Information Leakage in Index Coding With Sensitive and Nonsensitive Messages

Yucheng Liu, Lawrence Ong, *Senior Member, IEEE*, Parastoo Sadeghi, *Senior Member, IEEE*, Sarah Johnson, *Senior Member, IEEE*, Joerg Kliewer, *Senior Member, IEEE*, and Phee Lep Yeoh, *Member, IEEE*

***Abstract*—Index coding can be viewed as a compression problem with multiple decoders with side information. In such a setup, an encoder compresses a number of messages into a common codeword such that every decoder can decode its requested messages with the help of knowing some other messages as side information. In this paper, we study how much information is leaked to a guessing adversary observing the codeword in index coding, where some messages in the system are sensitive and others are not. The non-sensitive messages can be used by the encoder in a manner similar to secret keys to mitigate leakage of the sensitive messages to the adversary. We first characterize the optimal information leakage rate of a given index coding problem by the optimal compression rate of a related problem, which is constructed by adding an extra decoder with certain parameters to the original problem. Both the achievability and converse of the characterization are derived from a graph-theoretic perspective based on confusion graphs (Alon et al. 2008). In particular, the achievable coding scheme is a randomized mapping exploiting certain symmetrical properties of the confusion graph. Our second main result is a practical deterministic linear coding scheme, developed from the rank minimization method based on fitting matrices (Bar-Yossef et al. 2011). The linear scheme leads to an upper bound on the optimal leakage rate, which is proved to be tight over all deterministic scalar linear codes. While it is shown through an example that simultaneously achieving optimal compression and leakage rates is not always possible, time-sharing between different schemes could be used to balance the compression and leakage rates. Finally, we show how our results can be applied to different variants of index coding.**

***Index Terms*—Index coding, compression, guessing, information leakage, graph theory.**

Manuscript received 15 April 2022; revised 14 September 2022; accepted 15 December 2022. Date of publication 11 January 2023; date of current version 13 June 2023. This work was supported in part by the ARC Discovery Scheme under Grant DP190100770; in part by the U.S. National Science Foundation under Grant 2201824 and Grant 1815322; and in part by the ARC Future Fellowship under Grant FT190100429. This article was presented in part at the IEEE International Symposium on Information Theory (ISIT) 2022 [DOI: 10.1109/ISIT50566.2022.9834747]. *(Corresponding author: Yucheng Liu.)*

Yucheng Liu was with the School of Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia. He is now with the Advanced Technology Group, Dolby Australia, McMahons Point, NSW 2060, Australia (e-mail: liuyc34@gmail.com).

Lawrence Ong and Sarah Johnson are with the School of Engineering, The University of Newcastle, Callaghan, NSW 2308, Australia.

Parastoo Sadeghi is with the School of Engineering and Information Technology, The University of New South Wales, Canberra, ACT 2600, Australia.

Joerg Kliewer is with the Department of Electrical and Computer Engineering, New Jersey Institute of Technology, Newark, NJ 07102 USA.

Phee Lep Yeoh is with the School of Electrical and Information Engineering, University of Sydney, Sydney, NSW 2006, Australia.

Digital Object Identifier 10.1109/JSAIT.2022.3232126

I. Introduction

INDEX coding [2], [3] is a canonical problem in network information theory, which can be viewed essentially as a compression problem with one encoder and multiple decoders with side information. The encoder observes a number of messages and tries to efficiently compress them into a codeword such that every decoder can decode its wanted messages from the codeword with the help of its own side information. In this work, we study information leakage in index coding when the codeword is eavesdropped on by a guessing adversary. The adversary observes the codeword and tries to maximize the probability of correctly guessing its messages of interest within a certain number of trials.

Our goal is to minimize the leakage to this adversary, which is defined as the ratio between the adversary's probability of successful guessing *after and before* observing the codeword [4], [5]. This way of measuring information leakage was originally introduced as the min-entropy leakage [4]. A similar leakage metric was independently explored in a different setup [5], where the adversary is interested in guessing some randomized function of the messages rather than the messages themselves. Since their introductions, such information leakage measurements and their variants have been studied extensively from both the information-theoretic and computer science perspectives [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16].

Information leakage in single-encoder single-decoder compression systems has been studied considering multiple leakage metrics, under the assumption that the source code is deterministic and a random secret key is shared between the encoder and the decoder [17]. Leakage to a guessing adversary for single-encoder single-decoder source compression has been analysed [18], where the compression level is proven using random-coding arguments (to satisfy some rate-distortion measure).

On single-encoder multiple-decoder compression systems, we recently studied [19] the information leakage in index coding, with an assumption that the encoder aims to protect all of the messages against the adversary. This assumption is applied in most existing works in index coding [20], [21], [22], [23], [24] where security and privacy were investigated.

However, in many practical circumstances, some messages may be sensitive while others are not, and thus secrecy

2641-8770 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

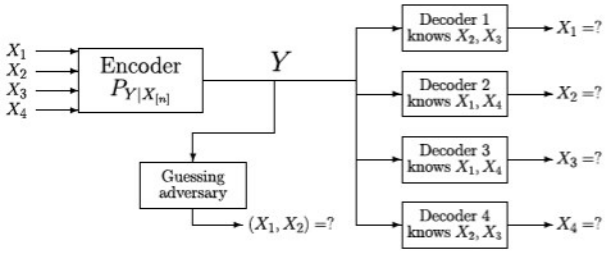

Fig. 1. Consider four i.i.d. uniform binary messages X_i, $i \in [4]$, where X_1, X_2 are sensitive and X_3, X_4 are non-sensitive. A guessing adversary eavesdrops the codeword Y and tries to guess the sensitive messages. When guessing blindly (without knowing Y), the probability of the adversary correctly guessing (X_1, X_2) is only $1/4$. To satisfy the legitimate decoders, the encoder can generate Y as $Y = (X_1 \oplus X_2, X_3 \oplus X_4)$. However, such Y leads to certain amount of information leakage as the adversary's correct guessing probability when observing it becomes $1/2$. To simultaneously satisfy the decoders and prevent any information leakage, the encoder can generate $\tilde{Y} = (X_1 \oplus X_3, X_2 \oplus X_4)$. In this way, the sensitive messages are perfectly protected against the adversary using non-sensitive messages, where the probability of the adversary correctly guessing (X_1, X_2) remains $1/4$ even after observing Y.

loss should be measured over only sensitive messages. For example, consider a file storage system where some files are private and need to be protected from any adversary, while the other files are non-private and thus need not be protected. These files, whether private or non-private, may be requested by different clients, each already have some other files as side information. The goal is to satisfy the clients' requirements while keeping the private files as safe as possible from the adversary. In another example, a video streaming server provides both free and paid videos. The server aims to deliver the requested contents to the legitimate receivers and at the same time, protects the paid videos from an adversary. A similar setting can also be motivated from the adversary's perspective: the encoder may know that the adversary is only interested in a subset of messages and thus protect these messages.

The distinction between sensitive and non-sensitive messages enables the encoder to treat the non-sensitive messages like secret keys and design a smart (possibly randomized) coding scheme to simultaneously satisfy the legitimate decoders and mitigate information leakage to the adversary. Figure 1 serves as a toy example showing how the encoder can reduce the information leakage by smartly designing coding schemes.

A. Organization and Contributions

In Section II, we describe the system model of index coding in detail, explicitly define our measure of information leakage of sensitive messages to the adversary, and provide necessary mathematical preliminaries.

In Section III, we characterize the optimal information leakage rate of any given problem using the optimal compression rate of another related problem. The related problem is constructed from the original problem by adding an extra decoder that requests to decode all the non-sensitive messages and knows the others as side information. The result is derived from a graph-theoretic perspective utilizing confusion graphs [25], which entirely captures the decoding requirements imposed by the decoders in the system. In particular, the achievability of the optimal leakage rate is proved by constructing a randomized coding scheme, designed based upon certain symmetries of the confusion graph of index coding. The result reveals a connection and tradeoff between compression and secrecy.

In Section IV, we propose a more practical, deterministic linear coding scheme, based on the rank minimization method over fitting matrices [3]. The scheme is proved to yield optimal leakage rate over all deterministic scalar linear codes. Examples are given to show the efficacy of the scheme. We also construct an example to show a negative result that the optimal compression and leakage rates cannot always be simultaneously achieved. To address this issue, a discussion on time-sharing between different coding schemes is included.

We investigate certain variants of the problem setup in Section V. In Section V-A, we consider the case where the goal is to minimize information leakage to the adversary while satisfying certain compression rate constraints. A simple extension of the linear coding scheme proposed in Section IV is given to suit the case. In Section V-B, we study information leakage in *pliable* index coding [26], where there is no predetermined desired message sets at the decoders and each decoder is satisfied whenever it can decode any messages not in its side information set. This variant is relevant to content distribution networks. It turns out that results similar to that for index coding hold for pliable index coding.

Finally, in Section VI, we conclude the paper with a few remarks and open problems.

Notation: For non-negative integers a and b, $[a]$ denotes the set $\{1, 2, \ldots, a\}$, and $[a : b]$ denotes the set $\{a, a+1, \ldots, b\}$. If $a > b$, $[a : b] = \emptyset$. For a finite set A, $|A|$ denotes its cardinality. For two sets A and B, $A \times B$ denotes their Cartesian product. For a sequence of sets $A_1, A_2, \ldots, A_t$, we may simply use $\prod_{j\in[t]} A_j$ to denote their Cartesian product. For any discrete random variable Z over an alphabet $\mathcal{Z}$ with probability distribution P_Z, we denote realizations with the small letter $z \in \mathcal{Z}$. For any $K \subseteq \mathcal{Z}$, $P_Z(K) \doteq \sum_{z\in K} P_Z(z)$.

II. Problem Formulation

A. System Model

We consider an encoder that has n uniformly distributed and independent messages X_i, $i \in [n]$. Each message is a sequence of length t that takes values from $\mathcal{X}^t$ for some finite field $\mathcal{X} = \mathbb{F}_q$. For any $S \subseteq [n]$, set $X_S \doteq (X_i, i \in S)$, $x_S \doteq (x_i, i \in S)$, and $\mathcal{X}_S \doteq \mathcal{X}^{|S|t}$. Thus $X_{[n]}$ denotes the tuple of all n messages, and $x_{[n]} \in \mathcal{X}_{[n]}$ denotes a realization of the message n-tuple. By convention, $X_\emptyset = x_\emptyset = \mathcal{X}_\emptyset = \emptyset$.

