf{X} = 0, \\ 0 & \text{otherwise} \end{cases}$$

Since $\mathbf{A}$ is full rank, $\delta[\mathbf{Y}{-}\mathbf{A}\mathbf{X}] \neq 0$, that is, $\mathbf{Y} - \mathbf{A}\mathbf{X} = 0$, for the unique value $\mathbf{X} = \mathbf{A}^{-1}\mathbf{Y}$. Therefore, (3) becomes

$$\Pr\{\mathbf{A} \mid \mathbf{Y}\} = f_{\boldsymbol{A}}(\mathbf{A})\, f_{\boldsymbol{X}}\left(\mathbf{A}^{-1}\mathbf{Y}\right), \tag{4}$$

where $f_{\boldsymbol{A}}(\cdot)$ and $f_{\boldsymbol{X}}(\cdot)$ are the *a priori* probability mass functions (pmf) of $\boldsymbol{A}$ and $\boldsymbol{X}$ respectively. Since $\mathbf{A}$ is of full rank, $f_{\boldsymbol{A}}$ is non-zero only on the subset $\mathcal{A} \subset \mathbb{F}_q^{G \times G}$ of full-rank matrices of size G. Thus, (2) becomes

$$\widehat{\mathbf{A}} = \arg \max_{\mathbf{A} \in \mathcal{A}} f_{\boldsymbol{A}}(\mathbf{A})\, f_{\boldsymbol{X}}\left(\mathbf{A}^{-1}\mathbf{Y}\right) \tag{5}$$

This estimation problem may be solved by searching the inverse of the encoding matrix, *i.e.*, the decoding matrix, $\mathbf{W}{=}\mathbf{A}^{-1}$. If $\mathcal{A}$ and $f_{\boldsymbol{A}}(\mathbf{A})$ are known, then

$$\mathcal{W} = \left\{\mathbf{W} \in \mathbb{F}_q^{G \times G} \mid \mathbf{W}^{-1} \in \mathcal{A}\right\} = \mathcal{A} \tag{6}$$

and $f_{\boldsymbol{W}}(\mathbf{W}){=}f_{\boldsymbol{A}}(\mathbf{W}^{-1})$. Then, (5) can be transformed in the equivalent problem of determining

$$\widehat{\mathbf{W}} = \arg \max_{\mathbf{W} \in \mathcal{A}} f_{\boldsymbol{W}}(\mathbf{W})\, f_{\boldsymbol{X}}(\mathbf{W}\mathbf{Y}) \tag{7}$$

Solving directly (7) has a worst-case complexity $O\left(q^{G \times G}\right)$, which is intractable, even for relatively small values of G. The following sections will show how the *a priori* information coming from the structure of the headers of the source messages may be exploited to break down this complexity.

III. STRUCTURE OF UNCODED PACKET HEADERS

At transmitter side, each layer of a protocol stack usually adds some header to the packets it receives from the upper layer [11]. This encapsulation facilitates packet processing at receiver side, but may lead to a significant redundancy, as recognized in [10].

We assume that headers of source messages consists of fields, *i.e.*, set of bits belonging to one of the following classes. *Constant*: fields of this class consist of bits of

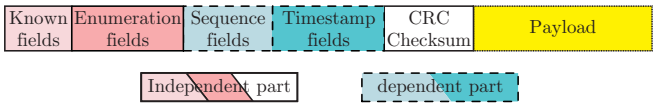

Fig. 1. Structure of a generic header.

constant and *known* value during the transmission, *e.g.*, protocol version.

Enumeration: the values taken by the bits of fields of this class belongs to a *finite* and *known* set of values, *e.g.*, port number, content type.

Sequence: bits of such field represent an integer increasing by a *predictable* amount for each packet, *e.g.*, sequence number.

Timestamp: bits of such fields represent a number increasing from packet to packet according to some known distribution, *e.g.*, packet timestamp, due-date.

Control: bits of such field are used to detect errors in other parts of packet.

Other: gathers all other bits which are not exploited in the proposed estimation technique.

