Coded Cooperative Data Exchange for a Secret Key

Thomas A. Courtade
EECS Department, University of California, Berkeley
Email: courtade@eecs.berkeley.edu

Thomas R. Halford
Trellisware Technologies, San Diego, CA
Email: thalford@trellisware.com

***Abstract*—We consider a cooperative data exchange problem with the goal of generating a secret key. Specifically, we investigate the number of public transmissions required for a set of clients to agree on a secret key with probability one, subject to the constraint that it remains private from an eavesdropper.**

Although the problems are closely related, we prove that secret key generation with fewest linear transmissions is NP-hard, while it is known that the analogous problem in traditional cooperative data exchange can be solved in polynomial time. In doing this, we completely characterize the best-possible performance of linear coding schemes, and also prove that linear codes can be strictly suboptimal.

I. Introduction

In an asymptotic setting, the relationship between secret key (SK) generation and communication for omniscience was revealed in the pioneering work [1]. In [1], Csiszár and Narayan left characterizing the minimum communication rate required to generate a maximum-rate SK as an open problem.

In [2], El Rouayheb *et al*. introduced a non-asymptotic, combinatorial version of Csiszár's communication for omniscience problem, which they called *(coded) cooperative data exchange* (CCDE). Since its introduction, this problem has received significant attention (cf. [3]–[5] and references therein).

In [5], it was shown that the omniscience-secrecy relations in [1] translate nicely to the combinatorial CCDE setting. Building on this, we investigate the number of public transmissions required for a set of clients to agree on a SK with probability one, subject to the constraint that it remains private from an eavesdropper. In doing so, we address a combinatorial analog of Csiszár's open problem.

Related to the present work is [6], which studies the minimum communication *rate* required to generate a SK in the two-terminal, asymptotic setting. Specifically, a multi-letter expression characterizing this minimum rate is given in terms of the *r-rounds interactive common information*. Despite the similarity in spirit, the asymptotic setting of [6] and the combinatorial setting of the present paper are considerably different in nature, and the proof techniques used are orthogonal.

Finally, the problem of weakly secure network coding is also related to our setting. The primary distinction is that we only aim to generate a SK, and do not require that the nodes communicate for omniscience. We refer the reader to [7] and references therein for details.

This work is supported by DARPA-DSO under contracts W15P7T-12-C-5013 and W911QX-13-C-0010. T. Courtade is supported in part by the NSF Center for Science of Information under grant agreement CCF-0939370.

Approved for public release; distribution is unlimited. DISCLAIMER: The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Our Contributions

Given the close connection between CCDE and SK generation, we show two surprising results. First, we prove that finding an optimal (linear) coding scheme for SK generation is NP-hard, while it is known that the analogous CCDE problem is in P. In doing this, we completely characterize the attainable performance for linear coding schemes. Second, despite linear codes being optimal for CCDE, we demonstrate that they can be strictly suboptimal for SK generation. Several ancillary results are also proved.

This paper is organized as follows. Section II formally defines our system model and reviews relevant results on CCDE. In Section III, we state and prove our main results. Section IV delivers concluding remarks.

II. System Model and Preliminaries

We first establish basic notation. Throughout, we use calligraphic notation to denote sets. For two sets $\mathcal{A} \subset \mathcal{B}$, we write $\mathcal{B}\backslash\mathcal{A}$ to denote those elements in $\mathcal{B}$, but not in $\mathcal{A}$. If $\mathcal{A}$ is a singleton set (i.e., $\mathcal{A} = \{a\}$), then we often use the notation $\mathcal{B} - a \triangleq \mathcal{B}\backslash\{a\}$ for convenience. We define $\mathbb{Z}$ to be the set of integers. For positive $m \in \mathbb{Z}$, we use the shorthand notation $[m] \triangleq \{1, 2, \ldots, m\}$. Finally, for random variables X, Y, we write $I(X;Y)$ for the mutual information between X and Y.

A. System Model

Throughout, we consider networks defined by a set of n clients $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$, a positive integer m, and a family of finite sets $\{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n\}$ (each $\mathcal{I}_j \subseteq [m]$ and $\cup_{j=1}^{n} \mathcal{I}_j = [m]$) in the following way. Define the random (column) vector $\underline{\mathrm{X}} \triangleq [X_1, X_2, \ldots, X_m]^T$, where each X_i is a discrete random variable with equiprobable distribution on a finite field $\mathbb{F}$, and $(X_1, X_2, \ldots, X_m)$ are mutually independent[1]. The random variables $\{X_i\}_{i=1}^{m}$ are called *messages*, and $\{X_i : i \in \mathcal{I}_j\}$ is the set of messages initially held by client $c_j \in \mathcal{C}$. In other words, $\mathcal{I}_j$ defines the indices of messages initially held by client c_j, for $j = 1, \ldots, n$. Throughout, n will always denote the number of clients; since the sets $\mathcal{I}_j$ are always indexed by $j \in [n]$, we will use the shorthand notation $\{\mathcal{I}_j\}$ to denote the family $\{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_n\}$.

We adopt the communication model which is standard in index coding and CCDE problems. That is, we consider transmission schemes consisting of a finite number of communication rounds. In each round, a single client broadcasts an element of $\mathbb{F}$ (which can be a function of the messages initially held by that client and all previous transmissions) to all other clients over an error-free channel. It is further assumed that

[1]For technical reasons, we assume $|\mathbb{F}| > 2n$.

all clients have knowledge of the index sets $\mathcal{I}_1, \ldots, \mathcal{I}_n$, and thus follow a protocol which is mutually agreed upon. We will elaborate on the definition of a transmission protocol in the next subsection.