The encoder encodes the n messages to some codeword Y. There are m decoders. Decoder $i \in [m]$ wants to decode messages X_{W_i} from Y for some $W_i \subseteq [n]$ and has X_{A_i} as side information for some $A_i \subseteq [n] \backslash W_i$. We allow degenerate decoders in the system, who want nothing (i.e., $W_i = \emptyset$). A degenerate decoder can always decode what it wants by definition.

More formally, a $(t, M, f, \mathbf{g})$ *index code* is defined by

- One *stochastic* encoding function $f : \mathcal{X}^{nt} \to \{1, 2, \ldots, M\}$ at the encoder that maps each message tuple

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

$x_{[n]} \in \mathcal{X}^{nt}$ to a codeword $y \in \{1, 2, \ldots, M\}$ according to a conditional probability distribution $P_{Y|X_{[n]}}$, and

- *m deterministic* decoding functions $\mathbf{g} = (g_i, i \in [m])$, one for each decoder $i \in [m]$, such that $g_i : \{1, 2, \ldots, M\} \times \mathcal{X}^{|A_i|t} \to \mathcal{X}^{|W_i|t}$ maps the codeword y and the side information x_{A_i} to some estimated sequence $\hat{x}_{W_i}$.

We say a $(t, M, f, \mathbf{g})$ index code is *valid* if and only if (iff) every decoder can perfectly decode its wanted messages (i.e., $x_{W_i} = \hat{x}_{W_i}, \forall i \in [m]$). We call $t^{-1} \log_q M$ the compression rate of the $(t, M, f, \mathbf{g})$ index code. We say a compression rate R is achievable iff there exists a valid $(t, M, f, \mathbf{g})$ code such that

$$R \geq t^{-1} \log_q M. \tag{1}$$

The optimal compression rate β, also referred to as the broadcast rate,[1] can be defined as [27]

$$\beta = \lim_{t \to \infty} \min_{\text{valid } (t, M, f, \mathbf{g}) \text{ code}} R. \tag{2}$$

Any index coding instance is described by the parameter tuple $(n, m, (W_i, i \in [m]), (A_i, i \in [m]))$.

B. Confusion Graph

Any index coding instance can also be characterized by a family of confusion graphs, $(\Gamma_t, t \in \mathbb{Z}^+)$ [25]. For a given sequence length t, the confusion graph Γ_t is an undirected graph defined on the message sequence tuple alphabet $\mathcal{X}_{[n]}$. That is, the vertex set $V(\Gamma_t) = \mathcal{X}_{[n]}$. Vertex $x_{[n]}$ in Γ_t corresponds to the realization $x_{[n]}$. Any two different vertices $x_{[n]}, z_{[n]}$ are adjacent in Γ_t iff $x_{W_i} \neq z_{W_i}$ and $x_{A_i} = z_{A_i}$ for some decoder $i \in [m]$. We call any pair of vertices satisfying this condition *confusable*. Hence, $E(\Gamma_t) = \{\{x_{[n]}, z_{[n]}\} : x_{W_i} \neq z_{W_i}$ and $x_{A_i} = z_{A_i}$ for some $i \in [m]\}$.

For correct decoding at all decoders, any two realizations $x_{[n]}, z_{[n]}$ can be mapped to the same codeword y with nonzero probabilities iff they are not confusable [25]. See Figure 2 below for a toy example of the confusion graph of a 3-message 3-decoder index coding instance. For the definitions of basic graph-theoretic notions, see any textbook on graph theory (e.g., Scheinerman and Ullman [28]).

We may denote an index coding problem whose confusion graphs are $(\Gamma_t, t \in \mathbb{Z}^+)$ simply as Γ. Consider any set $J \subseteq [n]$. The subproblem induced by message subset J is characterized by the tuple $(|J|, m, (W_i \cap J, i \in [m]), (A_i \cap J, i \in [m]))$. Let $\Gamma(J)$ and $\Gamma_t(J)$ denote the subproblem induced by J and the confusion graph of message length t of the subproblem, respectively.

The broadcast rate $\beta(\Gamma)$ can be characterized by the confusion graphs $(\Gamma_t, t \in \mathbb{Z}^+)$ as [27, Sec. 3.2]

$$\beta(\Gamma) = \lim_{t \to \infty} \frac{1}{t} \log_q \chi(\Gamma_t) = \lim_{t \to \infty} \frac{1}{t} \log_q \chi_{\mathrm{f}}(\Gamma_t), \tag{3}$$

where $\chi(\cdot)$ and $\chi_{\mathrm{f}}(\cdot)$ respectively denote the chromatic number and fractional chromatic number of a graph.

[1] It has been shown [27, Lemma 1.1] that the broadcast rate is independent of the underlying alphabet $\mathcal{X}$.

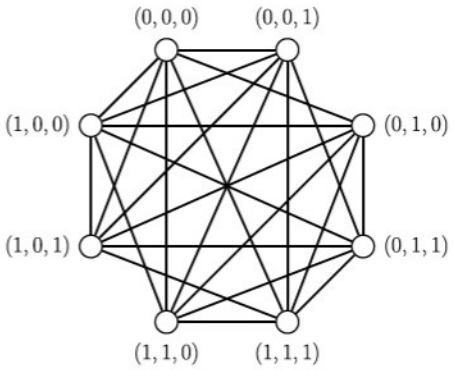

Fig. 2. The confusion graph Γ_1 with $t = 1$ and $q = 2$ for the index coding instance $(3, 3, (\{1\}, \{2\}, \{3\}), (\emptyset, \{3\}, \{2\}))$. Note that, for example, $x_{[n]} = (0, 0, 0)$ and $z_{[n]} = (0, 0, 1)$ are confusable because $x_3 = 0 \neq z_3 = 1$ and $x_{A_3} = x_2 = 0 = z_2 = z_{A_3}$. Suppose $(0, 0, 0)$ and $(0, 0, 1)$ are mapped to the same codeword y with certain nonzero probabilities. Then upon observing this y, decoder 3 will not be able to tell whether the value for X_3 is 0 or 1 based on its side information of $X_2 = 0$. For this graph, it can be easily verified that the independence number is 2, and that the chromatic number equals the fractional chromatic number, both of which are equal to 4.

C. Information Leakage Measurement

We assume the codeword Y is eavesdropped by a guessing adversary, who knows a subset of messages X_K as side information. The rest of the messages X_{K^c} are divided into two groups, where the sensitive messages are denoted by X_S and the non-sensitive ones are denoted by X_U. The information leakage from Y to the adversary will be measured only over the sensitive messages X_S. Note that K, S, U are non-overlapping and $[n] = K \cup S \cup U$. Let k, s, u denote the cardinality of sets K, S, U, respectively.

Upon observing the codeword, the adversary tries to guess the value of X_S according to the maximum likelihood rule within a number of trials. In other words, the adversary always guesses the most probable message tuple realization, and if not correct, then the second most probable realization, and so on, until it exhausts the number of guesses it can make.

Before we go into the details of the leakage measure, we define the following notation which will be used repeatedly later. Consider any valid $(t, M, f, \mathbf{g})$ index code. For any y value and any $J_1, J_2 \subseteq [n]$, let

$$\mathcal{X}_{J_1}(x_{J_2}, y) = \left\{ x_{J_1} \in \mathcal{X}_{J_1} : P_{Y, X_{J_1 \cup J_2}}(y, x_{J_1 \cup J_2}) > 0 \right\} \tag{4}$$

denote the set of x_{J_1} values jointly possible with (x_{J_2}, y). For example, $\mathcal{X}_S(x_K, y)$ denotes the set of x_S values jointly possible with (x_K, y). That is, upon observing a certain side information and codeword tuple (x_K, y), the adversary has only the values in $\mathcal{X}_S(x_K, y)$ left to guess from.

We characterize the adversary's number of guesses using a function of sequence length, $c : \mathbb{Z}^+ \to \mathbb{Z}^+$. It is natural to assume c to be a non-decreasing function of t. Consider any function c such that for any encoding scheme f and generated codeword Y,

$$c(t) > \max_{x_K, y} |\mathcal{X}_S(x_K, y)|, \qquad t = 1, 2, 3, \ldots \tag{5}$$

The right hand side of (5) denotes the maximum number of possible x_S values given a (x_K, y) tuple. Thus, (5) means

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

that, upon observing any (x_K, y), the number of guesses the adversary can make is larger than the number of possible values of sensitive messages. Hence, the adversary is guaranteed to successfully guess the sensitive messages and the problem becomes trivial. Therefore, we consider only the case where there exists some encoding scheme f such that $c(t)$ is upper-bounded by $\max_{x_K,y} |\mathcal{X}_S(x_K, y)|$.

Consider any valid $(t, M, f, \mathbf{g})$ index code. Before eavesdropping the codeword Y, the expected probability of the adversary successfully guessing x_S within $c(t)$ number of guesses, denoted by $\mathrm{P_s}(X_K)$, is

$$\mathbb{E}_{X_K}\left[\max_{J\subseteq\mathcal{X}_S:|J|\le c(t)} \sum_{x_S\in J} P_{X_S|X_K}(x_S|X_K)\right].$$

The expected successful guessing probability after observing Y, denoted by $\mathrm{P_s}(X_K, Y)$, is

$$\mathbb{E}_{Y,X_K}\left[\max_{\substack{J\subseteq\mathcal{X}_S:\\ |J|\le c(t)}} \sum_{x_S\in J} P_{X_S|Y,X_K}(x_S|Y, X_K)\right].$$

The leakage to the adversary, denoted by L, is defined as the logarithm of the ratio between the expected probabilities of the adversary successfully guessing x_S *after and before* observing Y. That is,

$$L \doteq \log_q \frac{\mathrm{P_s}(X_K,Y)}{\mathrm{P_s}(X_K)}. \tag{6}$$

It should be noted that the definition of L carries a clear operational meaning. A leakage of $L = \ell$ simply means that observing the codeword increases the adversary's probability of successfully guessing the value of sensitive messages by q^ℓ times. Generally speaking, whenever we consider a hard decision (i.e., guessing) adversary, the ratio between the correct guessing probabilities after and before eavesdropping describes the amount of the increase in the adversarial power, and thus can be a reasonable measure of the information leakage.

Remark 1: The idea of measuring leakage as the ratio of the adversary's successful guessing probabilities has been introduced and explored in various contexts [4], [5]. The min-entropy leakage [4] quantifies the leakage when the adversary tries to guess the information source in one try, while the maximal leakage [5] is a measurement of the worst-case one-try leakage of any function of the source. Our definition is closer to the min-entropy leakage in the sense that the adversary is interested in the messages themselves rather than some functions of the messages, yet with the key difference that we allow the adversary to make multiple guesses. Note that various versions of multiple-guess leakage have been formulated and studied in different settings [5, Sec. III], [13], [29], [30], [31].