The main idea of the proposed Protocol-Assisted Network Decoding Algorithm (PANDA) is to build first the set of *admissible* candidate rows of $\mathbf{W}$. Such rows should lead to estimated messages (network-decoded packets) with headers *compliant* with the syntax imposed by the protocol encapsulation: each field of the message headers has to comply with the constraints of the class it belongs to. Second, a combinatorial search is performed among the compliant candidate rows to find the most probable matrix $\mathbf{W}$ such that the headers of the G estimated messages of a generation jointly satisfy *all* constraints imposed by the protocol.

For that purpose, each message $\mathbf{x}_i$ is partitioned into an *independent part* $\mathbf{x}_i^{(\mathrm{I})}$ and a *dependent part* $\mathbf{x}_i^{(\mathrm{D})}$. This partition is assumed to be the same for all $\mathbf{x}_i$s. The independent part of a message gathers all fields which compliance with the protocol can be verified independently of the other messages. The dependent part gathers the fields which consistency has to be verified on all network decoded messages.

The values of $\mathbf{x}_i^{(\mathrm{I})}$ are assumed statistically independent from $\mathbf{x}_j^{(\mathrm{I})}$, $j \neq i$, and the dependent and independent parts of the messages are assumed mutually independent. Thus

$$\Pr\{\mathbf{x}_1 \dots \mathbf{x}_G\} = \Pr\{\mathbf{x}_1^{(\mathrm{D})} \dots \mathbf{x}_G^{(\mathrm{D})}\} \Pr\{\mathbf{x}_1^{(\mathrm{I})}\} \cdots \Pr\{\mathbf{x}_G^{(\mathrm{I})}\}. \quad (8)$$

To extract the different fields of the messages, consider the $L \times N_{\mathrm{F}}$ *projection matrix* $\mathbf{\Pi}_{\mathrm{F}} = \left(\mathbf{e}_{i_1}^\top, \mathbf{e}_{i_2}^\top \dots \mathbf{e}_{i_{N_{\mathrm{F}}}}^\top\right)$, where $\mathbf{e}_i$ is the i-th row of the $L \times L$ identity matrix $\mathbf{I}_L$, and $i_1, i_2, \dots, i_{N_{\mathrm{F}}}$ are the indices of the symbols in $\mathbb{F}_q$ of the field F of the message. Therefore, once the indices of the dependent and independent parts of $\mathbf{X}$ have been identified and provided that all fields are symbol-aligned, $\mathbf{x}_i^{(\mathrm{I})} = \mathbf{x}_i \mathbf{\Pi}_{\mathrm{I}}$ and $\mathbf{x}_i^{(\mathrm{D})} = \mathbf{x}_i \mathbf{\Pi}_{\mathrm{D}}$.

Using this notation, we define $P_{\boldsymbol{X}}^{(\mathrm{D})}(\mathbf{X})$ as $\Pr\{\boldsymbol{X}\mathbf{\Pi}_{\mathrm{D}} = \mathbf{X}\mathbf{\Pi}_{\mathrm{D}}\}$ and $p_{\boldsymbol{x}_i}^{(\mathrm{I})}(\mathbf{x}_i)$ as $\Pr\{\boldsymbol{X}_i\mathbf{\Pi}_{\mathrm{I}} = \mathbf{x}_i\mathbf{\Pi}_{\mathrm{D}}\}$. If the $\boldsymbol{X}_i\mathbf{\Pi}_{\mathrm{I}}$ are also identically distributed, there exists a pmf $p_{\boldsymbol{X}}^{(\mathrm{I})}(\cdot)$ such that $p_{\boldsymbol{x}_i}^{(\mathrm{I})}(\cdot) = p_{\boldsymbol{X}}^{(\mathrm{I})}(\cdot), \forall i$.

Due to the constraints imposed on the various fields of the message headers, many values of $\mathbf{X}$ are not compliant. In other words, $P_{\boldsymbol{X}}^{(\mathrm{D})}(\cdot)$ and $p_{\boldsymbol{X}}^{(\mathrm{I})}(\cdot)$ may vanish for many randomly-chosen $\mathbf{X}$. Since, in order to have $f_{\boldsymbol{X}}(\mathbf{X}) > 0$, each $\boldsymbol{X}_i\mathbf{\Pi}_{\mathrm{P}}$ must have non-zero probability for all parts P, *i.e.*, in order for a matrix $\mathbf{X}$ to be compliant, each of the parts of its rows has to be compliant, studying $p_{\boldsymbol{X}}^{(\mathrm{I})}$ allows us to eliminate many candidates rows $\mathbf{w}$, thus reducing considerably the number of candidate decoding matrices.

