Information Leakage in Index Coding With Sensitive and Non-Sensitive Messages

Yucheng Liu[†], Lawrence Ong[†], Phee Lep Yeoh[§], Parastoo Sadeghi[‡], Joerg Kliewer[*], and Sarah Johnson[†],
[†]The University of Newcastle, Australia (emails: {yucheng.liu, lawrence.ong, sarah.johnson}@newcastle.edu.au)
[§]University of Sydney, Australia (email: phee.yeoh@sydney.edu.au)
[‡]University of New South Wales, Canberra, Australia (email: p.sadeghi@unsw.edu.au)
[*]New Jersey Institute of Technology, USA (email: jkliewer@njit.edu)

2022 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-2159-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ISIT50566.2022.9834747

***Abstract*—Information leakage to a guessing adversary in index coding is studied, where some messages in the system are sensitive and others are not. The non-sensitive messages can be used by the server like secret keys to mitigate leakage of the sensitive messages to the adversary. We construct a deterministic linear coding scheme, developed from the rank minimization method based on fitting matrices (Bar-Yossef et al. 2011). The linear scheme leads to a novel upper bound on the optimal information leakage rate, which is proved to be tight over all deterministic scalar linear codes. We also derive a converse result from a graph-theoretic perspective, which holds in general over all deterministic and stochastic coding schemes.**

I. Introduction

Index coding [1], [2] is a canonical problem in network information theory, where a server tries to efficiently broadcast messages to multiple receivers with side information. In this work, we study the information leakage in index coding when the broadcast codeword is eavesdropped by a guessing adversary. The adversary tries to maximize the probability of correctly guessing its messages of interest within a single trial. 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 [3]–[5].

Our previous work [6] studied the information leakage with an underlying assumption that the server aims to protect *all* of the messages against the adversary. Indeed, such assumption holds in most existing works [7]–[11] where the security and privacy aspects of index coding are investigated. Nevertheless, in many practical circumstances, some messages may be sensitive while the others are non-sensitive, and thus secrecy loss should be measured over only those sensitive messages. For example, consider a data storage system with a number of messages. Some messages are private and needs to be protected from any adversary, while the other messages are non-private and thus need not be protected. There are a number of clients, each of which may request access to some messages, whether private or non-private, and may have stored some other messages already as side information. The goal is to satisfy the clients' requirements while keeping the private messages as safe as possible from the adversary. A similar setting can also be motivated from the adversary's perspective as it may only be interested in a subset of messages and thus these messages should be considered as sensitive by the server. This distinction between sensitive and non-sensitive messages enables the server to treat the non-sensitive messages like secret keys and design a smart way of coding to simultaneously satisfy the receivers and mitigate information leakage to the adversary. Figure 1 serves as a toy example showing how the server can reduce the information leakage by smartly designing coding schemes.

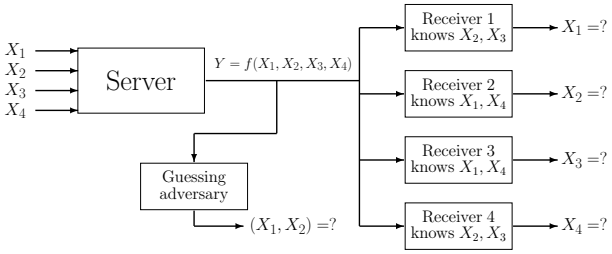

Figure 1. There are four binary messages $X_i, i \in [4]$, where X_1, X_2 are sensitive and X_3, X_4 are non-sensitive. A guessing adversary eavesdrops the broadcast codeword Y and tries to guess the sensitive messages. When guessing blindly (without knowing Y), the probability of the adversary correctly guessing (X_1, X_2) is only $1/4$. To satisfy the legitimate receivers, the server 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 receivers and prevent any information leakage, the server can broadcast $\tilde{Y} = (X_1 \oplus X_3, X_2 \oplus X_4)$. In this way, the sensitive messages are protected against the adversary using non-sensitive messages.

We study the problem of both achievability and converse. In Section II, we describe the system model and provide necessary preliminaries. In Section III, we propose a practical linear coding scheme based on the rank minimization method over fitting matrices [2]. The scheme is proved to yield optimal (minimum) information leakage rate over all deterministic scalar linear codes. In Section IV, we develop a general lower bound on the optimal leakage rate that holds over any coding schemes, even stochastic ones. To show the lower bound, we use the confusion graph representation of

This work was supported by the ARC Discovery Scheme DP190100770; the US National Science Foundation Grant CNS-1815322; and the ARC Future Fellowship FT190100429.

index coding problems [12] and apply graph-theoretic tools.