Remark 2: Although carrying different operational meanings, given the fact that the messages are uniformly distributed in our index coding system, the leakage L in (6) turns out to be equal to the min-entropy leakage [4] from X_S to Y given X_K, which is also equal to the maximal leakage [5] from X_S to Y given side information X_K. Given such relationships, it can be shown that L in (6) is zero iff the codeword Y and the sensitive messages X_S are stochastically independent given the side information X_K [5, Corollary 2].

Remark 3: In *secure* index coding [20], [21], [22], [23], [24], the encoder tries to achieve perfect secrecy against the adversary while satisfying the legitimate decoders' requirements. However, such a strictly secure coding scheme may not always exist, even with the help of secret keys [23], [24].

Given the definition of L in (6), the leakage rate of the $(t, M, f, \mathbf{g})$ index code is

$$\mathcal{L} = t^{-1}L \tag{7}$$

and the optimal leakage rate for the problem can then be defined as

$$\mathcal{L}^* \doteq \lim_{t\to\infty} \inf_{(t,M,f,\mathbf{g})\text{ codes}} \mathcal{L}. \tag{8}$$

Remark 4: The choice of the base of the logarithm in (6) is arbitrary. Selecting a different base will incur a correction term when calculating the leakage rate (7).

III. A Characterization of $\mathcal{L}^*$

In this section, we present our first main result in Theorem 1, which is a characterization of the optimal leakage rate $\mathcal{L}^*$. Detailed graph-theoretic proofs are given right after the theorem, which are then followed by a few discussions of interesting points. We also identify two classes of index coding instances that satisfy certain structural properties, leading to further simplifications of the result in Theorem 1.

For any index coding problem Γ, we define a related problem $\tilde{\Gamma}$ by adding an extra decoder to Γ. The extra decoder is indexed by $m+1$ (as there are m original decoders in Γ), which knows side information indexed by the set $A_{m+1} = K \cup S$ and requests to decode messages indexed by the set $W_{m+1} = U$.

We have the following theorem.

Theorem 1: For any index coding problem Γ, we have

$$\mathcal{L}^*(\Gamma) = \beta\Big(\tilde{\Gamma}(S\cup U)\Big) - u. \tag{9}$$

We first prove an intermediate result regarding $\tilde{\Gamma}$, which will prove useful in later proofs.

Lemma 1: Consider any valid $(t, M, f, \mathbf{g})$ index code for Γ. We have

$$\max_{x_K,y} |\mathcal{X}_S(x_K, y)| \le \alpha\Big(\tilde{\Gamma}_t(S\cup U)\Big), \tag{10}$$

where $\alpha(\cdot)$ denotes the independence number of a graph.

Proof: Consider an arbitrary (x_K, y) realization. By the definition of the confusion graph, $\mathcal{X}_{S\cup U}(x_K, y)$ is an independent set in the induced subgraph $\Gamma_t(S\cup U)$. Consider a vertex subset $\mathcal{I}$ of $\mathcal{X}_{S\cup U}(x_K, y)$ such that for every $x_S \in \mathcal{X}_S(x_K, y)$, there is exactly one vertex $v_{S\cup U} \in \mathcal{I}$ satisfying $v_S = x_S$. Thus $|\mathcal{I}| = |\mathcal{X}_S(x_K, y)|$. Also note that since $\mathcal{I} \subseteq \mathcal{X}_{S\cup U}(x_K, y)$, $\mathcal{I}$ is also an independent set in $\Gamma_t(S\cup U)$. For a visualization of the construction of $\mathcal{I}$, see the schematic graph in Figure 3.

As $\mathcal{I}$ is also an independent set in $\Gamma_t(S\cup U)$, any two vertices in $\mathcal{I}$ are not confusable at any decoder $i \in S\cup U$ of the subproblem $\Gamma(S\cup U)$. Note that the extra decoder

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

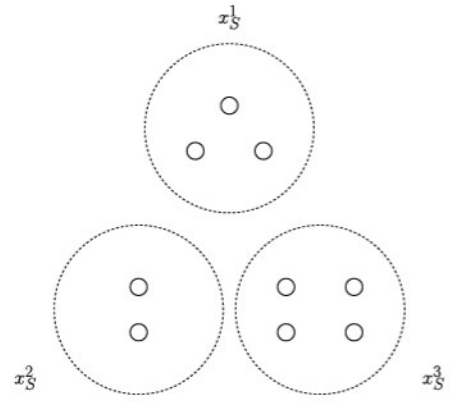

Fig. 3. Vertices in $\mathcal{X}_{S\cup U}(x_K, y)$ are denoted by nodes in the figure. As $\mathcal{X}_{S\cup U}(x_K, y)$ is an independent set in $\Gamma_t(S \cup U)$, there is no edge among the vertices in $\mathcal{X}_{S\cup U}(x_K, y)$. Note that the figure is for illustrative purpose only; there is no limit on the number of vertices in $\mathcal{X}_{S\cup U}(x_K, y)$. We partition the nodes into subgroups according to their x_S values, and each subgroup is denoted by a dashed circle with its corresponding x_S value marked beside it. To construct $\mathcal{I}$, one needs only to arbitrarily pick one node from each dashed circle. It is clear that $|\mathcal{X}_S(x_K, y)|$, denoting the number of distinct x_S values in $\mathcal{X}_{S\cup U}(x_K, y)$, is equal to the number of dashed circles in the graph, which is also equal to $|\mathcal{I}|$.

$m + 1$ knows X_S as side information. Since any two vertices in $\mathcal{I}$ have different x_S values by construction, they are not confusable at the extra decoder $m + 1$ either. In conclusion, any two vertices in $\mathcal{I}$ are not confusable at any decoders in the subproblem $\tilde{\Gamma}(S \cup U)$. In other words, any two vertices in $\mathcal{I}$ are not adjacent in $\tilde{\Gamma}_t(S \cup U)$, and hence $\mathcal{I}$ is also an independent set of $\tilde{\Gamma}_t(S \cup U)$. Therefore, we must have $|\mathcal{I}| \le \alpha(\tilde{\Gamma}_t(S \cup U))$. Recall that $|\mathcal{I}| = |\mathcal{X}_S(x_K, y)|$, and thus we have $|\mathcal{X}_S(x_K, y)| \le \alpha(\tilde{\Gamma}_t(S\cup U))$. Since (x_K, y) is arbitrary, we have (10). ∎

Proof of the Converse of Theorem 1: Consider any valid $(t, M, f, \mathbf{g})$ index code for the index coding problem Γ. We have

$$\begin{aligned}
\mathrm{P_s}(X_K, Y) &= \sum_{x_K,y} \max_{J\subseteq\mathcal{X}_S:|J|\le c(t)} \sum_{x_S\in J} P_{Y,X_{K,S}}\big(y, x_{K,S}\big) \\
&\overset{(a)}{=} \sum_{x_K,y} \max_{J\subseteq\mathcal{X}_S(x_K,y):|J|\le c(t)} \sum_{x_S\in J} P_{Y,X_{K,S}}\big(y, x_{K,S}\big) \\
&\overset{(b)}{\ge} \sum_{x_K,y} \frac{\sum_{J\subseteq\mathcal{X}_S(x_K,y):|J|=c(t)^-} \sum_{x_S\in J} P_{Y,X_{K,S}}\big(y, x_{K,S}\big)}{\big|\big\{J \subseteq \mathcal{X}_S(x_K, y) : |J| = c(t)^-\big\}\big|} \\
&\overset{(c)}{=} \sum_{x_K,y} \frac{\binom{|\mathcal{X}_S(x_K,y)|-1}{c(t)^- -1} \sum_{x_S\in\mathcal{X}_S(x_K,y)} P_{Y,X_{K,S}}\big(y, x_{K,S}\big)}{\binom{|\mathcal{X}_S(x_K,y)|}{c(t)^-}} \\
&= \sum_{x_K,y} \frac{c(t)^-}{|\mathcal{X}_S(x_K, y)|} \sum_{x_S\in\mathcal{X}_S(x_K,y)} P_{Y,X_{K,S}}\big(y, x_{K,S}\big) \\
&\overset{(d)}{\ge} \frac{c(t)}{\alpha\Big(\tilde{\Gamma}_t(S \cup U)\Big)} \sum_{x_K,y} \sum_{x_S\in\mathcal{X}_S(x_K,y)} P_{Y,X_{K,S}}\big(y, x_{K,S}\big) \\
&= \frac{c(t)}{\alpha\Big(\tilde{\Gamma}_t(S \cup U)\Big)},
\end{aligned} \tag{11}$$

where $c(t)^- = \min\{c(t), |\mathcal{X}_S(x_K, y)|\}$ and

- (a) follows from the fact that for any x_K, y, $P_{Y,X_{K,S}}(y, x_{K,S}) = 0$ for any $x_S \notin \mathcal{X}_S(x_K, y)$ by the definition in (4);
- (b) follows since the maximum is larger than average;
- (c) follows since each $x_S \in \mathcal{X}_S(x_K, y)$ appears in exactly $\binom{|\mathcal{X}_S(x_K,y)|-1}{c(t)^- -1}$ subsets of $\mathcal{X}_S(x_K, y)$ of size $c(t)^-$;
- (d) follows from Lemma 1 and the following arguments
  - if $c(t) \le |\mathcal{X}_S(x_K, y)|$, then $\frac{c(t)^-}{|\mathcal{X}_S(x_K,y)|} = \frac{c(t)}{|\mathcal{X}_S(x_K,y)|} \ge \frac{c(t)}{\alpha(\tilde{\Gamma}_t(S\cup U))}$,
  - otherwise we have $c(t) > |\mathcal{X}_S(x_K, y)|$ and $\frac{c(t)^-}{|\mathcal{X}_S(x_K,y)|} = 1 \ge \frac{c(t)}{\alpha(\tilde{\Gamma}_t(S\cup U))}$, where the last inequality is due to the assumption that $c(t) \le \max_{x_K,y} |\mathcal{X}_S(x_K, y)| \le \alpha(\tilde{\Gamma}_t(S \cup U))$ stated in Section II-C.