For that purpose, consider the sets

$$\mathcal{C}_{\boldsymbol{X}}^{(\mathrm{I})} = \left\{\mathbf{x} \in \mathbb{F}_q^L \mid p_{\boldsymbol{X}}^{(\mathrm{I})}(\mathbf{x}) > 0\right\} \quad (9)$$

of all messages compliant with the constraints on the independent parts and

$$\mathcal{C}_{\boldsymbol{X}}^{(\mathrm{D})} = \left\{\mathbf{X} \in \mathbb{F}_q^{G \times L} \mid P_{\boldsymbol{X}}^{(\mathrm{D})}(\mathbf{X}) > 0\right\}. \quad (10)$$

of all matrices of G messages compliant with the constraints on de dependent parts.

These sets are used to derive the set of compliant candidate row vectors

$$\mathcal{W}_{\mathbf{y}}^{(\mathrm{I})} = \left\{\mathbf{w} \in \mathbb{F}_q^G \mid \mathbf{w}\mathbf{Y} \in \mathcal{C}_{\boldsymbol{X}}^{(\mathrm{I})}\right\}, \quad (11)$$

and the set of candidate decoding matrices

$$\mathcal{W}_{\mathbf{Y}}^{(\mathrm{D})} = \left\{\mathbf{W} \in \mathcal{W} \mid \mathbf{W}\mathbf{Y} \in \mathcal{C}_{\boldsymbol{X}}^{(\mathrm{D})}\right\} \quad (12)$$

The true decoding matrix belongs to $\mathcal{W}_{\mathbf{Y}}^{(\mathrm{D})}$ and has rows taken from $\mathcal{W}_{\mathbf{y}}^{(\mathrm{I})}$. The search for $\mathbf{W}$ is thus done in

$$\mathcal{W}_{\mathbf{Y}} = \mathcal{W}_{\mathbf{Y}}^{(\mathrm{D})} \cap \bigotimes_{i=1}^{G} \mathcal{W}_{\mathbf{y}}^{(\mathrm{I})}, \quad (13)$$

where $\bigotimes_{i=1}^{G} \mathcal{W}_{\mathbf{y}}^{(\mathrm{I})}$ is the set of all $G \times G$ matrices with G rows taken from $\mathcal{W}_{\mathbf{y}}^{(\mathrm{I})}$.

The estimation problem (7) can then be rewritten as

$$\widehat{\mathbf{W}} = \arg \max_{\mathbf{W} \in \mathcal{W}_{\mathbf{Y}}} f_{\boldsymbol{W}}(\mathbf{W}) P_{\boldsymbol{X}}^{(\mathrm{D})}(\mathbf{W}\mathbf{Y}) \prod_{i=1}^{G} p_{\boldsymbol{X}}^{(\mathrm{I})}(\mathbf{w}_i \mathbf{Y}) \quad (14)$$

where $\mathbf{w}_i$ is the i-th row of $\mathbf{W}$.

IV. PANDA

PANDA, the proposed network decoding algorithm, is summarized in Algorithm IV. It takes as input a $G \times L$ matrix $\mathbf{Y}$ satisfying the hypotheses introduced in Section II.

In the first loop (Line 6 to 12), the independent set of constraints is used to build the set of compliant decoding rows $\mathcal{W}_{\mathbf{y}}^{(\mathrm{I})}$.

At Line 20, the set $\mathcal{W}_{\mathbf{y}}^{G}$ of $G \times G$ matrices of full rank is generated from $\mathcal{W}_{\mathbf{y}}^{(\mathrm{I})}$. Since $\mathcal{W}_{\mathbf{Y}}^{(\mathrm{D})} \subseteq \mathcal{W}$, from (13), one knows that $\mathcal{W}_{\mathbf{Y}} \subseteq \mathcal{W}_{\mathbf{Y}}^{(\mathrm{D})} \subseteq \mathcal{W}_{\mathbf{y}}^{G}$.