B. Transmission Protocols Defined

For a network defined by $\{\mathcal{I}_j\}$, a transmission protocol $\mathbf{P}$ (or simply, a protocol $\mathbf{P}$) consisting of t communication rounds is defined by n encoding functions $\{\mathsf{f}_1, \mathsf{f}_2, \ldots, \mathsf{f}_n\}$, and a t-tuple $(i_1, i_2, \ldots, i_t)$, where $i_k \in [n]$ indicates which client transmits during communication round k. More specifically, during communication round k, client c_{i_k} transmits

$$\mathsf{f}_{i_k}\Big(\{X_j : j \in \mathcal{I}_{i_k}\}, k, \{\mathsf{f}_{i_\ell}\}_{\ell=1}^{k-1}\Big) \in \mathbb{F}, \tag{1}$$

where we have abbreviated the transmitted symbols in rounds $\ell \in [k-1]$ by $\{\mathsf{f}_{i_\ell}\}_{\ell=1}^{k-1}$. For a given transmission protocol $\mathbf{P}$ requiring t communication rounds, we let $\mathbf{T}(\underline{\mathrm{X}}, \mathbf{P}) \in \mathbb{F}^t$ be the column vector with k^{th} entry equal to $\mathsf{f}_{i_k}\big(\{X_j : j \in \mathcal{I}_{i_k}\}, k, \{\mathsf{f}_{i_\ell}\}_{\ell=1}^{k-1}\big)$. Letting $\|\cdot\|$ be the length function, we have $\|\mathbf{T}(\underline{\mathrm{X}}, \mathbf{P})\| = t$. Note that $\mathbf{T}(\underline{\mathrm{X}}, \mathbf{P})$ is a random variable since it is a function of the random vector $\underline{\mathrm{X}}$. Generally, the transmission protocol under consideration will be clear from context. Hence, we abbreviate $\mathbf{T}(\underline{\mathrm{X}}) \triangleq \mathbf{T}(\underline{\mathrm{X}}, \mathbf{P})$ for convenience when there is no ambiguity.

A transmission protocol is said to be *linear* (over $\mathbb{F}$) if the encoding functions $\{\mathsf{f}_1, \mathsf{f}_2, \ldots, \mathsf{f}_n\}$ are of the form

$$\mathsf{f}_{i_k}\big(\{X_j : j \in \mathcal{I}_{i_k}\}, k, \{\mathsf{f}_{i_\ell}\}_{\ell=1}^{k-1}\big) = \sum_j \alpha_j^{(k)} X_j, \tag{2}$$

where $\alpha_j^{(k)} \in \mathbb{F}$ can be interpreted as the encoding coefficient for message j during communication round k. In this case, we can express $\mathbf{T}(\underline{\mathrm{X}}) = A\underline{\mathrm{X}}$, where $A \in \mathbb{F}^{t \times m}$ assuming the definitions $t \triangleq \|\mathbf{T}(\underline{\mathrm{X}})\|$ and $m \triangleq |\cup_j \mathcal{I}_j|$. Hence, the *encoding matrix* A provides a succinct description of a linear transmission protocol. Note that the order of transmissions corresponding to a linear protocol is inconsequential.

C. Transmission Protocols for Omniscience

Before proceeding, let $\mathsf{M}^\star(\{\mathcal{I}_j\})$ denote the optimal value of the following integer linear program (ILP), which arises in the CCDE problem (see [5] for details).

$$\begin{aligned} \text{minimize :} \quad & \sum_{j \in [n]} a_j \\ \text{subject to :} \quad & \sum_{j \in \mathcal{S}} a_j \geq \left| \bigcap_{j \in \bar{\mathcal{S}}} \bar{\mathcal{I}}_j \right| \quad \forall \text{ nonempty } \mathcal{S} \subset [n] \\ & a_j \in \mathbb{Z} \quad \text{for all } j \in [n], \end{aligned} \tag{3}$$

where $\bar{\mathcal{I}}_i \triangleq (\cup_j \mathcal{I}_j) \setminus \mathcal{I}_i$ and $\bar{\mathcal{S}} \triangleq [n] \setminus \mathcal{S}$. Though offered here without explanation, the quantity $\mathsf{M}^\star(\{\mathcal{I}_j\})$ will play an important role in our treatment and will be needed shortly.

A transmission protocol $\mathbf{P}$ is said to achieve *omniscience* if there exist decoding functions $\{\mathsf{g}_1, \mathsf{g}_2, \ldots, \mathsf{g}_n\}$ which satisfy

$$\mathsf{g}_j(\{X_i : i \in \mathcal{I}_j\}, \mathbf{T}(\underline{\mathrm{X}}, \mathbf{P})) = \underline{\mathrm{X}} \quad \text{for each } j \in [n] \tag{4}$$

with probability 1.

Theorem 1 (See [5, Theorem 2])**.** *If a protocol $\mathbf{P}$ achieves omniscience, then $\|\mathbf{T}(\underline{\mathrm{X}}, \mathbf{P})\| \geq \mathsf{M}^\star(\{\mathcal{I}_j\})$. Conversely, there always exists a linear protocol $\mathbf{P}_{\mathrm{L}}$ that achieves omniscience and has $\|\mathbf{T}(\underline{\mathrm{X}}, \mathbf{P}_{\mathrm{L}})\| = \mathsf{M}^\star(\{\mathcal{I}_j\})$.*

Theorem 1 addresses the central issue in the CCDE problem, which primarily investigates the number of transmissions required to achieve omniscience.