Finally, we have

$$\begin{aligned}
\mathcal{L}^*(\Gamma) &= \lim_{t\to\infty} \inf_{(t,M,f,\mathbf{g}) \text{ codes}} \frac{1}{t} \log_q \frac{\mathrm{P_s}(X_K, Y)}{\mathrm{P_s}(X_K)} \\
&\ge \lim_{t\to\infty} \frac{1}{t} \log_q \frac{\frac{c(t)}{\alpha\left(\tilde{\Gamma}_t(S\cup U)\right)}}{\frac{c(t)}{|\mathcal{X}|^{ts}}} \qquad (12) \\
&= \lim_{t\to\infty} \frac{1}{t} \log_q \frac{|\mathcal{X}|^{ts}}{\alpha\Big(\tilde{\Gamma}_t(S \cup U)\Big)} \\
&= \lim_{t\to\infty} \frac{1}{t} \log_q \frac{\big|V\Big(\tilde{\Gamma}_t(S \cup U)\Big)\big| \cdot |\mathcal{X}|^{-tu}}{\alpha\Big(\tilde{\Gamma}_t(S \cup U)\Big)} \qquad (13) \\
&= \lim_{t\to\infty} \frac{1}{t} \log_q \chi_{\mathrm{f}}\Big(\tilde{\Gamma}_t(S \cup U)\Big) - u \qquad (14) \\
&= \beta\Big(\tilde{\Gamma}(S \cup U)\Big) - u, \qquad (15)
\end{aligned}$$

where (12) follows since (11) holds for any valid $(t, M, f, \mathbf{g})$ code and $\mathrm{P_s}(X_K) = \frac{c(t)}{|\mathcal{X}|^{ts}}$, (13) follows from $|V(\tilde{\Gamma}_t(S \cup U))| = |\mathcal{X}|^{t(s+u)}$, (14) follows since confusion graphs are vertex-transitive [28], and (15) follows from (3). ∎

Proof of the Achievability of Theorem 1: We construct a randomized coding scheme based on the confusion graph.

Consider any sufficiently large t. Consider any independent set $\mathcal{I} \subseteq \mathcal{X}_{S\cup U}$ in $\tilde{\Gamma}_t(S \cup U)$ such that $|\mathcal{I}| = \alpha(\tilde{\Gamma}_t(S \cup U))$.

Notice that by adding decoder $m + 1$ to Γ to construct $\tilde{\Gamma}$, we are essentially adding more edges to the confusion graph (i.e., $E(\Gamma_t) \subseteq E(\tilde{\Gamma}_t)$). Because $\Gamma_t(S \cup U)$ has fewer edges than $\tilde{\Gamma}_t(S \cup U)$, we know that $\mathcal{I}$ is also an independent set of $\Gamma_t(S \cup U)$.

Consider a given $x_{S\cup U}$ value. For any $z_{S\cup U} \in \mathcal{I}$, we generate $\tilde{z}_{S\cup U}$ by adding $z_{S\cup U}$ and $x_{S\cup U}$ symbol-wise. Collecting all such $\tilde{z}_{S\cup U}$ generated from some $z_{S\cup U} \in \mathcal{I}$, we obtain another independent set in $\Gamma_t(S \cup U)$. We denote this set by $\tilde{\mathcal{I}}(x_{S\cup U})$ because it is generated from $\mathcal{I}$ and the given $x_{S\cup U}$ value. Clearly, the size of $\tilde{\mathcal{I}}(x_{S\cup U})$ is also $\alpha(\tilde{\Gamma}_t(S \cup U))$.

In this way, using all possible $x_{S\cup U}$ values, we can generate $|\mathcal{X}|^{t(s+u)}$ independent sets $\tilde{\mathcal{I}}(x_{S\cup U})$, each of size $\alpha(\tilde{\Gamma}_t(S\cup U))$. Moreover, it can be verified that every vertex in $\Gamma_t(S \cup U)$ appears in exactly $\alpha(\tilde{\Gamma}_t(S \cup U))$ generated independent sets.

For each $x_{[n]} = (x_K, x_S, x_U)$, suppose there is a unique codeword. For each value $z_{S\cup U}$ in the independent set $\tilde{\mathcal{I}}(x_{S\cup U})$,

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

the message tuple realization $(x_K, z_{S\cup U})$ is mapped to this corresponding codeword with probability $1/\alpha(\tilde{\Gamma}_t(S\cup U))$. In such a way, we construct a valid randomized mapping scheme since no confusable realizations are mapped to the same codeword and thus each decoder $i \in [m]$ can decode its wanted messages.

With the coding scheme described above, the leakage can be computed as

$$\begin{aligned}
L &= \frac{1}{t}\log_q\Bigg[\Bigg(\sum_{x_K,y}\max_{J\subseteq\mathcal{X}_S(x_K,y):|J|\le c(t)} \sum_{x_S\in J} P_{Y,X_{K\cup S}}(y, x_{K\cup S})\Bigg)\cdot\frac{|\mathcal{X}|^{ts}}{c(t)}\Bigg] \\
&= \frac{1}{t}\log_q\Bigg[\Bigg(\sum_{x_K,y}\max_{J\subseteq\mathcal{X}_S(x_K,y):|J|\le c(t)}\sum_{x_S\in J} \sum_{x_U\in\mathcal{X}_U(x_{K\cup S},y)} P_{Y|X_{[n]}}\big(y|x_{[n]}\big)\cdot P_{X_{[n]}}\big(x_{[n]}\big)\Bigg)\cdot\frac{|\mathcal{X}|^{ts}}{c(t)}\Bigg] \\
&\overset{(a)}{=} \frac{1}{t}\log_q\Big(|\mathcal{Y}|\cdot c(t)\cdot\frac{1}{\alpha\Big(\tilde{\Gamma}_t(S\cup U)\Big)}\cdot\frac{1}{|\mathcal{X}|^{tn}}\cdot\frac{|\mathcal{X}|^{ts}}{c(t)}\Big) \\
&\overset{(b)}{=} \frac{1}{t}\log_q\frac{|\mathcal{X}|^{ts}}{\alpha\Big(\tilde{\Gamma}_t(S\cup U)\Big)} \\
&\overset{(c)}{=} \beta\Big(\tilde{\Gamma}(S\cup U)\Big) - u,
\end{aligned}$$

where (a) follows since for every y there is a unique x_K value by the coding scheme construction, and that $P_{Y|X_{[n]}}(y|x_{[n]}) = \frac{1}{\alpha(\tilde{\Gamma}_t(S\cup U))}$ for jointly possible y, x_K, x_S, and x_U, (b) follows since $|\mathcal{Y}| = |\mathcal{X}_{[n]}| = |\mathcal{X}|^{tn}$, and (c) follows from the same arguments as in (13)-(15) as t is sufficiently large. This completes the proof of the achievability of the theorem. ■

Remark 5: It can be shown using the generalized maximal acyclic induced subgraph (MAIS) bound [3] that the result in Theorem 1 is always non-negative.

Remark 6: It is worth noticing that the actual characterization of $\mathcal{L}^*$ is independent of the number of guesses the adversary can make as specified by the function c.

Example 1: Consider the toy example in Figure 1, where $n = m = 4$, $W_i = i, i \in [4]$, $A_1 = \{2,3\}$, $A_2 = \{1,4\}$, $A_3 = \{1,4\}$, $A_4 = \{2,3\}$, and $K = \emptyset$, $S = \{1,2\}$, $U = \{3,4\}$. We have $\mathcal{L}^*(\Gamma) = \beta(\tilde{\Gamma}(S\cup U)) - u = \beta(\tilde{\Gamma}) - 2$. To find $\beta(\tilde{\Gamma})$, we first lower bound it using the generalized MAIS bound [3], which gives $\beta(\tilde{\Gamma}) \ge |\{1,4\}| = 2$ as decoders 1 and 4 do not know each other's wanted message as side information. On the other hand, the coding scheme $\tilde{Y} = (X_1 \oplus X_3, X_2 \oplus X_4)$ can satisfy the decoding requirement at every decoder, including the extra decoder $i = m+1 = 5$ who knows $X_S = X_{1,2}$ and wants to decode $X_U = X_{3,4}$. This indicates an upper bound of 2 on $\beta(\tilde{\Gamma})$, which, combined with the lower bound, gives $\beta(\tilde{\Gamma}) = 2$ and thus $\mathcal{L}^*(\Gamma) = 2 - 2 = 0$.

Remark 7: An intuition behind the characterization of $\mathcal{L}^*(\Gamma)$ using the broadcast rate of a related problem $\tilde{\Gamma}$ is as follows. For simplicity of exposition, we assume $K = \emptyset$, $[n] = S\cup U$, and Theorem 1 states that $\mathcal{L}^*(\Gamma) = \beta(\tilde{\Gamma}) - u$. At first glance, some sort of connection between data compression and secrecy preservation for a given problem seems expected, as compression removes redundancy from the source and thus less information is exposed to the adversary. In particular, with a guessing adversary, better compression in general means more message realizations being mapped to each codeword, making the adversary less likely to guess the true realization from its observation of the codeword. However, with a more careful look, the goal of data compression is to compress all the messages in the system to satisfy the decoders, while the goal of secrecy preservation is to protect the *sensitive* messages from the adversary. That is, for compression, the sender wants to map as many $x_{[n]}$ values as possible to each y, while for reducing information leakage, the sender needs to map more x_S to each y. These two goals do not necessarily perfectly align for a given problem Γ. Therefore, we establish connections between the compression and leakage by constructing an auxiliary problem $\tilde{\Gamma}$, for which the two goals do align. For $\tilde{\Gamma}$, the extra decoder requires that X_U must be a deterministic function of (Y, X_S). Consequently, for a given y, the number of $x_{[n]} = (x_S, x_U)$ being mapped to it equals to the number of x_S being mapped to it. This means that the compression goal becomes equivalent to the secrecy preserving goal for $\tilde{\Gamma}$, which makes the characterization of $\mathcal{L}^*(\tilde{\Gamma})$ using $\beta(\tilde{\Gamma})$ intuitively understandable. As a final step, one needs to link $\mathcal{L}^*(\tilde{\Gamma})$ back to $\mathcal{L}^*(\Gamma)$, which turn out to be equal. The intuition behind this equivalence comes from the fact that the existence of the extra decoder does not impose any extra decoding constraint on the number of x_S being mapped to each y, since it does not require any message in X_S to be decoded.

Given Theorem 1, various bounds and properties on β established in the literature [27] can be directly applied to $\mathcal{L}^*$. Arbabjolfaei and Kim [32] showed that the broadcast rate of an index coding instance can be computed from the broadcast rates of its subproblems if certain structural properties hold among the decoders' side information sets. These properties were proved using confusion graphs in the scope of *unicast* index coding, where every decoder wants one unique message. We can apply similar techniques to our *multicast* problem setup, and simplify the characterization of $\mathcal{L}^*$ in Theorem 1 when certain structural properties hold.