Thus, in the second loop (Line 14 to 20), the set of compliant matrices $\mathcal{W}_{\mathbf{Y}}$ is build, selecting from $\mathcal{W}_{\mathbf{y}}^{G}$ all matrices leading to decoded messages compliant with the dependent constraints. The most probable matrix of $\mathcal{W}_{\mathbf{Y}}$ is then taken as the MAP estimate $\widehat{\mathbf{W}}$ at Line 21.

Finally, $\widehat{\mathbf{W}}$ is used in Line 22 to decode the received set of packets $\mathbf{Y}$, thus providing an estimate $\widehat{\mathbf{X}}$ of $\mathbf{X}$.

If we assume that the maximization is performed in linear time w.r.t. the size of its search domain, the time θ required to separate the sources is

$$\theta = O\left(q^{G} + \left|\mathcal{W}_{\mathbf{y}}^{G}\right| + \left|\mathcal{W}_{\mathbf{Y}}\right|\right),$$

where $|\cdot|$ denotes the cardinal number of a set. The first iteration performs a full exploration of $\mathbb{F}_q^G$ containing q^G elements. The second iteration is performed over $\mathcal{W}_{\mathbf{y}}^{G}$, whose size in turn depends on the size of $\mathcal{W}_{\mathbf{y}}^{(\mathrm{I})}$. The maximization finally is performed over the elements of $\mathcal{W}_{\mathbf{Y}}$.

Notice that, if all the vectors in $\mathbb{F}_q^G$ are admissible decoding rows, then $\left|\mathcal{W}_{\mathbf{y}}^{G}\right| \approx q^{G\times G}$. In other words, the worst case scenario is when all full-rank matrices have to be tested against the dependent constraints. Conversely, the most favorable scenario is when only the rows of $\mathbf{A}^{-1}$ are admissible. In this case, the size of $\mathcal{W}_{\mathbf{y}}^{G}$ is upper bounded by $G!$, while a smaller set can be generated, *e.g.*, by taking the packets sequence numbers into account.

In conclusion, the computing time of PANDA depends mostly on the constraints imposed on the independent part of the messages.

Algorithm 1 Protocol-Assisted Network-Decoding Algorithm

1: **function** $\widehat{\mathbf{X}} = \textsc{Decode}(\mathbf{Y})$
2: **Input**: $(G \times L)$ encoded matrix $\mathbf{Y}$.
3: **Output**: $(G \times L)$ separated source matrix $\widehat{\mathbf{X}}$.
4: $\mathcal{W}_{\mathbf{y}}^{(\mathrm{I})} \leftarrow \emptyset$; ▷ Set of admissible decoding rows
5: $\mathcal{W}_{\mathbf{Y}} \leftarrow \emptyset$; ▷ Set of admissible decoding matrices
6: **for all** $\mathbf{w} \in \mathbb{F}_q^G$ **do**
7: $\mathbf{x} \leftarrow \mathbf{w}\mathbf{Y}$;
8: **if** $\mathbf{x}\mathbf{\Pi}_{\mathrm{I}} \in \mathcal{C}^{(\mathrm{I})}$ **then**
9: $\mathcal{W}_{\mathbf{y}}^{(\mathrm{I})} \leftarrow \mathcal{W}_{\mathbf{y}}^{(\mathrm{I})} \cup \{\mathbf{w}\}$;
10: $p_{\mathbf{w}} \leftarrow p_{\boldsymbol{X}}^{(\mathrm{I})}(\mathbf{x})$;
11: **end if**
12: **end for**
13: $\mathcal{W}_{\mathbf{y}}^{G} \leftarrow \bigotimes_{i=1}^{G} \mathcal{W}_{\mathbf{y}}^{(\mathrm{I})} \cap \mathcal{W}$;
14: **for all** $\mathbf{W} \in \mathcal{W}_{\mathbf{y}}^{G}$ **do**
15: $\mathbf{X} \leftarrow \mathbf{W}\mathbf{Y}$;
16: **if** $\mathbf{X}\mathbf{\Pi}_{\mathrm{D}} \in \mathcal{C}^{(\mathrm{D})}$ **then**
17: $\mathcal{W}_{\mathbf{Y}} \leftarrow \mathcal{W}_{\mathbf{Y}} \cup \{\mathbf{W}\}$;
18: $P_{\mathbf{W}} \leftarrow P_{\boldsymbol{X}}^{(\mathrm{D})}(\mathbf{X}) \prod_{i=1}^{G} p_{\mathbf{w}_i}$;
19: **end if**
20: **end for**
21: $\widehat{\mathbf{W}} \leftarrow \arg\max_{\mathbf{W}\in\mathcal{W}_{\mathbf{Y}}} P_{\mathbf{W}}$;
22: $\widehat{\mathbf{X}} \leftarrow \widehat{\mathbf{W}}\mathbf{Y}$;
23: **end function**