D. Transmission Protocols for Secret Keys

A transmission protocol (with corresponding transmission sequence $\mathbf{T}(\underline{\mathrm{X}})$) generates a secret key (SK) if there exist decoding functions $\{\mathsf{k}_1, \mathsf{k}_2, \ldots, \mathsf{k}_n\}$ which satisfy the following three properties:

(i) For all $j \in [n]$, and with probability 1,
$$\mathsf{k}_j(\{X_i : i \in \mathcal{I}_j\}, \mathbf{T}(\underline{\mathrm{X}})) = \mathsf{k}_1(\{X_i : i \in \mathcal{I}_1\}, \mathbf{T}(\underline{\mathrm{X}}))\,.$$
(ii) $\mathsf{k}_1(\{X_i : i \in \mathcal{I}_1\}, \mathbf{T}(\underline{\mathrm{X}}))$ is equiprobable on $\mathbb{F}$.
(iii) $I\left(\mathsf{k}_1(\{X_i : i \in \mathcal{I}_1\}, \mathbf{T}(\underline{\mathrm{X}}))\,; \mathbf{T}(\underline{\mathrm{X}})\right) = 0$.

In words, requirement (iii) guarantees that the public transmissions $\mathbf{T}(\underline{\mathrm{X}})$ reveal no information about $\mathsf{k}_1(\{X_i : i \in \mathcal{I}_1\}, \mathbf{T}(\underline{\mathrm{X}}))$. Requirement (i) asserts that all clients $c_j \in \mathcal{C}$ can compute $\mathsf{k}_1(\{X_i : i \in \mathcal{I}_1\}, \mathbf{T}(\underline{\mathrm{X}}))$. For these reasons, $\mathsf{k}_1(\{X_i : i \in \mathcal{I}_1\}, \mathbf{T}(\underline{\mathrm{X}}))$ is called a *secret key*. Naturally, a secret key should be equiprobable on its domain to make guessing difficult, thus motivating requirement (ii).

It is not immediately clear whether any protocol $\mathbf{P}$ generates a SK. However, it turns out that such protocols exist in great abundance. In particular, the existence of protocols that generate a SK depends solely on the family $\{\mathcal{I}_j\}$.

Theorem 2 (See [5, Theorem 6])**.** *For a network defined by $\{\mathcal{I}_j\}$, there exists a protocol $\mathbf{P}$ which generates a SK if and only if*

$$|\cup_j \mathcal{I}_j| \geq \mathsf{M}^\star(\{\mathcal{I}_j\}) + 1. \tag{5}$$

Despite the fact that $\mathsf{M}^\star(\{\mathcal{I}_j\})$ corresponds to the optimal value of an ILP, it can be computed in time polynomial in $|\cup_j \mathcal{I}_j|$ (see [3], [5]). Therefore, for any family $\{\mathcal{I}_j\}$, we can efficiently test whether (5) holds. Hence, the essential remaining question is: "*How many transmissions are needed to generate a SK?*"

To this end, let $\mathcal{P}(\{\mathcal{I}_j\})$ denote the set of protocols for $\{\mathcal{I}_j\}$ that generate a SK, and define

$$\mathsf{S}(\{\mathcal{I}_j\}) \triangleq \min\Big\{\|\mathbf{T}(\underline{\mathrm{X}}, \mathbf{P})\| : \mathbf{P} \in \mathcal{P}(\{\mathcal{I}_j\})\Big\}. \tag{6}$$

That is, $\mathsf{S}(\{\mathcal{I}_j\})$ is the minimum number of transmissions needed to generate a SK. Similarly, let $\mathcal{P}_{\mathrm{L}}(\{\mathcal{I}_j\})$ denote the set of linear protocols for $\{\mathcal{I}_j\}$ that generate a SK, and define

$$\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\}) \triangleq \min\Big\{\|\mathbf{T}(\underline{\mathrm{X}}, \mathbf{P})\| : \mathbf{P} \in \mathcal{P}_{\mathrm{L}}(\{\mathcal{I}_j\})\Big\}. \tag{7}$$

In words, $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$ is the minimum number of transmissions required to generate a SK when we restrict our attention to linear protocols. If $\{\mathcal{I}_j\}$ does not satisfy (5), then we set $\mathsf{S}(\{\mathcal{I}_j\}) = \mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\}) = \infty$.

Remark 1. *We will often write "$\{\mathcal{I}_j\}$ generates a SK" instead of the more accurate, but cumbersome, "For the network defined by $\{\mathcal{I}_j\}$, there exists a protocol $\mathbf{P}$ which generates a SK" whenever* (5) *holds.*

III. Main Results

In this section, we investigate the number of transmissions required to generate a SK. In particular, we completely characterize $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$, and make progress toward characterizing $\mathsf{S}(\{\mathcal{I}_j\})$.

As demonstrated in the previous section, the CCDE and SK-generation problems are closely connected through the quantity $\mathsf{M}^{\star}(\{\mathcal{I}_j\})$. Since Theorem 1 and the tractability of ILP (3) essentially resolve the CCDE problem, it is natural to conjecture that a similar result should hold for $\mathsf{S}(\{\mathcal{I}_j\})$ and $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$. Unfortunately, there is a fundamental difference between the problems, which is revealed by the following two negative results[2].

Theorem 3. *For some families* $\{\mathcal{I}_j\}$, $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\}) > \mathsf{S}(\{\mathcal{I}_j\})$.

Theorem 4. *Computing* $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$ *is NP-hard.*

For the CCDE problem, Theorem 1 asserts that linear protocols achieve optimal performance. Furthermore, the number of transmissions required by linear protocols is easily computed. For the problem of SK generation, the opposite is true. That is, linear protocols can be suboptimal, and the number of transmissions required by linear protocols is generally difficult to compute. This situation is parallel to that of multicast network coding and index coding. The two problems are closely related (cf. [8]), but exhibit the same dichotomy.