In the following we consider the case when the set of decoders can be separated into two parts, one part requesting to decode messages within set $X_{S\cup K}$ and the other part requesting to decode messages within set $X_{U\cup K}$, and the decoders in the former part know nothing in X_U as side information. Intuitively speaking, in such a case, the non-sensitive messages X_U are not known by the decoders requesting sensitive messages in X_S, and thus X_U become "useless" as they cannot be effectively used like secret keys to be combined with X_S since doing so violates the decoding requirements. For a visualizing example, see Figure 4(a). In the figure, we consider unicast index coding problems where $n = m = 4$ and decoder i wants message i for every $i \in [4]$. With this unicast condition, the problem can be represented by a directed graph, commonly known as the side information graph [27, Sec. 1.2]. Every node in the graph denotes a unique message and an edge from node i to j means that decoder j knows message X_i as side information.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

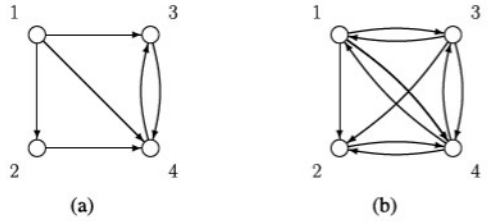

Fig. 4. Visualizing examples for the two special scenarios under consideration: (a) the side information graph of a four-message four-decoder unicast index coding problem, where decoders requesting messages in $X_S = X_{1,2}$ know nothing in $X_U = X_{3,4}$ as side information; (b) the side information graph of another four-message four-decoder unicast index coding problem, where decoders requesting messages in $X_S = X_{1,2}$ know everything in $X_U = X_{3,4}$ as side information, and vice versa.

In Figure 4(a), suppose $S = \{1, 2\}$ and $U = \{3, 4\}$. It can be seen that neither decoder 1 nor 2 knows $X_{3,4}$.

Corollary 1: Consider any problem Γ such that every decoder $i \in [m]$ satisfies either $W_i \subseteq U \cup K$ or $W_i \cup A_i \subseteq S \cup K$. We have

$$\mathcal{L}^*(\Gamma) = \beta(\Gamma(S)). \tag{16}$$

Proof: Consider the problem $\tilde{\Gamma}(S \cup U)$. One way is to split the problem into two subproblems $\tilde{\Gamma}(S)$ and $\tilde{\Gamma}(U)$, and separately compress the messages in S and U. Thus, we have

$$\beta\Big(\tilde{\Gamma}(S \cup U)\Big) \leq \beta\Big(\tilde{\Gamma}(S)\Big) + \beta\Big(\tilde{\Gamma}(U)\Big) \overset{(a)}{=} \beta(\Gamma(S)) + u,$$

where (a) follows from $\beta(\tilde{\Gamma}(S)) = \beta(\Gamma(S))$ since the extra decoder $m+1$ is a degenerate decoder who wants nothing in $\tilde{\Gamma}(S)$, and $\beta(\tilde{\Gamma}(U)) = u$ since the extra decoder $m+1$ needs to decode all the messages in X_U and has no side information at all in $\tilde{\Gamma}(U)$. In the following, we show that $\beta(\tilde{\Gamma}(S \cup U)) \geq \beta(\Gamma(S)) + u$.

The decoders in $\tilde{\Gamma}(S \cup U)$ can be divided into two parts, one consisting of those requesting messages within X_U and the other consisting of those requesting messages within X_S. We simply call the former the U part and the latter the S part. We add some side information to the decoders in the U part in $\tilde{\Gamma}(S \cup U)$ to construct another problem Γ' with messages $X_{S \cup U}$ such that every decoder in the U part knows X_S. Clearly, $\beta(\tilde{\Gamma}(S \cup U)) \geq \beta(\Gamma')$. In Γ', every decoder in the U part knows all X_S as side information, and no decoder in the S part knows anything in X_U as side information as specified in the corollary statement. Hence, for any t, any two realizations $x_{S \cup U}, z_{S \cup U} \in \mathcal{X}_{S \cup U}$ are confusable iff x_S and z_S are confusable for $\Gamma'_t(S)$ or $x_S = z_S$ and x_U and z_U are confusable for $\Gamma'_t(U)$. Therefore, the confusion graph Γ'_t is the lexicographic product [33] of the two confusion graphs $\Gamma'_t(S)$ and $\Gamma'_t(U)$. Then, we have [28, Cor. 3.4.5]

$$\chi_{\mathrm{f}}\big(\Gamma'_t\big) = \chi_{\mathrm{f}}\big(\Gamma'_t(S)\big) \cdot \chi_{\mathrm{f}}\big(\Gamma'_t(U)\big).$$

Combining the above result and (3), we have

$$\begin{aligned}
\beta\Big(\tilde{\Gamma}(S \cup U)\Big) &\geq \beta\big(\Gamma'\big) \\
&= \lim_{t \to \infty} \frac{1}{t} \log_q \chi_{\mathrm{f}}\big(\Gamma'_t\big) \\
&= \lim_{t \to \infty} \frac{1}{t} \Big(\log_q \chi_{\mathrm{f}}\big(\Gamma'_t(S)\big) + \log_q \chi_{\mathrm{f}}\big(\Gamma'_t(U)\big) \Big) \\
&= \beta\big(\Gamma'(S)\big) + \beta\big(\Gamma'(U)\big) \\
&= \beta\Big(\tilde{\Gamma}(S)\Big) + \beta\Big(\tilde{\Gamma}(U)\Big) \\
&= \beta(\Gamma(S)) + u.
\end{aligned}$$

Therefore, we have proved $\beta(\tilde{\Gamma}(S \cup U)) = \beta(\Gamma(S)) + u$, and thus by Theorem 1 we have $\mathcal{L}^*(\Gamma) = \beta(\tilde{\Gamma}(S \cup U)) - u = \beta(\Gamma(S)) + u - u = \beta(\Gamma(S))$. ■

Remark 8: An extreme case of the scenario considered in Corollary 1 is when there are no non-sensitive messages in the system (i.e., $U = \emptyset$). In such a case, the extra decoder in $\tilde{\Gamma}(S \cup U)$ becomes degenerate as it knows all the messages in the system and wants nothing, and the result in Corollary 1 reduces to that in [19, Corollary 1].

Consider another scenario when the set of decoders can again be separated into two parts, one part requesting messages within $X_{S \cup K}$ while knowing all X_U, and the other part requesting messages within $X_{U \cup K}$ while knowing all X_S. In such a case, the non-sensitive messages X_U can be exploited to the maximum extent, since they can be "safely" combined with the sensitive messages X_S without worrying about the decoding requirements being violated. A visualizing example can be found in Figure 4(b). Note that a special case of this scenario is when X_U are common random keys shared among the encoder and all the decoders.

Corollary 2: Consider any problem Γ such that every decoder $i \in [m]$ satisfies one of the following two conditions: 1) $W_i \subseteq U \cup K$ and $S \subseteq A_i$; 2) $W_i \subseteq S \cup K$ and $U \subseteq A_i$. We have

$$\mathcal{L}^*(\Gamma) = \max\{\beta(\Gamma(S)) - u, 0\}. \tag{17}$$

Proof: We call the set of decoders in $\tilde{\Gamma}(S \cup U)$ requesting messages within X_U the U part, and the rest of decoders the S part. Decoder $m+1$ belongs to the U part. Every decoder in the U part knows X_S as side information, and every decoder in the S part knows X_U as side information. Then, for any t, any two realizations $x_{S \cup U}, z_{S \cup U} \in \mathcal{X}_{S \cup U}$ are confusable iff one of the following conditions is satisfied: 1) $x_U = z_U$ and x_S and z_S are confusable in $\tilde{\Gamma}_t(S)$; 2) $x_S = z_S$ and x_U and z_U are confusable in $\tilde{\Gamma}_t(U)$. Therefore, the confusion graph $\tilde{\Gamma}_t(S \cup U)$ is the Cartesian product [28] of the two confusion graphs $\tilde{\Gamma}_t(S)$ and $\tilde{\Gamma}_t(U)$. Hence, we have [34, Lemma 2.6] $\chi_{\mathrm{f}}(\tilde{\Gamma}_t(S \cup U)) = \max\{\chi_{\mathrm{f}}(\tilde{\Gamma}_t(S)), \chi_{\mathrm{f}}(\tilde{\Gamma}_t(U))\}$, which together with (3) leads to $\beta(\tilde{\Gamma}(S \cup U)) = \max\{\beta(\tilde{\Gamma}(S)), \beta(\tilde{\Gamma}(U))\}$. Therefore, by Theorem 1, we have $\mathcal{L}^*(\Gamma) = \beta(\tilde{\Gamma}(S \cup U)) - u = \max\{\beta(\tilde{\Gamma}(S)) - u, \beta(\tilde{\Gamma}(U)) - u\} = \max\{\beta(\Gamma(S)) - u, 0\}$. ■

IV. A Deterministic Linear Index Code

In the following, we construct a deterministic linear index code based on the minrank method and fitting matrices, which were developed for the original index coding problem without adversary [3]. We then show that the proposed scheme achieves the optimal leakage rate over all valid deterministic scalar linear index codes. Throughout the section, we set $t = 1$ as we are considering only scalar linear codes. However, note that the message alphabet size q can be arbitrary.

Unless otherwise stated, we use bold-faced capital letters to denote matrices and vectors. In particular, $\mathbf{X}_{[n]} =$

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

$[X_1\ X_2 \dots\ X_n]^T$. Let $r(\cdot)$ denote the rank of a matrix over the Galois field $\mathbb{F}_q$.

Note that for any decoder i requiring more than one message (i.e., $|W_i| \geq 2$), we can transform the problem into an equivalent problem by removing decoder i and adding $|W_i|$ new decoders, where every new decoder has the same side information set A_i and each decoder wants a unique message in the set W_i. Therefore, we can, without loss of generality, always assume $W_i = \{w_i\}$ is a singleton set for every decoder $i \in [m]$.

For any given problem, a size $m \times n$ fitting matrix $\mathbf{M}$ with elements in Galois field $\mathbb{F}_q$ is a matrix such that for any decoder $i \in [m]$,

$$\begin{aligned} &\mathbf{M}_{ij} = 1, \quad \text{for } j = w_i, \\ &\mathbf{M}_{ij} = 0, \quad \text{for any message } j \in [n] \backslash (W_i \cup A_i). \end{aligned}$$

As $\mathbf{M}_{ij}$ can be any element in $\mathbb{F}_q$ if $j \in A_i$, there can be multiple fitting matrices for a given problem. For example, for the problem with $n = 5$ messages and $m = 3$ decoders, where $W_i = \{i\}, \forall i \in [3]$, $A_1 = \{3, 4\}$, $A_2 = \{1, 4, 5\}$, $A_3 = \{2, 5\}$, any matrix of the following form is a fitting matrix,

$$\begin{bmatrix} 1 & 0 & ? & ? & 0 \\ ? & 1 & 0 & ? & ? \\ 0 & ? & 1 & 0 & ? \end{bmatrix}$$

where "?" can be any (and different) elements $\mathbb{F}_q$.