V. ANALYSIS OF COMMONLY USED FIELDS

This section focuses on some of the fields commonly used in protocols such as TCP, UDP, or RTP and that may be exploited by PANDA.

The constraints imposed by a given field F eliminates combinations of rows of $\mathbf{Y}$ that are not compliant. This reduces the search space for $\mathbf{W}$.

If F belongs to the independent part of the messages, the set of messages $\mathbf{x}$ that are linear combinations of the rows of $\mathbf{Y}$ and are compliant with F has an average number of elements

$$S^{(\mathrm{I,F})} = \mathrm{E}\left[\left|\mathrm{rowspan}\left(\boldsymbol{Y}\right) \cap \mathcal{C}_{\boldsymbol{X}}^{(\mathrm{I,F})}\right|\right], \tag{15}$$

where $\mathcal{C}_{\boldsymbol{X}}^{(\mathrm{I,F})}$ is the set of messages compliant with F.

If F belongs to the dependent part of the messages, the set of generations $\mathbf{X}$ whose rows are linear combinations of the rows of $\mathbf{Y}$ and are compliant with F has an average number of elements

$$S^{(\mathrm{D,F})} = \mathrm{E}\left[\left|\mathrm{span}\left(\boldsymbol{Y}\right) \cap \mathcal{C}_{\boldsymbol{X}}^{(\mathrm{D,F})}\right|\right], \tag{16}$$

where $\mathcal{C}_{\boldsymbol{X}}^{(\mathrm{D,F})}$ is the set of generations compliant with F. In (16), $\mathrm{span}\left(\boldsymbol{Y}\right)$ represents the set of all $G\times L$ matrices whose rows are linear combinations of the rows of $\boldsymbol{Y}$.

In both cases, the expectation, denoted $\mathrm{E}\left[\cdot\right]$, is taken over all mixing matrices $\boldsymbol{A}$ and source matrices $\boldsymbol{X}$.

This paper details only three classes of field: constant, control, and timestamp fields. The remaining classes are only briefly commented at the end of this section. Their impact will be detailed in an extended version of this paper.

V-A. Constant field

Encapsulated source messages $\mathbf{x}_i$ frequently include a field, *i.e.*, a set of bits, that is *constant* and perfectly known by the receiver prior to reception. Fields such as the protocol version, the content type, or the source ID, are constant throughout the generation and known after the initial handshake phase and belong to this class of fields, denoted K in what follows. The bits of K are not necessarily contiguous. Nevertheless, we assume that there are enough contiguous bits of K to form an integer number $N_{\mathrm{K}} \geq 1$ of well-aligned symbols in $\mathbb{F}_q$. The case of constant but isolated bits, can be reduced to the case of the enumeration field, mentioned later.

Let $\mathbf{k}$ be the $1 \times N_{\mathrm{K}}$ non-null vector containing the concatenated constant symbols. Denote $\mathbf{\Pi}_{\mathrm{K}}$ the projection

matrix extracting the field K from $\mathbf{x}_i$, *i.e.*,

$$\mathbf{\Pi}_{\mathrm{K}}\mathbf{x}_i = \mathbf{k},\ i = 1, \ldots, G. \tag{17}$$

This field introduces an independent constraint on each message. The set of messages consistent with (17) is

$$\mathcal{C}_{\boldsymbol{x}}^{(\mathrm{I,K})} = \left\{\mathbf{x} \in \mathbb{F}_q^L \mid \mathbf{x}\mathbf{\Pi}_{\mathrm{K}} = \mathbf{k}\right\}. \tag{18}$$