Despite the negative results offered by Theorems 3 and 4, we can characterize several properties of $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$, $\mathsf{S}(\{\mathcal{I}_j\})$, and $\mathsf{M}^{\star}(\{\mathcal{I}_j\})$. Some of these are demonstrated in the following results, which are needed as we progress toward proving Theorem 4. A complete characterization of $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$ will be given in Theorem 5.

Lemma 1. *If* $\{\mathcal{I}_j\}$ *generates a SK, then*

$$\mathsf{S}(\{\mathcal{I}_j\}) \le \mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\}) \le \mathsf{M}^{\star}(\{\mathcal{I}_j\})\,. \tag{8}$$

Proof. By definition, $\mathsf{S}(\{\mathcal{I}_j\}) \le \mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$ since $\mathcal{P}_{\mathrm{L}}(\{\mathcal{I}_j\}) \subseteq \mathcal{P}(\{\mathcal{I}_j\})$. The second inequality follows from the proof of [5, Theorem 6], in which a linear transmission protocol $\mathbf{P}$ is constructed that generates a SK with $\|\mathbf{T}(\underline{\mathrm{X}}), \mathbf{P})\| = \mathsf{M}^{\star}(\{\mathcal{I}_j\})$ communication rounds. □

We say that $\{\mathcal{J}_j\}$ is a *subfamily* of $\{\mathcal{I}_j\}$ if there is a set $\mathcal{S} \subset \cup_j \mathcal{I}_j$ such that $\mathcal{J}_j = \mathcal{I}_j \backslash \mathcal{S}$ for all $j \in [n]$.

Lemma 2. *If* $\{\mathcal{J}_j\}$ *is a subfamily of* $\{\mathcal{I}_j\}$*, then*

$$\mathsf{M}^{\star}(\{\mathcal{J}_j\}) \le \mathsf{M}^{\star}(\{\mathcal{I}_j\})\,, \tag{9}$$

$$\mathsf{S}(\{\mathcal{J}_j\}) \ge \mathsf{S}(\{\mathcal{I}_j\})\,, \text{ and} \tag{10}$$

$$\mathsf{S}_{\mathrm{L}}(\{\mathcal{J}_j\}) \ge \mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})\,. \tag{11}$$

Proof. By De Morgan's law, it is easy to verify that

$$\left|\bigcap_{j\in\bar{\mathcal{S}}} \bar{\mathcal{I}}_j\right| \ge \left|\bigcap_{j\in\bar{\mathcal{S}}} \bar{\mathcal{J}}_j\right| \quad \text{for all nonempty } \mathcal{S} \subset [n], \tag{12}$$

where $\bar{\mathcal{I}}_i \triangleq (\cup_j \mathcal{I}_j)\backslash\mathcal{I}_i$ and $\bar{\mathcal{J}}_i \triangleq (\cup_j \mathcal{J}_j)\backslash\mathcal{J}_i$. Therefore, the constraints in ILP (3) are relaxed, and $\mathsf{M}^{\star}(\{\mathcal{J}_j\}) \le \mathsf{M}^{\star}(\{\mathcal{I}_j\})$ by definition.

To show (10), observe that any transmission protocol which generates a SK for the subfamily $\{\mathcal{J}_j\}$ also generates a SK for the family $\{\mathcal{I}_j\}$ by ignoring the set of messages $\{X_i : i \notin \cup_j \mathcal{J}_j\}$. Hence, it follows that $\mathsf{S}(\{\mathcal{J}_j\}) \ge \mathsf{S}(\{\mathcal{I}_j\})$. If $\{\mathcal{J}_j\}$ can not generate a SK, the inequality trivially holds. This argument also proves (11). □

Lemma 2 demonstrates monotonicity, but offers no insight into whether inequalities (9)-(11) are tight. The following lemma identifies settings under which (11) holds with equality, and will prove useful later on.

Lemma 3. *If* $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\}) < \mathsf{M}^{\star}(\{\mathcal{I}_j\})$*, then there exists some* $\ell \in \cup_j \mathcal{I}_j$ *for which* $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j - \ell\}) = \mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$.

Proof. Define $m \triangleq |\cup_j \mathcal{I}_j|$. By definition, there is a linear transmission protocol $\mathbf{P}_{\mathrm{L}}$ which generates a SK in $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$ communication rounds. Let $\mathbf{T}(\underline{\mathrm{X}}, \mathbf{P}_{\mathrm{L}}) = A\underline{\mathrm{X}}$ be the sequence of transmissions made by $\mathbf{P}_{\mathrm{L}}$, and let $\{\mathsf{k}_1, \ldots, \mathsf{k}_n\}$ be valid decoding functions.

Since $\|\mathbf{T}(\underline{\mathrm{X}}, \mathbf{P}_{\mathrm{L}})\| = \mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\}) < \mathsf{M}^{\star}(\{\mathcal{I}_j\})$, Theorem 1 asserts that the protocol $\mathbf{P}_{\mathrm{L}}$ can not achieve omniscience. Therefore, by a possible permutation of clients, we can assume without loss of generality that there is no function g_1 for which

$$\mathsf{g}_1(\{X_i : i \in \mathcal{I}_1\}, A\underline{\mathrm{X}}) = \underline{\mathrm{X}} \text{ with probability 1.} \tag{13}$$

As a consequence, there must exist a nonzero vector $\underline{\mathrm{v}}$ such that $A\underline{\mathrm{v}} = 0$, and $v_i = 0$ for all $i \in \mathcal{I}_1$. Indeed, if there is no such $\underline{\mathrm{v}}$, then client c_1 can solve a full-rank system of equations to recover $\underline{\mathrm{X}}$, yielding a contradiction. Since $\underline{\mathrm{v}}$ is not identically zero, there is some ℓ for which $v_\ell \neq 0$.