If the encoder generates a codeword $\mathbf{Y}$ by multiplying a fitting matrix $\mathbf{M}$ by the message vector $\mathbf{X}_{[n]}$, every decoder $i \in [m]$ can recover its wanted message because the i-th element of $\mathbf{Y}$ is a linear combination of only X_{w_i} and some of its side information. Moreover, note that any row of $\mathbf{M}$ can be generated by $r(\mathbf{M})$ independent rows of $\mathbf{M}$. Thus, to satisfy the decoders, the encoder needs only to generate a codeword Y that contains the linear combinations of the messages with coefficients from $r(\mathbf{M})$ independent rows of $\mathbf{M}$. In this way, for a given problem, the minimum rank over all the fitting matrices establishes an upper bound on its broadcast rate, which has been proved to be optimal over all deterministic scalar linear codes [3].

For analysis of information leakage based on the fitting matrix framework, we can split $\mathbf{M}$ into three submatrices formed by different groups of columns in $\mathbf{M}$ according to sets K, S, and U. For brevity, we simply write

$$\mathbf{M} = [\mathbf{K} \quad \mathbf{S} \quad \mathbf{U}].$$

The following theorem characterizes the minimal leakage rate among all codes based on the fitting matrix framework. Moreover, we show that this leakage rate is in fact optimal for all deterministic scalar linear codes.

Theorem 2: For any index coding problem, there exists a deterministic scalar linear index code that yields the following leakage rate,

$$\mathcal{L} = \min_{\mathbf{M}} (r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U})). \tag{18}$$

Furthermore, this result is leakage-wise rate optimal for all deterministic scalar linear codes.

Proof: We first show the achievability. Consider any fitting matrix $\mathbf{M} = [\mathbf{K} \quad \mathbf{S} \quad \mathbf{U}]$ and any encoding matrix $\mathbf{E}$ formed by a set of row vectors of $\mathbf{M}$ such that the row space of $\mathbf{M}$ is the same as that of $\mathbf{E}$ and thus by observing $\mathbf{Y} = \mathbf{EX}$ every decoder can decode its wanted message.

Recall that $t = 1$ as we are considering scalar linear codes. Moreover, for every (x_K, y), $|\mathcal{X}_S(x_K, y)|$ remains the same since the code is linear, and thus we always have $c(1) \leq |\mathcal{X}_S(x_K, y)|$.

We have

$$\begin{aligned} \mathcal{L} &= \log_q \Bigg(\Bigg(\sum_{x_K, y} \max_{J \in \mathcal{X}_S(x_K, y): |J| \leq c(1)} \\ &\qquad \sum_{x_S \in J} P_{Y, X_{K \cup S}}(y, x_{K,S}) \Bigg) \cdot \frac{|\mathcal{X}|^s}{c(1)} \Bigg) \\ &= \log_q \Bigg(\frac{|\mathcal{X}|^s}{c(1)} \cdot \Bigg(\sum_{x_K, y} \max_{J \subseteq \mathcal{X}_S(x_K, y): |J| \leq c(1)} \\ &\qquad \sum_{x_S \in J} \sum_{x_U \in \mathcal{X}_U(x_{K \cup S}, y)} P_{Y|X_{[n]}}(y|x_{[n]}) \cdot P_{X_{[n]}}(x_{[n]}) \Bigg) \Bigg) \\ &\overset{(a)}{=} \log_q \Bigg(\frac{|\mathcal{X}|^s}{c(1)} \cdot \Bigg(\sum_{x_K, y} \max_{J \subseteq \mathcal{X}_S(x_K, y): |J| \leq c(1)} \\ &\qquad \sum_{x_S \in J} \sum_{x_U} \mathbb{1}(\mathbf{Y} = \mathbf{EX}) \cdot \frac{1}{|\mathcal{X}|^n} \Bigg) \Bigg) \\ &\overset{(b)}{=} \log_q \Bigg(\frac{|\mathcal{X}|^s}{c(1)} \cdot c(1) \cdot \frac{1}{|\mathcal{X}|^n} \cdot \\ &\qquad \Bigg(\sum_{x_K, y} \max_{x_S} \sum_{x_U} \mathbb{1}(\mathbf{Y} = \mathbf{EX}) \Bigg) \Bigg) \\ &= \log_q \Bigg(q^{-k-u} \sum_{x_K} \sum_{y} \max_{x_S} \sum_{x_U} \mathbb{1}(\mathbf{Y} = \mathbf{EX}) \Bigg) \\ &\overset{(c)}{=} \log_q \Bigg(q^{-k-u} \sum_{x_K} \sum_{y} q^{u - r(\mathbf{U})} \Bigg) \\ &\overset{(d)}{=} r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U}), \end{aligned}$$

where (a) is due to the code being deterministic, where $\mathbb{1}(\cdot)$ is the indicator function, (b) follows from the fact that we have $c(1) \leq |\mathcal{X}(x_K, y)|$ for every x_K, y, (c) follows from the fact that given any fixed y and $x_{K \cup S}$, there are $q^{u - r(\mathbf{U})}$ possible x_U values that satisfy $\mathbf{Y} = \mathbf{EX}$, and (d) follows from the fact that given any fixed x_K, there are $q^{r([\mathbf{S} \quad \mathbf{U}])}$ possible y values. Therefore, by minimizing over all fitting matrices, the leakage rate in (18) can be achieved.

Now we prove the converse part of the theorem. Suppose a $\ell \times n$ matrix $\tilde{\mathbf{E}}$ of Galois field $\mathbb{F}_q$ is the encoding matrix of an arbitrary valid deterministic scalar linear index code. Note that $\tilde{\mathbf{E}}$ need not be a fitting matrix. We split $\tilde{\mathbf{E}}$ into three submatrices formed by different groups of columns according to sets K, S, and U as

$$\tilde{\mathbf{E}} = \begin{bmatrix} \tilde{\mathbf{K}} & \tilde{\mathbf{S}} & \tilde{\mathbf{U}} \end{bmatrix}.$$

Following a similar argument as in the achievability proof of the theorem, we can show that the leakage rate caused by the

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

codeword $\mathbf{Y} = \tilde{\mathbf{E}}\mathbf{X}$ is

$$\mathcal{L}_{\tilde{\mathbf{E}}} = r\left(\begin{bmatrix}\tilde{\mathbf{S}} & \tilde{\mathbf{U}}\end{bmatrix}\right) - r\left(\tilde{\mathbf{U}}\right).$$

It remains to show that $\mathcal{L}_{\tilde{\mathbf{E}}}$ is lower bounded by (18).

According to the proof of [3, Th. 1], there exists some fitting matrix $\mathbf{M} = [\mathbf{K} \quad \mathbf{S} \quad \mathbf{U}]$ of the problem such that the row vectors of $\mathbf{M}$ lie in the row space of $\tilde{\mathbf{E}}$. In other words, there exists some $n \times \ell$ matrix $\mathbf{B}$ such that $\mathbf{B}\tilde{\mathbf{E}} = \mathbf{M}$, or, if we only consider the submatrices according to sets S and U,

$$\mathbf{B}\begin{bmatrix}\tilde{\mathbf{S}} & \tilde{\mathbf{U}}\end{bmatrix} = [\mathbf{S} \quad \mathbf{U}].$$

Also, there exists some matrix $\mathbf{D}$ such that

$$\tilde{\mathbf{U}} = \begin{bmatrix}\tilde{\mathbf{S}} & \tilde{\mathbf{U}}\end{bmatrix}\mathbf{D}, \qquad \mathbf{U} = [\mathbf{S} \quad \mathbf{U}]\mathbf{D} = \mathbf{B}\begin{bmatrix}\tilde{\mathbf{S}} & \tilde{\mathbf{U}}\end{bmatrix}\mathbf{D},$$

and hence,

$$\begin{aligned} r([\mathbf{S} \quad \mathbf{U}]) + r\left(\tilde{\mathbf{U}}\right) &= r\left(\mathbf{B}\begin{bmatrix}\tilde{\mathbf{S}} & \tilde{\mathbf{U}}\end{bmatrix}\right) + r\left(\begin{bmatrix}\tilde{\mathbf{S}} & \tilde{\mathbf{U}}\end{bmatrix}\mathbf{D}\right) \\ &\overset{(a)}{\leq} r\left(\begin{bmatrix}\tilde{\mathbf{S}} & \tilde{\mathbf{U}}\end{bmatrix}\right) + r\left(\mathbf{B}\begin{bmatrix}\tilde{\mathbf{S}} & \tilde{\mathbf{U}}\end{bmatrix}\mathbf{D}\right) = r\left(\begin{bmatrix}\tilde{\mathbf{S}} & \tilde{\mathbf{U}}\end{bmatrix}\right) + r(\mathbf{U}), \end{aligned}$$

where (a) is due to Frobenius inequality [35]. By reorganizing the above result, we have

$$\mathcal{L}_{\tilde{\mathbf{E}}} = r\left(\begin{bmatrix}\tilde{\mathbf{S}} & \tilde{\mathbf{U}}\end{bmatrix}\right) - r\left(\tilde{\mathbf{U}}\right) \geq r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U}),$$

which, together with the fact that $\mathbf{M}$ is a fitting matrix, completes the proof. ∎

Remark 9: Following similar arguments as in the proof of Theorem 2, we can show that when the mutual information[2] $I(X_S; Y|X_K)$ is used as the leakage metric, $\min_{\mathbf{M}}(r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U}))$ still characterizes the minimal leakage rate over all deterministic scalar linear index codes.

The following example shows the efficacy of the proposed linear coding scheme.

Example 2: Consider the problem Γ with $n = 5$ binary messages and $m = 5$ decoders with $W_i = \{i\}, i \in [m]$ and

$$A_1 = \{4, 5\}, A_2 = \{1\}, A_3 = \{2\}, A_4 = \{3\}, A_5 = \{4\}.$$

Assume for the adversary,

$$K = \{5\}, \quad S = \{1, 3\}, \quad U = \{2, 4\}.$$

The fitting matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{19}$$

achieves the broadcast rate of $\beta = r(\mathbf{M}) = 4$. For the leakage rate, we have

$$\mathcal{L} = r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U}) = 3 - 2 = 1,$$

which is indeed the optimal leakage rate as can be verified by Theorem 1. Therefore, the linear code given by (19) is optimal in both leakage and compression senses for this problem.