$S^{(\mathrm{I,K})}$ is obtained by determining the number of vectors $\mathbf{w}$ in $\mathbb{F}_q^G$ such that $\mathbf{w}\mathbf{Y} \in \mathcal{C}_{\boldsymbol{x}}^{(\mathrm{I,K})}$. The row vector $\mathbf{w}\mathbf{Y}$ can be rewritten as

$$\mathbf{w}\mathbf{Y} = \mathbf{w}(\mathbf{A}\mathbf{X}) = (\mathbf{w}\mathbf{A})\mathbf{X} = \mathbf{v}\mathbf{X} \tag{19}$$

A necessary condition to have $\mathbf{w}$ equal to one of the rows of $\mathbf{A}^{-1}$ is that $\mathbf{w}\mathbf{Y} = \mathbf{x}_i$, *i.e.*, that $\mathbf{v} = \mathbf{e}_i$. In order for the constraint to be satisfied, one should have

$$\begin{aligned} \mathbf{v}\mathbf{X}\mathbf{\Pi}_{\mathrm{K}} &= \mathbf{e}_i\mathbf{X}\mathbf{\Pi}_{\mathrm{K}} \\ \mathbf{v}\begin{pmatrix} \mathbf{k} \\ \vdots \\ \mathbf{k} \end{pmatrix} &= \mathbf{k} \\ \left(\sum_{j=1}^{G} v_j\right)\mathbf{k} &= \mathbf{k} \end{aligned} \tag{20}$$

Since $\mathbf{k} \neq 0$, one deduces from (20) that

$$\sum_{j=1}^{G} v_j = \sum_{j=1}^{G} (\mathbf{w}\mathbf{A})_j = 1.$$

The vectors $\mathbf{w}$ in $\mathbb{F}_q^G$ such that $\sum_{j=1}^{G}(\mathbf{w}\mathbf{A})_j = 1$ form a subspace of dimension $G-1$, thus $S^{(\mathrm{I,K})} = q^{G-1}$, independently of the size N_{K} of the constant field.

V-B. Control field

Usually, source messages include a control sequence, checksum or cyclic redundancy check (CRC), to verify its integrity. The corresponding field is denoted C and is computed from a subset D of the symbols in the message.

For a given checksum/CRC function $\mathcal{H}(\cdot)$, the induced independent constraint set is

$$\mathcal{C}_{\boldsymbol{x}}^{(\mathrm{I,C})} = \left\{\mathbf{x} \in \mathbb{F}_q^L \mid \mathbf{x}\mathbf{\Pi}_{\mathrm{C}} = \mathcal{H}(\mathbf{x}\mathbf{\Pi}_{\mathrm{D}})\right\}. \tag{21}$$

The checksum/CRC function induces a binning of the possible values of the input $\mathbf{x}\mathbf{\Pi}_{\mathrm{D}}$. Messages in the same bin share the same control value. The binning of CRCs achieves maximum distance separability of the inputs. Most checksums used in practice also empirically show good separability.

$S^{(\mathrm{I,C})}$ depends on the considered $\mathcal{H}(\cdot)$. Provided that the input vectors are uniformly divided into q^{N_C} bins, one has $\Pr\left\{\mathbf{v}\mathbf{X} \in \mathcal{C}_{\boldsymbol{x}}^{(\mathrm{I,C})}\right\} \approx q^{-N_{\mathrm{C}}}$ and

$$S^{(\mathrm{I,C})} = \mathrm{E}\left[\left|\mathrm{rowspan}\,(\boldsymbol{Y}) \cap \mathcal{C}_{\boldsymbol{x}}^{(\mathrm{I,C})}\right|\right] = O\left(q^{G-N_{\mathrm{C}}}\right). \tag{22}$$

V-C. Timestamp field

For several protocols, the messages in $\mathbf{X}$ include a timestamp field T of N_{T} symbols such that, in two consecutive messages of a generation, the difference $\Delta\mathrm{T}$ of their values follows a distribution $f_{\Delta\mathsf{T}}(\cdot)$, assumed to be known. The value of T for the first message in the generation is drawn from a generally unknown distribution $f_{\mathsf{T}}(\cdot)$.