Define $\hat{X}_\ell \equiv 0$, and $\hat{X}_i \triangleq X_i$ for $i \in \cup_j(\mathcal{I}_j - \ell)$. Also, define vectors $\hat{\underline{\mathrm{X}}} \triangleq [\hat{X}_1, \hat{X}_2, \ldots, \hat{X}_m]^T$ and $\underline{\mathrm{X}}' \triangleq \hat{\underline{\mathrm{X}}} + X_\ell \cdot \underline{\mathrm{v}}$. First, we note that

$$\mathsf{k}_j\left(\{\hat{X}_i : i \in \mathcal{I}_j\}, A\hat{\underline{\mathrm{X}}}\right) = \mathsf{k}_1\left(\{\hat{X}_i : i \in \mathcal{I}_1\}, A\hat{\underline{\mathrm{X}}}\right)$$

for all $j \in [n]$ since

$$\mathsf{k}_j(\{X_i : i \in \mathcal{I}_j\}, A\underline{\mathrm{X}}) = \mathsf{k}_1(\{X_i : i \in \mathcal{I}_1\}, A\underline{\mathrm{X}})$$

with probability 1 by definition.

Next, observe that:

$$\begin{aligned} &I\left(\mathsf{k}_1\left(\{\hat{X}_i : i \in \mathcal{I}_1\}, A\hat{\underline{\mathrm{X}}}\right); A\hat{\underline{\mathrm{X}}}\right) \\ &= I\left(\mathsf{k}_1\left(\{\hat{X}_i + X_\ell \cdot v_i : i \in \mathcal{I}_1\}, A\underline{\mathrm{X}}'\right); A\underline{\mathrm{X}}'\right) \qquad (14) \\ &= I\left(\mathsf{k}_1(\{X_i : i \in \mathcal{I}_1\}, A\underline{\mathrm{X}}); A\underline{\mathrm{X}}\right) = 0 \qquad (15) \end{aligned}$$

In the above,

- (14) follows since $A\underline{\mathrm{X}}' = A(\hat{\underline{\mathrm{X}}} + X_\ell \cdot \underline{\mathrm{v}}) = A\hat{\underline{\mathrm{X}}}$, and $v_i = 0$ for all $i \in \mathcal{I}_1$.
- (15) follows from the (crucial) observation that $\underline{\mathrm{X}}'$ and $\underline{\mathrm{X}}$ are equal in distribution, and by definition of A and k_1.

[2]The proofs of Theorems 3 and 4 are deferred until the end of this section.

Finally, by similar reasoning, we note that the random variable $\mathsf{k}_1\Big(\{\hat{X}_i : i \in \mathcal{I}_1\}, A\hat{\underline{\mathbf{X}}}\Big)$ is equiprobable on $\mathbb{F}$ since

$$\begin{aligned}
&\mathsf{k}_1\Big(\{\hat{X}_i : i \in \mathcal{I}_1\}, A\hat{\underline{\mathbf{X}}}\Big) \\
&= \mathsf{k}_1\Big(\{\hat{X}_i + X_\ell \cdot v_i : i \in \mathcal{I}_1\}, A\underline{\mathbf{X}}'\Big) \qquad (16)\\
&\overset{d}{=} \mathsf{k}_1(\{X_i : i \in \mathcal{I}_1\}, A\underline{\mathbf{X}})\,, \qquad (17)
\end{aligned}$$

and $\mathsf{k}_1(\{X_i : i \in \mathcal{I}_1\}, A\underline{\mathbf{X}})$ is equiprobable on $\mathbb{F}$ by definition. In (17), $\overset{d}{=}$ indicates equality in distribution.

Therefore, we can conclude that a SK can be generated by the subfamily $\{\mathcal{I}_j - \ell\}_j$ by applying the protocol $\mathbf{P}_{\mathrm{L}}$ and fixing $X_\ell \equiv 0$. This proves that $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j - \ell\}) \le \mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$. By Lemma 2, the reverse inequality also holds. □

In order to proceed, we will need to introduce *critical* families. To this end, let $\tau \ge 1$ be an integer. A family $\{\mathcal{I}_j\}$ is τ-*critical* if the following hold:

(i) $|\cup_j \mathcal{I}_j| - \mathsf{M}^\star(\{\mathcal{I}_j\}) = \tau$, and
(ii) $\mathsf{M}^\star(\{\mathcal{I}_j - i\}) = \mathsf{M}^\star(\{\mathcal{I}_j\})$ for all $i \in \cup_j \mathcal{I}_j$.

It is interesting to note that τ-criticality of $\{\mathcal{I}_j\}$ can be efficiently tested since $\mathsf{M}^\star(\{\mathcal{I}_j\})$ is computable in polynomial time. 1-critical families have a threshold property: families $\{\mathcal{I}_j\}$ that are 1-critical generate a secret key, and no proper subfamilies of $\{\mathcal{I}_j\}$ generate SKs. This is a consequence of Theorem 2 and the definition of 1-criticality.

A *minimum τ-critical subfamily* $\{\mathcal{J}_j^\star\}$ of $\{\mathcal{I}_j\}$ satisfies

$$\big|\cup_j \mathcal{J}_j^\star\big| \le |\cup_j \mathcal{J}_j| \qquad (18)$$

for all other τ-critical subfamilies $\{\mathcal{J}_j\}$ of $\{\mathcal{I}_j\}$. Note that if $\{\mathcal{I}_j\}$ is 1-critical, then $\{\mathcal{I}_j\}$ is its own unique minimum 1-critical subfamily.