[2]Note that mutual information between sensitive variables and codeword has been commonly used as a leakage metric in the literature of information-theoretic secrecy [17], [36], [37], [38], [39].

As discussed in Remark 7, for an arbitrary problem, the goals of compression and leakage prevention do not necessarily align. In the following we show by a simple two-decoder example that it is not always possible to simultaneously achieve the optimal compression and leakage rates.

Example 3: Consider the problem Γ with $n = 4$ binary messages and $m = 2$ decoders, where

$$W_1 = \{1\}, W_2 = \{2\}, A_1 = \{2, 3\}, A_2 = \{1, 4\}.$$

Assume for the adversary,

$$K = \emptyset, \qquad S = \{1, 2\}, \qquad U = \{3, 4\}.$$

The fitting matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

achieves the broadcast rate of $\beta = r(\mathbf{M}) = 1$ while resulting in a leakage rate of $\mathcal{L} = r(\mathbf{M}) - r(\mathbf{U}) = 1 - 0 = 1$. The following fitting matrix

$$\tilde{\mathbf{M}} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

gives $\mathcal{L} = r(\tilde{\mathbf{M}}) - r(\tilde{\mathbf{U}}) = 2 - 2 = 0$, indicating that zero leakage (i.e., perfect secrecy) can be achieved for the problem. However, $\tilde{\mathbf{M}}$ leads to a suboptimal compression rate of $r(\tilde{\mathbf{M}}) = 2$. In fact, it can be shown using Shannon-type inequalities [40] that the compression rate of any index code that attains zero leakage is at least 2, implying that $\beta = 1$ and $\mathcal{L}^* = 0$ can never be simultaneously achieved for this problem.

Remark 10: Although not always optimal, the fitting matrix based coding scheme can achieve good performance for both compression and leakage rates for some index coding instances. On the other hand, the randomized coding scheme described in the proof of Theorem 1 always achieves optimal leakage rate, at the cost of a rather high compression rate of n. Hence, it is natural to ask if we could use time-sharing between these two schemes (or any possible schemes) to achieve a better trade-off between compression and leakage rates. Indeed, it can be verified that the time-sharing property (either deterministic or stochastic) holds not only for the compression rate R but also for the leakage rate $\mathcal{L}$. Therefore, time-sharing between different schemes to balance the compression and leakage rates is possible.

V. Extensions

In this section, we study information leakage in two varied settings. In Section V-A, we consider the scenario where a rate constraint is imposed on the compression, and information leakage is to be minimized under this constraint. In Section V-B, we look at the pliable index coding problem [26], and show how we can establish results similar to those we derive for index coding by utilizing the connection between the two problems.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

A. Minrank Coding Scheme Under a Compression Rate Constraint

In this scenario, the encoder attempts to minimize the information leakage rate subject to a maximum permitted compression rate. We formalize the problem as finding

$$\mathcal{L}^*_{R_\Delta} \doteq \lim_{t\to\infty} t^{-1} \inf_{(t,M,f,\mathbf{g}) \text{ codes s.t. } t^{-1}\log_q M \le R_\Delta} L.$$

for some given compression rate threshold R_Δ.

Note that the parameter R_Δ should always satisfy $R_\Delta \ge \beta$, since otherwise there will be no index codes that satisfy both the decoding requirements and the compression rate constraint. On the other hand, we can assume without loss of generality that $R_\Delta \le n$. Because otherwise the randomized achievable scheme described in the achievability proof of Theorem 1, which is of compression rate n, can be applied, and thus $\mathcal{L}^*_\Delta = \mathcal{L}^*$.

While a characterization of $\mathcal{L}^*_{R_\Delta}$ in general remains to be investigated, the fitting-matrix-based coding scheme can be extended to suit this case in a straightforward manner.

Proposition 1: For any index coding problem, there exists a deterministic scalar linear $(t, M, f, \mathbf{g})$ index code that satisfies the compression rate constraint $t^{-1}\log_q M \le R_\Delta$ and yields the following leakage rate,

$$\mathcal{L} = \min_{\mathbf{M}: r(\mathbf{M}) \le R_\Delta} (r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U})). \tag{20}$$

Furthermore, this result is leakage-wise rate optimal under the compression rate constraint for all deterministic scalar linear codes.

Proof: To show the achievability, simply consider any fitting matrix $\mathbf{M} = [\mathbf{K} \quad \mathbf{S} \quad \mathbf{U}]$ such that $r(\mathbf{M}) \le R_\Delta$. Then by similar arguments in the achievability proof of Theorem 2, it can be shown that the information leakage caused by $\mathbf{M}$ is $\mathcal{L} = r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U})$. Therefore, by minimizing over all fitting matrices $\mathbf{M}$ satisfying $r(\mathbf{M}) \le R_\Delta$, the leakage rate in (20) can be achieved.

To show the converse part of the proposition, consider an arbitrary valid deterministic scalar linear index code satisfying the compression rate constraint, whose encoding matrix is denoted by matrix $\tilde{\mathbf{E}}$ of Galois field $\mathbb{F}_q$. Thus, $r(\tilde{\mathbf{E}}) \le R_\Delta$.

According to the proof of [3, Th. 1], there exists some fitting matrix $\mathbf{M} = [\mathbf{K} \quad \mathbf{S} \quad \mathbf{U}]$ of the problem such that the row vectors of $\mathbf{M}$ lie in the row space of $\tilde{\mathbf{E}}$. Hence, we have

$$r(\mathbf{M}) \le r\left(\tilde{\mathbf{E}}\right) \le R_\Delta.$$

Using similar arguments as in the proof of the converse part of Theorem 2, it can be shown that the leakage rate caused by $\tilde{\mathbf{E}}$ is lower bounded as

$$\mathcal{L}_{\tilde{\mathbf{E}}} \ge r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U}),$$

which, together with the fact that $\mathbf{M}$ is a fitting matrix such that $r(\mathbf{M}) \le r(\tilde{\mathbf{E}}) \le R_\Delta$, completes the proof. ■

B. Pliable Index Coding

In some applications, each decoder may be interested in decoding any message it does not know as side information. This *pliable* version of index coding was first formalized and studied by Brahma and Fragouli [26], and then further investigated in a number of subsequent works [41], [42], [43], [44], [45], [46]. Technically speaking, the major difference between pliable index coding and index coding is that for the pliable version, the desired message sets at the decoders are not pre-determined and each decoder is satisfied whenever it can decode some messages not in its side information set. That is, the encoder can encode based on its own choice of desired message sets for the decoders, and this flexibility leads to more encoding opportunities to possibly achieve a lower compression rate and less information leakage.

Any pliable index coding instance can be denoted by a parameter tuple $\Pi = (n, m, (A_i, i \in [m]))$. While the encoder has the flexibility to choose the decoding message tuple for the decoders, once a decoding message tuple $(W_i, i \in [m])$ is chosen, the remaining problem becomes a normal index coding problem.[3] The broadcast rate (i.e., the optimal compression rate) of Π can be defined as

$$\beta(\Pi) \doteq \min_{(W_i, i\in[m]): W_i \subseteq ([n]\setminus A_i), i\in[m]} \beta\big((n, m, (A_i, i \in [m]), (W_i, i \in [m]))\big),$$

which can also be computed from the confusion graph perspective as

$$\begin{aligned}\beta(\Pi) &= \min_{\substack{(W_i, i\in[m]):\\ W_i \subseteq ([n]\setminus A_i), i\in[m]}} \lim_{t\to\infty} \frac{1}{t}\log_q \chi(\Gamma_t) \\ &= \min_{\substack{(W_i, i\in[m]):\\ W_i \subseteq ([n]\setminus A_i), i\in[m]}} \lim_{t\to\infty} \frac{1}{t}\log_q \chi_{\mathrm{f}}(\Gamma_t),\end{aligned}$$

where Γ_t denotes the confusion graph corresponding to the index coding problem $(n, m, (A_i, i \in [m]), (W_i, i \in [m]))$ with message length t.

Our goal is to characterize the optimal information leakage rate to a guessing adversary in pliable index coding, which we can simply define as

$$\mathcal{L}^*(\Pi) \doteq \min_{(W_i, i\in[m]): W_i \subseteq ([n]\setminus A_i), i\in[m]} \mathcal{L}^*\big((n, m, (A_i, i \in [m]), (W_i, i \in [m]))\big).$$

The above definitions imply that we can characterize $\mathcal{L}^*(\Pi)$ using our established results on $\mathcal{L}^*(\Gamma)$, minimized over all possible index coding problems Γ generated from Π and some decoding message tuple $(W_i, i \in [m])$. We have the following results analogous to Theorems 1 and 2.

Proposition 2: For any pliable index coding problem Π, we have

$$\mathcal{L}^*(\Pi) = \min_{(W_i, i\in[m]): W_i \subseteq ([n]\setminus A_i), i\in[m]} \beta\left(\tilde{\Gamma}(S \cup U)\right) - u, \tag{21}$$

where, given a decoding message tuple $(W_i, i \in [m])$, $\tilde{\Gamma}$ denotes the index coding problem constructed by adding an extra decoder to the index coding problem $(n, m, (W_i, i \in [m]), (A_i, i \in [m]))$. The extra decoder is indexed by $m + 1$

[3] Without loss of generality, we can assume that $|W_i| = 1$ for every $i \in [m]$.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

and knows side information indexed by the set $A_{m+1} = K \cup S$ and wants messages indexed by the set $W_{m+1} = U$.[4]

Proof: Combining the definition of $\mathcal{L}^*(\Pi)$ and Theorem 1 directly gives

$$\begin{aligned}\mathcal{L}^*(\Pi) &= \min_{(W_i, i\in[m]):W_i\subseteq([n]\setminus A_i), i\in[m]} \mathcal{L}^*\big((n, m, (A_i, i \in [m]), (W_i, i \in [m]))\big) \\ &= \min_{(W_i, i\in[m]):W_i\subseteq([n]\setminus A_i), i\in[m]} \Big(\beta\Big(\tilde{\Gamma}(S \cup U)\Big) - u\Big) \\ &= \min_{(W_i, i\in[m]):W_i\subseteq([n]\setminus A_i), i\in[m]} \beta\Big(\tilde{\Gamma}(S \cup U)\Big) - u,\end{aligned}$$

which completes the proof. ∎

Proposition 3: For any pliable index coding problem Π, there exists a deterministic scalar linear index code that yields the following leakage rate,

$$\mathcal{L} = \min_{(W_i, i\in[m]):W_i\subseteq([n]\setminus A_i), i\in[m]} \min_{\mathbf{M}} (r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U})), \tag{22}$$

where, given a decoding message tuple $(W_i, i \in [m])$, $\mathbf{M} = [\mathbf{K} \quad \mathbf{S} \quad \mathbf{U}]$ denotes any fitting matrix for the index coding problem $(n, m, (W_i, i \in [m]), (A_i, i \in [m]))$. Furthermore, this result is leakage-wise rate optimal for all deterministic scalar linear codes.