The difference $\Delta\mathrm{T}$ is usually not evaluated in $\mathbb{F}_q^{N_{\mathrm{T}}}$, but we can define a function $\mathcal{T} : \mathbb{F}_q^{N_{\mathrm{T}}} \times \mathbb{F}_q^{N_{\mathrm{T}}} \mapsto \mathbb{F}_q^{N_{\mathrm{T}}}$ that maps the result of this difference in $\mathbb{F}_q$.

The timestamp field does not, in general, induce constraints. Nevertheless, consider for all $i \in \{2, \ldots, G\}$ the quantity $\Delta_{\mathbf{X}}\mathrm{T}_i = \mathcal{T}(\mathbf{x}_i\mathbf{\Pi}_{\mathrm{T}}, \mathbf{x}_{i-1}\mathbf{\Pi}_{\mathrm{T}})$. The associated *a priori* pmfs

$$\begin{aligned} p_{\boldsymbol{x}}^{(\mathrm{I,T})}(\mathbf{x}) &= f_{\mathsf{T}}(\mathbf{x}\mathbf{\Pi}_{\mathrm{T}}) \\ P_{\boldsymbol{X}}^{(\mathrm{D,T})}(\mathbf{X}) &= \prod_{i=2}^{G} f_{\Delta\mathsf{T}}(\Delta_{\mathbf{X}}\mathrm{T}_i) \end{aligned}$$

are useful to discriminate among generations compliant with the other fields, and to re-order the messages.

V-D. Remaining fields

The enumeration field gathers symbols that take values in a set $\mathcal{E}$, whose size is much smaller than the number of elements that could be represented considering all combinations of bits for the symbols of the field. This is the case of fields such as protocol source port, destination port, or content type. The enumeration field may also contain symbols in $\mathbb{F}_q$ for which some (but not all) bits are known. In this case, the set consist in all the symbols sharing the same constant pattern of bits.

For this field, the constraint in (15) may be written as

$$\mathcal{C}_{\boldsymbol{x}}^{(\mathrm{I,E})} = \left\{\mathbf{x} \in \mathbb{F}_q^L \mid \mathbf{x}\mathbf{\Pi}_{\mathrm{E}} \in \mathcal{E}\right\}. \tag{23}$$

which depends on three factors: (i) the size of $\mathcal{E}$ (ii) its rank and (iii) the possibility for the same value to appear in different messages. In fact, having a smaller number of possible values allows to discard a higher number of inadmissible decoding rows, providing that few linear combinations of the values in $\mathcal{E}$ also belong to the set.

The sequence field is an integer value that is increased by a constant and known amount in two consecutive messages of a generation. Except when the sequence field of the last packet of the previous generation is known, this field does not provide in general a restriction on the set of candidate decoding rows. It does, however, provide useful information about the set of candidate decoding matrices. In particular, the sequence field allows to restrict the search to generations such that their sequence numbers are progressive. This can considerably reduce the average size of the set of candidates.

	G	N_{gen}	$\lvert\mathcal{W}_{\mathbf{y}}^{G}\rvert$	$\lvert\mathcal{W}_{\mathbf{Y}}\rvert$	η	θ
PANDA	2	1000	2.0	2.0	100 %	0.015 s
	4	1000	132.4	4926.5	100 %	1.45 s
ICA	2	1000	n.a.	n.a.	0 %	0.020 s
	4	1000	n.a.	n.a.	0 %	6.43 s

Table I. Performance of PANDA compared to an entropy-based ICA algorithm in terms of average success rate η and per-generation computing time θ. For PANDA, the average size of the sets $\mathcal{W}_{\mathbf{y}}^{G}$ and $\mathcal{W}_{\mathbf{Y}}$ is also reported.

VI. EXPERIMENTAL RESULTS

Consider a source generating sequences of L uniformly distributed bytes encapsulated with a RTP/UDP-like header. In the header, we consider

- a one-byte K field, with value 255,
- a two-byte T field expressed in ms, with uniformly distributed initial value and with $\boldsymbol{X}\boldsymbol{\Pi}_{\text{T}}$ distributed as a Gaussian variable with mean $\mu = 500$ ms and standard deviation $\sigma = 50$ ms, rounded to the nearest ms.
- a two-byte C field, evaluated as detailed in [12].