The following Theorem demonstrates that minimum 1-critical subfamilies completely characterize $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$.

Theorem 5. *If $\{\mathcal{I}_j\}$ generates a SK, then*

$$\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\}) = \mathsf{M}^\star\big(\{\mathcal{J}_j^\star\}\big) = \big|\cup_j \mathcal{J}_j^\star\big| - 1, \qquad (19)$$

where $\{\mathcal{J}_j^\star\}$ is a minimum 1-critical subfamily of $\{\mathcal{I}_j\}$.

Proof. By inductively applying Lemma 3, we can find a subfamily $\{\mathcal{T}_j\}$ of $\{\mathcal{I}_j\}$ for which

$$\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\}) = \mathsf{M}^\star(\{\mathcal{T}_j\})\,. \qquad (20)$$

Let $\{\mathcal{J}_j\}$ be any 1-critical subfamily of $\{\mathcal{T}_j\}$. We have the following chain of inequalities

$$\begin{aligned}
\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\}) \le \mathsf{S}_{\mathrm{L}}\big(\{\mathcal{J}_j^\star\}\big) &\le \mathsf{M}^\star\big(\{\mathcal{J}_j^\star\}\big) && (21)\\
&\le \mathsf{M}^\star(\{\mathcal{J}_j\}) && (22)\\
&\le \mathsf{M}^\star(\{\mathcal{T}_j\}) && (23)\\
&= \mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})\,. && (24)
\end{aligned}$$

The above steps can be justified as follows:

- (21) follows from Lemmas 1 and 2.
- By definition of τ-criticality, (18) is equivalent to $\mathsf{M}^\star\big(\{\mathcal{J}_j^\star\}\big) \le \mathsf{M}^\star(\{\mathcal{J}_j\})$. Thus, (22) follows since $\{\mathcal{J}_j^\star\}$ is a minimum 1-critical subfamily of $\{\mathcal{I}_j\}$, and $\{\mathcal{J}_j\}$ is a 1-critical subfamily of $\{\mathcal{I}_j\}$.
- (23) follows from Lemma 2.
- (24) is a repeat of (20).

This proves that $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\}) = \mathsf{M}^\star\big(\{\mathcal{J}_j^\star\}\big)$. Recalling the definition of 1-criticality completes the proof. □

Theorem 5 implies that $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$ is easily computed if we can identify a minimum 1-critical subfamily of $\{\mathcal{I}_j\}$. By Theorem 4, we know this must be NP-hard. In order to prove this, we require the following lemma which lends a hypergraph interpretation to 1-criticality. For a hypergraph $H = (\mathcal{V}, \mathcal{E})$, recall that an edge set $\mathcal{E}' \subseteq \mathcal{E}$ is a *minimal connected dominating edge set* if the subhypergraph $H' = (\mathcal{V}, \mathcal{E}')$ is connected, and the removal of any edge from $\mathcal{E}'$ disconnects H'.

Lemma 4. *Consider a hypergraph $H = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = \mathcal{C}$, and edge set $\mathcal{E} = \cup_j \mathcal{I}_j$. Define H as follows: a vertex $c_j \in \mathcal{V}$ is contained in the edge $e \in \mathcal{E}$ if and only if $e \in \mathcal{I}_j$. H is connected if and only if*

$$\mathsf{M}^\star(\{\mathcal{I}_j\}) < |\cup_j \mathcal{I}_j|\,. \qquad (25)$$

In particular, $\{\mathcal{I}_j\}$ is 1-critical if and only if $\mathcal{E}$ is a minimal connected dominating edge set.

Proof. First, suppose H is not connected. By definition, there must exist a nontrivial partition $\mathcal{V} = (\mathcal{S}, \bar{\mathcal{S}})$ such that there is no edge $e \in \mathcal{E}$ which contains vertices from both $\mathcal{S}$ and $\bar{\mathcal{S}}$. Stated another way, $(\cup_{j\in\mathcal{S}} \mathcal{I}_j) \cap (\cup_{j\in\bar{\mathcal{S}}} \mathcal{I}_j) = \emptyset$. Hence, ILP (3) includes the two constraints

$$\sum_{j\in\mathcal{S}} a_j \ge \left|\bigcap_{j\in\bar{\mathcal{S}}} \bar{\mathcal{I}}_j\right| = \left|\bigcup_{j\in\mathcal{S}} \mathcal{I}_j\right| \qquad (26)$$

$$\sum_{j\in\bar{\mathcal{S}}} a_j \ge \left|\bigcap_{j\in\mathcal{S}} \bar{\mathcal{I}}_j\right| = \left|\bigcup_{j\in\bar{\mathcal{S}}} \mathcal{I}_j\right|, \qquad (27)$$

the sum of which imply $\mathsf{M}^\star(\{\mathcal{I}_j\}) \ge |\cup_j \mathcal{I}_j|$. By taking the contrapositive, we have proven

$$\mathsf{M}^\star(\{\mathcal{I}_j\}) < |\cup_j \mathcal{I}_j| \implies H \text{ is connected.} \qquad (28)$$