II. Problem Formulation

A. System Model and Preliminaries

We consider a server containing 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 server encodes the n messages to some codeword Y and broadcasts it to m receivers via a noiseless channel. Receiver $i \in [m]$ wants to know messages X_{W_i} and has X_{A_i} as side information. We allow degenerated receivers in the system, who wants nothing (i.e., $W_i = \emptyset$). Such a receiver can always decode what it wants perfectly by definition.

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

- One *stochastic* encoder $f : \mathcal{X}^{nt} \to \{1, 2, \ldots, M\}$ at the server that maps each message tuple $x_{[n]} \in \mathcal{X}^{nt}$ to a codeword $y \in \{1, 2, \ldots, M\}$, and
- m *deterministic* decoders $\mathbf{g} = (g_i, i \in [m])$, one for each receiver $i \in [m]$, 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 receiver can perfectly decode its wanted messages. A compression rate R is achievable iff there exists a valid $(t, M, f, \mathbf{g})$ code such that $R \geq (\log_q M)/t$. The optimal rate β, also called the *broadcast rate*, can be defined as [13]

$$\beta = \lim_{t\to\infty} \min_{\text{valid } (t,\, M,\, f,\, \mathbf{g}) \text{ code}} \frac{\log_q M}{t}. \tag{1}$$

Remark 1: Broadcast rate can also be defined allowing vanishing decoding error. However, for index coding, the zero-error and vanishing-error broadcast rates are equal [14].

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, $(\Gamma_t, t \in \mathbb{Z}^+)$ [12]. For a given sequence length t, the confusion graph Γ_t is an undirected graph defined on the message 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 there exists some receiver $i \in [m]$ such that $x_{W_i} \neq z_{W_i}$ and $x_{A_i} = z_{A_i}$. We call any pair of vertices satisfying this condition *confusable* at receiver i, or simply confusable. Hence, the edge set $E(\Gamma_t) = \{\{x^t_{[n]}, z^t_{[n]}\} : x_{W_i} \neq z_{W_i}$ and $x_{A_i} = z_{A_i}$ for some $i \in [m]\}$.

For correct decoding at all receivers, any two values $x_{[n]}, z_{[n]}$ can be mapped to the same codeword y with nonzero probabilities iff they are not confusable [12]. See Figure 2 below for a toy example of an index coding instance and its confusion graph. For the definitions for basic graph-theoretic notions, see any textbook on graph theory (e.g., Scheinerman and Ullman [15]).

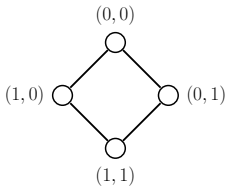

Figure 2. The confusion graph Γ_1 with $t = 1$ for the index coding instance $(2, 2, (\{1\}, \{2\}), (\{2\}, \{1\}))$. Note that, for example, $x_{[n]} = (0, 0)$ and $z_{[n]} = (0, 1)$ are confusable at receiver 2, because $x_{W_2} = x_2 = 0 \neq z_{W_2} = z_2 = 1$ and $x_{A_2} = x_1 = 0 = z_1 = z_{A_2}$. Suppose $(0, 0)$ and $(0, 1)$ are mapped to the same codeword y with certain nonzero probabilities. Then upon receiving this y, receiver 2 will not be able to tell whether the value for X_2 is 0 or 1 based on its side information of $X_1 = 0$.

To simplify the notation, we may use Γ to denote an index coding instance characterized by $(\Gamma_t, t \in \mathbb{Z}^+)$. We denote the broadcast rate of the problem Γ as $\beta(\Gamma)$ when this dependence needs to be emphasized.

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 itself 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

$$\beta(\Gamma) = \lim_{t\to\infty} \frac{1}{t} \log_q \chi(\Gamma_t) = \lim_{t\to\infty} \frac{1}{t} \log_q \chi_{\mathrm{f}}(\Gamma_t), \tag{2}$$

where $\chi(\cdot)$ and $\chi_{\mathrm{f}}(\cdot)$ respectively denote the chromatic number and fractional chromatic number of a graph. The proof of (2) can be found in [13, Section 3.2].