The source generates N_{gen} generations of G messages. The NC operations in $\mathbb{F}_{2^8}$ are simulated mixing each generation with a $G \times G$ matrix $\mathbf{A}$ chosen uniformly among the matrices of full rank. The matrix $\mathbf{Y}$ of mixed packets is assumed received without error at destination.

PANDA is compared to the entropy-based ICA technique introduced in [8], which searches the G linearly-independent demixing rows that minimize the empirical entropy of the decoded packets. The results are provided in Table I for both algorithms for G=2 and G=4, with L=2 kBytes. In both cases, PANDA is able to recover all messages, without permutation ambiguity, thanks to the T field.

The entropy-based ICA technique is completely inadequate for this task. Even for a generation size of 2 and a packet length of 2 kB, this technique has been unable to correctly reconstruct a single packet from a stream of 2000 packets. This is due mostly to the uniformity of the bits within the packet, and is rendered even most severe by the relatively large size of the finite field.

The average computing time per generation is evaluated on an Intel Core2 Duo E6400 at 2.13 GHz. For larger generations, more fields have to be taken into account to reduce the search space for the admissible lines of the $\mathbf{W}$ matrix. Moreover, as indicated in Section IV, the combinatorial part may be significantly simplified by accounting for the content of the T field, and by selecting candidate matrices leading to a correctly ordered T field in sequence of decoded messages.

VII. CONCLUSIONS AND FUTURE WORK

In the context of random linear NC, this paper presents a new technique for network decoding when the NC coefficients are not transmitted. At the receiver side, decoding is performed by MAP estimation, using the *a priori* knowledge on the structure of the headers included by the transmission protocol *before* NC. The proposed decoding algorithms first identifies all admissible rows that are compliant with the constraints imposed on the packet headers. Then, it determines the matrices formed by these admissible rows that are globally consistent with the protocol. This reduces significantly the search complexity w.r.t. a brute-force search.

Our theoretical study, supported by our experimental results, show that this approach is able to recover 100 % of the transmitted packets correctly. If the packets are all of the same length, no NC overhead is required. This result would not be achievable with traditional source separation techniques based on independent component analysis, due to the relatively short length of the packets and the quasi-uniformity of the compressed data.

REFERENCES

[1] T. Ho, M. Médard, J. Shi, M. Effros, and D. Karger, "On randomized network coding," in *Proc. IEEE ISIT*, Kanagawa, Japan, 2003, pp. 11–20.

[2] P. Chou, Y. Wu, and K. Jain, "Practical network coding," in *Proc. of Allerton Conf. on Commun. Control and Comput.*, Monticello, IL, USA, Oct. 2003.

[3] M. Jafari, L. Keller, C. Fragouli, and K. Argyraki, "Compressed network coding vectors," in *Proc. of IEEE Int. Symp. Inf. Theory*, Seoul, Republic of Korea, 2009, pp. 109–113.

[4] N. Thomos and P. Frossard, "Toward one symbol network coding vectors," *IEEE Commun. Lett.*, vol. 16, no. 11, pp. 1860–1863, 2012.

[5] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*, 1st ed. Academic Press, 2010.

[6] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Elsevier J. on Neural Networks*, vol. 13, no. 4–5, pp. 411–430, 2000.

[7] A. Yeredor, "Independent component analysis over galois fields of prime order," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5342–5359, 2011.

[8] H. W. Gutch, P. Gruber, A. Yeredor, and F. J. Theis, "ICA over finite field—separability and algorithms," *Signal Processing*, vol. 92, no. 8, pp. 1796–1808, 2012.

[9] I.-D. Nemoianu, C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "On a practical approach to source separation over finite fields for network coding applications," in *Proc. IEEE ICASSP*, Vancouver, Canada, 2013, pp. 1335–1339.

[10] P. Duhamel and M. Kieffer, *Joint source-channel decoding: A cross-layer perspective with applications in video broadcasting*. Academic Press, 2009.

[11] J. F. Kurose and K. W. Ross, *Computer networking: a top-down approach featuring the Internet*. Pearson Education, 2004.

[12] J. Postel, "User Datagram Protocol," Internet Engineering Task Force, Aug. 1980. [Online]. Available: http://www.ietf.org/rfc/rfc768.txt