Next, suppose H is connected, and assume without loss of generality that $\mathcal{E} = \cup_j \mathcal{I}_j \triangleq \{1, 2, \ldots, m\}$. Since H is connected, there is a transmission protocol for which the entries of $\mathbf{T}(\underline{\mathbf{X}})$ are precisely $\{X_1 + X_j\}_{j=2}^m$. Indeed, by connectivity of H, there must be some client c initially holding X_1 and some X_e (say, X_2 without loss of generality), and can therefore transmit $X_1 + X_2$ during the first communication round. By induction, assume that $\{X_1 + X_j\}_{j=2}^{m-1}$ are transmitted during the first $m-2$ communication rounds (permuting indices of the X_i's if necessary). Again, by connectivity of H, there must be a client c' which initially holds X_m and X_k, where $k < m$. Hence, in communication round $m-1$, client c' can transmit $(X_1 + X_k) - (X_k - X_m) = X_1 + X_m$. Noting that

$$(X_1, X_1 + X_2, \ldots, X_1 + X_m) \overset{d}{=} (X_1, X_2, \ldots, X_m),$$

we have $I(X_1;\mathbf{T}(\underline{\mathrm{X}}))=0$. If client $c\in e\in\mathcal{E}$, then it can recover X_1 from the transmission X_1+X_e by simply subtracting X_e. Since H is connected, each $c\in\mathcal{V}$ belongs to some edge in $\mathcal{E}$, and therefore all clients can recover X_1 losslessly. Since X_1 is equiprobable on $\mathbb{F}$ by definition, we can conclude that $\{\mathcal{I}_j\}$ generates a SK. Theorem 2 asserts that we must have $\mathsf{M}^\star(\{\mathcal{I}_j\})<|\cup_j\mathcal{I}_j|$, and we have proven

$$\mathsf{M}^\star(\{\mathcal{I}_j\})<|\cup_j\mathcal{I}_j| \iff H \text{ is connected.} \tag{29}$$

We now prove the second claim. To this end, suppose $\{\mathcal{I}_j\}$ is 1-critical. Then $\mathsf{M}^\star(\{\mathcal{I}_j\})=|\cup_j\mathcal{I}_j|-1$, which implies H is connected (and thus $\mathcal{E}$ is dominating) by (29). Consider the subhypergraph $H'=(\mathcal{V},\mathcal{E}\backslash\{e\})$, which corresponds to the subfamily $\{\mathcal{I}_j-e\}$ of $\{\mathcal{I}_j\}$. Since $\{\mathcal{I}_j\}$ is 1-critical, we must have $\mathsf{M}^\star(\{\mathcal{I}_j-e\})=\mathsf{M}^\star(\{\mathcal{I}_j\})=|\cup_j\mathcal{I}_j|-1=|\cup_j(\mathcal{I}_j-e)|$. By (29), H' must be disconnected, and therefore $\mathcal{E}$ is a minimal connected dominating edge set.

On the other hand, suppose $\mathcal{E}$ is a minimal connected dominating edge set. Since H is connected, (29) implies

$$\mathsf{M}^\star(\{\mathcal{I}_j\})\le|\cup_j\mathcal{I}_j|-1. \tag{30}$$

Since $\mathcal{E}$ is minimal, for any $e\in\mathcal{E}$, $H'=(\mathcal{V},\mathcal{E}\backslash\{e\})$ is disconnected, and (29) implies

$$\mathsf{M}^\star(\{\mathcal{I}_j-e\})\ge|\cup_j(\mathcal{I}_j-e)|=|\cup_j\mathcal{I}_j|-1. \tag{31}$$

Applying Lemma 2, we must have $\mathsf{M}^\star(\{\mathcal{I}_j\})=\mathsf{M}^\star(\{\mathcal{I}_j-e\})$, and $|\cup_j\mathcal{I}_j|-\mathsf{M}^\star(\{\mathcal{I}_j\})=1$, which implies $\{\mathcal{I}_j\}$ is 1-critical. □

We are finally in a position to prove Theorem 4.

Proof of Theorem 4. Consider a hypergraph $H=(\mathcal{V},\mathcal{E})$ with vertex set $\mathcal{V}=\mathcal{C}$, and edge set $\mathcal{E}=\cup_j\mathcal{I}_j$. Define H as follows: a vertex $c_j\in\mathcal{V}$ is contained in the edge $e\in\mathcal{E}$ if and only if $e\in\mathcal{I}_j$.

We can assume $\{\mathcal{I}_j\}$ generates a SK. By Theorem 5 and Lemma 4, computing $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$ is equivalent to computing the the number of edges in a minimum connected dominating edge set (i.e., a minimal connected dominating edge set with fewest possible edges). It is easy to see that the NP-complete SET COVER DECISION PROBLEM is a special case[3]. □

Theorem 3 is proved by the following example: Let $n=7$, and consider the family $\{\mathcal{I}_j\}$ defined by $\mathcal{I}_1=\{1,2,3,4\}$, and $\mathcal{I}_2,\dots,\mathcal{I}_7$ are all $\binom{4}{2}$ distinct 2-element subsets of $\{1,2,3,4\}$. By direct computation, we find that $\{\mathcal{I}_j-\{1\}\}$ is a minimum 1-critical subfamily, and hence $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})=2$ by Theorem 5. Suppose $\mathbb{F}=\{0,1,\alpha,\beta\}^2=\mathsf{GF}(4)\times\mathsf{GF}(4)$. Thus, we can express $X_j=(X_j^{(1)},X_j^{(2)})$ for each $j=1,\dots,4$, where $X_j^{(1)},X_j^{(2)}$ are mutually independent, each equiprobable on $\mathsf{GF}(4)$. It is readily verified that the single transmission

$$\left(X_1^{(1)}+\alpha X_2^{(1)}+X_3^{(1)},\,X_1^{(1)}+\beta X_2^{(1)}+X_4^{(1)}\right)\in\mathbb{F} \tag{32}$$

by client c_1 permits reconstruction of the SK $\mathsf{k}_1(\{X_i:i\in\mathcal{I}_1\},\mathbf{T}(\underline{\mathrm{X}}))\triangleq(X_3^{(1)},X_4^{(1)})\in\mathbb{F}$ at all clients. Hence, we can conclude $1=\mathsf{S}(\{\mathcal{I}_j\})<\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})=\mathsf{M}^\star(\{\mathcal{I}_j\})=2$.