Proof: Fix any decoding message tuple $(W_i, i \in [m])$, Theorem 2 implies that there exists some deterministic scalar coding scheme based on fitting matrices that gives

$$\mathcal{L} = \min_{\mathbf{M}} (r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U})).$$

And this $\mathcal{L}$ is optimal for all deterministic scalar linear codes for the index coding problem associated with $(W_i, i \in [m])$. Then, minimizing over all possible decoding message tuples, we know the leakage rate in (22) is achievable, and is leakage-wise rate optimal for all deterministic scalar linear codes for the pliable index coding problem Π. ∎

VI. Conclusion

We studied information leakage of sensitive messages in index coding to a guessing eavesdropper. A characterization of the optimal leakage rate using optimal compression rate was developed from a graph-theoretic perspective. We also proposed a deterministic linear coding scheme utilizing the rank minimization technique based on fitting matrices. In the following, we conclude the paper with several open questions and concluding remarks that may motivate future studies.

- While some intuitions behind the result of Theorem 1 has been given in Remark 7, a deeper investigation of the relationship between Γ and $\tilde{\Gamma}$ may lead to further results.
- Extending the *scalar* linear coding scheme in Section IV to a general *vector* linear code should give better performance in both leakage and compression. However, this may incur higher computation and implementation complexity.
- A general characterization of the optimal leakage rate under an arbitrary compression rate constraint turns out to be quite challenging and remains open at the current stage.
- It may prove beneficial if the techniques utilized in this paper can be extended to be used in the study of information leakage in other related problems, such as locally repairable distributed storage, coded caching, and multi-terminal source coding.

[4]Note that the extra decoder is not "pliable". No matter what decoding message tuple $(W_i, i \in [m])$ is for the original decoders, the extra decoder always requests to decode X_U.

References

[1] Y. Liu, L. Ong, P. L. Yeoh, P. Sadeghi, J. Kliewer, and S. Johnson, "Information leakage in index coding with sensitive and non-sensitive messages," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Espoo, Finland, 2022, pp. 3256–3261.

[2] Y. Birk and T. Kol, "Informed-source coding-on-demand (ISCOD) over broadcast channels," in *Proc. IEEE INFOCOM*, Mar. 1998, pp. 1257–1264.

[3] Z. Bar-Yossef, Y. Birk, T. S. Jayram, and T. Kol, "Index coding with side information," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1479–1494, Mar. 2011.

[4] G. Smith, "On the foundations of quantitative information flow," in *Proc. Int. Conf. Found. Softw. Sci. Comput. Struct.*, 2009, pp. 288–302.

[5] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2020.

[6] C. Braun, K. Chatzikokolakis, and C. Palamidessi, "Quantitative notions of leakage for one-try attacks," *Electron. Notes Theor. Comput. Sci.*, vol. 249, pp. 75–91, Aug. 2009.

[7] J. Liao, L. Sankar, F. P. Calmon, and V. Y. F. Tan, "Hypothesis testing under maximal leakage privacy constraints," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2017, pp. 779–783.

[8] M. Karmoose, L. Song, M. Cardone, and C. Fragouli, "Privacy in index coding: k-limited-access schemes," *IEEE Trans. Inf. Theory*, vol. 66, no. 5, pp. 2625–2641, May 2020.

[9] A. R. Esposito, M. Gastpar, and I. Issa, "Learning and adaptive data analysis via maximal leakage," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2019, pp. 1–5.

[10] Y. Liu, N. Ding, P. Sadeghi, and T. Rakotoarivelo, "Privacy-utility trade-off in a guessing framework inspired by index coding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2020, pp. 926–931.

[11] R. Zhou, T. Guo, and C. Tian, "Weakly private information retrieval under the maximal leakage metric," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2020, pp. 1089–1094.

[12] B. Wu, A. B. Wagner, and G. E. Suh, "Optimal mechanisms under maximal leakage," in *Proc. IEEE Conf. Commun. Netw. Security (CNS)*, 2020, pp. 1–6.

[13] M. S. Alvim, K. Chatzikokolakis, C. Palamidessi, and G. Smith, "Measuring information leakage using generalized gain functions," in *Proc. IEEE 25th Comput. Security Found. Symp.*, 2012, pp. 265–279.

[14] B. Espinoza and G. Smith, "Min-entropy as a resource," *Inf. Comput.*, vol. 226, pp. 57–75, May 2013.

[15] M. S. Alvim, K. Chatzikokolakis, A. Mciver, C. Morgan, C. Palamidessi, and G. Smith, "Additive and multiplicative notions of leakage, and their capacities," in *Proc. IEEE 27th Comput. Security Found. Symp.*, 2014, pp. 308–322.

[16] G. Smith, "Recent developments in quantitative information flow (invited tutorial)," in *Proc. 30th Annu. ACM/IEEE Symp. Logic Comput. Sci.*, 2015, pp. 23–31.

[17] Y. Y. Shkel and H. V. Poor, "A compression perspective on secrecy measures," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2020, pp. 995–1000.

[18] I. Issa and A. B. Wagner, "Measuring secrecy by the probability of a successful guess," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3783–3803, Jun. 2017.

[19] Y. Liu, L. Ong, P. L. Yeoh, P. Sadeghi, J. Kliewer, and S. Johnson, "Information leakage in index coding," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2021, pp. 1–6.

[20] S. H. Dau, V. Skachek, and Y. M. Chee, "On the security of index coding with side information," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3975–3988, Jun. 2012.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.

[21] L. Ong, B. N. Vellambi, P. L. Yeoh, J. Kliewer, and J. Yuan, "Secure index coding: Existence and construction," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, 2016, pp. 2834–2838.
[22] L. Ong, J. Kliewer, and B. N. Vellambi, "Secure network-index code equivalence: Extension to non-zero error and leakage," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Vail, CO, USA, 2018, pp. 841–845.
[23] M. M. Mojahedian, M. R. Aref, and A. Gohari, "Perfectly secure index coding," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7382–7395, Nov. 2017.
[24] Y. Liu, Y.-H. Kim, B. N. Vellambi, and P. Sadeghi, "On the capacity region for secure index coding," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Guanzhou, China, Nov. 2018, pp. 1–5.
[25] N. Alon, E. Lubetzky, U. Stav, A. Weinstein, and A. Hassidim, "Broadcasting with side information," in *Proc. 49th Annu. IEEE Symp. Found. Comput. Sci. (FOCS)*, Oct. 2008, pp. 823–832.
[26] S. Brahma and C. Fragouli, "Pliable index coding," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 6192–6203, Nov. 2015.
[27] F. Arbabjolfaei and Y.-H. Kim, "Fundamentals of index coding," *Found. Trends Commun. Inf. Theory*, vol. 14, nos. 3–4, pp. 163–346, 2018.
[28] E. R. Scheinerman and D. H. Ullman, *Fractional Graph Theory: A Rational Approach to the Theory of Graphs*. New York, NY, USA: Courier Corp., 2011.
[29] M. H. R. Khouzani and P. Malacaria, "Generalized entropies and metric-invariant optimal countermeasures for information leakage under symmetric constraints," *IEEE Trans. Inf. Theory*, vol. 65, no. 2, pp. 888–901, Feb. 2019.
[30] M. Romanelli, K. Chatzikokolakis, C. Palamidessi, and P. Piantanida, "Estimating g-leakage via machine learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2020, pp. 697–716.
[31] G. R. Kurri, O. Kosut, and L. Sankar, "Evaluating multiple guesses by an adversary via a tunable loss function," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2021, pp. 2002–2007.
[32] F. Arbabjolfaei and Y.-H. Kim, "Structural properties of index coding capacity using fractional graph theory," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Hong Kong, Jun. 2015, pp. 1034–1038.
[33] R. Hammack, W. Imrich, and S. Klavžar, *Handbook of Product Graphs*. Boca Raton, FL, USA: CRC Press, 2011.
[34] G. Sabidussi, "Graphs with given group and given graph-theoretical properties," *Can. J. Math.*, vol. 9, no. 4, pp. 515–525, 1957.
[35] O. Baksalary and G. Trenkler, "On k-potent matrices," *Electron. J. Linear Algebra*, vol. 26, pp. 446–470, Jan. 2013.
[36] C. E. Shannon, "Communication theory of secrecy systems," *Bell Syst. Tech. J.*, vol. 28, no. 4, pp. 656–715, Oct. 1949.
[37] H. Yamamoto, "Coding theorems for Shannon's cipher system with correlated source outputs, and common information," *IEEE Trans. Inf. Theory*, vol. 40, no. 1, pp. 85–95, Jan. 1994.
[38] C. Schieler and P. Cuff, "Rate-distortion theory for secrecy systems," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7584–7605, Dec. 2014.
[39] T. Guo, R. Zhou, and C. Tian, "On the information leakage in private information retrieval systems," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2999–3012, 2020.
[40] R. W. Yeung, *Information Theory and Network Coding*. New York, NY, USA: Springer, 2008.
[41] T. Liu and D. Tuninetti, "Tight information theoretic converse results for some pliable index coding problems," *IEEE Trans. Inf. Theory*, vol. 66, no. 5, pp. 2642–2657, May 2020.
[42] L. Song and C. Fragouli, "A polynomial-time algorithm for pliable index coding," *IEEE Trans. Inf. Theory*, vol. 64, no. 2, pp. 979–999, Feb. 2018.
[43] L. Ong, B. N. Vellambi, and J. Kliewer, "Optimal-rate characterisation for pliable index coding using absent receivers," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2019, pp. 522–526.
[44] T. Liu and D. Tuninetti, "Private pliable index coding," in *Proc. IEEE Inf. Theory Workshop (ITW)*, 2019, pp. 1–5.
[45] T. Liu and D. Tuninetti, "Decentralized pliable index coding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2019, pp. 532–536.
[46] S. Sasi and B. S. Rajan, "Code construction for pliable index coding," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, 2019, pp. 527–531.

Authorized licensed use limited to: New Jersey Institute of Technology. Downloaded on December 18,2023 at 16:39:42 UTC from IEEE Xplore. Restrictions apply.