The above construction is a *vector-linear* transmission protocol, and cannot be realized by a protocol which is linear over $\mathbb{F}$. We remark that this construction can be generalized to increase the gap between $\mathsf{S}(\{\mathcal{I}_j\})$ and $\mathsf{S}_{\mathrm{L}}(\{\mathcal{I}_j\})$, but details are omitted due to space constraints.

IV. CONCLUDING REMARKS

Finding a minimum 1-critical subfamily $\{\mathcal{J}_j^\star\}$ of $\{\mathcal{I}_j\}$ is NP-hard, but we can easily find a 1-critical subfamily $\{\mathcal{J}_j\}$ satisfying $|\cup_j\mathcal{J}_j|\le(2+\ln(n-1))\left|\cup_j\mathcal{J}_j^\star\right|$ by using the greedy algorithm in [9]. Given the connection to the SET COVER PROBLEM pointed out in the proof of Theorem 4, this approximation is best-possible unless NP has $O(n^{O(\ln n)})$ time algorithms. With $\{\mathcal{J}_j\}$ in hand, we can implement the linear protocol described in the proof of Lemma 4 requiring $\mathsf{S}_{\mathrm{L}}(\{\mathcal{J}_j\})$ transmissions, which is guaranteed to be roughly within a $\ln n$ factor of optimal.

Although we have focused exclusively on protocols that generate a single SK, it is natural to consider protocols that generate τ independent secret keys. To this end, let $\mathsf{S}_{\mathrm{L}}^{(\tau)}(\{\mathcal{I}_j\})$ denote the minimum number of transmissions required by a linear protocol to generate τ independent secret keys. A slight modification of our argument yields:

Theorem 6. *Let $\tau\ge1$ be an integer. If there is a protocol $\mathbf{P}$ for $\{\mathcal{I}_j\}$ which generates τ independent secret keys, then*

$$\mathsf{S}_{\mathrm{L}}^{(\tau)}(\{\mathcal{I}_j\})=\mathsf{M}^\star\left(\{\mathcal{J}_j^\star\}\right)=\left|\cup_j\mathcal{J}_j^\star\right|-\tau, \tag{33}$$

where $\{\mathcal{J}_j^\star\}$ is a minimum τ-critical subfamily of $\{\mathcal{I}_j\}$.

Presumably, it is NP-hard to find a minimum τ-critical subfamily of $\{\mathcal{I}_j\}$ for $\tau\ge2$, which would lead to a natural generalization of Theorem 4. This requires a modification of Lemma 4, and appears challenging. Indeed, the structure of 1-critical families is succinctly captured by minimal connected dominating edge sets, but the structural class of τ-critical families is much richer for $\tau\ge2$. We leave this general hardness problem for future work.

REFERENCES

[1] I. Csiszár and P. Narayan, "Secrecy capacities for multiple terminals," *IEEE Trans. on Inf. Theory*, vol. 50, no. 12, pp. 3047 – 3061, Dec. 2004.

[2] S. El Rouayheb, A. Sprintson, and P. Sadeghi, "On coding for cooperative data exchange," in *IEEE Information Theory Workshop (ITW)*, Jan. 2010.

[3] N. Milosavljevic, S. Pawar, S. El Rouayheb, M. Gastpar, and K. Ramchandran, "Optimal deterministic polynomial-time data exchange for omniscience," *arXiv preprint:1108.6046 [cs.IT]*, Aug. 2011.

[4] M. Gonen and M. Langberg, "Coded cooperative data exchange problem for general topologies," in *Proc. of IEEE International Symposium on Information Theory (ISIT)*, July 2012, pp. 2606–2610.

[5] T. Courtade and R. Wesel, "Coded cooperative data exchange in multihop networks," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 1136–1158, 2014.

[6] H. Tyagi, "Common information and secret key capacity," *IEEE Trans. on Inf. Theory*, vol. 59, no. 9, pp. 5627–5640, 2013.

[7] M. Yan and A. Sprintson, "Algorithms for weakly secure data exchange," in *Network Coding (NetCod), 2013 Intl. Symp. on*. IEEE, 2013, pp. 1–6.

[8] M. Effros, S. El Rouayheb, and M. Langberg, "An equivalence between network coding and index coding," *arXiv:1211.6660*, Nov. 2012.

[9] W. Ren and Q. Zhao, "A note on 'Algorithms for connected set cover problem and fault-tolerant connected set cover problem'," *Theoretical Computer Science*, vol. 412, no. 45, pp. 6451–6454, 2011.

[3] Consider any subsets $\mathcal{A}_1,\mathcal{A}_2,\dots,\mathcal{A}_k$ whose union covers $\mathcal{U}$. For $u'\notin\mathcal{U}$, define $\mathcal{U}'=\mathcal{U}\cup\{u'\}$, and $\mathcal{A}_j'=\mathcal{A}_j\cup\{u'\}$ for $j\in[k]$. Clearly, $\{\mathcal{A}_{j_i}\}_{i=1}^m$ is a minimum cover of $\mathcal{U}$ if and only if $\{\mathcal{A}_{j_i}'\}_{i=1}^m$ is a minimum connected cover of $\mathcal{U}'$.