C. Information Leakage Metric

We assume the broadcast codeword is eavesdropped by a guessing adversary, knowing 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. Upon observing the broadcast codeword, the adversary makes a single guess on the value of X_S according to the maximum likelihood rule. 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.

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 is

$$\mathrm{P_s}(X_K) = \mathbb{E}_{X_K}\left[\max_{x_S} P_{X_S|X_K}(x_S|X_K)\right] = |\mathcal{X}|^{-ts},$$

and the expected successful guessing probability after observing Y is

$$\mathrm{P_s}(X_K, Y) = \mathbb{E}_{Y,X_K}\left[\max_{x_S} P_{X_S|Y,X_K}(x_S|Y, X_K)\right].$$

The leakage L is defined as the logarithm of the ratio between the expected probabilities of the adversary successfully guessing x_Q after and before observing Y [3], [4]. That is,

$$L \doteq \log_q \frac{\mathrm{P_s}(X_K, Y)}{\mathrm{P_s}(X_K)} \tag{3}$$

$$= \log_q \sum_{x_K, y} \max_{x_S} P_{Y, X_K | X_S}(y, x_K | x_S). \tag{4}$$

The leakage rate of the code is $\mathcal{L} = t^{-1}L$ and the optimal leakage rate can then be defined as

$$\mathcal{L}^* \doteq \lim_{t \to \infty} \inf_{(t, M, f, \mathbf{g}) \text{ codes}} \mathcal{L}. \tag{5}$$

Remark 2: The idea of measuring leakage as the ratio of the adversary's successful guessing probabilities has been introduced and explored in various contexts [3]–[5]. The leakage L is equal to the maximal leakage [5] from X_S to Y given side information X_K, which is also equal to the maximum min-entropy leakage [4] from X_S to Y given X_K.

III. 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 [2]. 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.

Unless otherwise stated, we use bold-faced capital letters to denote matrices and vectors, e.g., $\mathbf{X}_{[n]} = [X_1 \ \ \dots \ \ X_n]^T$. Let $r(\cdot)$ denote the rank of a matrix over the Galois field $\mathbb{F}_q$.

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

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

$$\mathbf{M}_{ij} = 1, \quad \text{for } j = w_i, \tag{6}$$

$$\mathbf{M}_{ij} = 0, \quad \text{for any message } j \in [n] \setminus (W_i \cup A_i). \tag{7}$$

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 $m = 3$ receivers, 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 form below is a fitting matrix,

$$\begin{bmatrix} 1 & 0 & ? & ? & 0 \\ ? & 1 & 0 & ? & ? \\ 0 & ? & 1 & 0 & ? \end{bmatrix} \tag{8}$$

where "?" means that the entry can be any element in $\mathbb{F}_q$.

If the server generates codeword $\mathbf{Y}$ by multiplying a fitting matrix $\mathbf{M}$ by the message vector $\mathbf{X}_{[n]} = [X_1 \ X_2 \ \dots \ X_n]^T$, every receiver $i \in [m]$ 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 decoding requirements at the receivers, the server needs only to transmit the linear combination 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, namely, the *minrank* value, establishes an upper bound on its broadcast rate, which has been proved to be optimal over all deterministic scalar linear codes [2].

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

$$\mathbf{M} = [\mathbf{K} \quad \mathbf{S} \quad \mathbf{U}]. \tag{9}$$

The following theorem characterizes the minimal leakage rate among all codes based on the fitting matrix framework.

Theorem 1: For any index coding problem, there exists a deterministic scalar linear index code that yields the following leakage rate,

$$\mathcal{L} = \min_{\mathbf{M}} (r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U})). \tag{10}$$

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 receiving $\mathbf{Y} = \mathbf{EX}$ every receiver can decode its wanted message. As $t = 1$, we have

$$\begin{aligned}
\mathcal{L} &= \log_q \sum_{x_K, y} \max_{x_S} P_{Y, X_K | X_S}(y, x_K | x_S) \\
&= \log_q \sum_{x_K, y} \max_{x_S} \sum_{x_U} P_{Y, X_K, X_U | X_S}(y, x_K, x_U | x_S) \\
&\overset{(a)}{=} \log_q \sum_{x_K, y} \max_{x_S} \sum_{x_U} P_{Y | X_{[n]}}(y | x_{[n]}) \cdot P_{X_{K \cup U}}(x_{K \cup U}) \\
&\overset{(b)}{=} \log_q \left(q^{-k-u} \sum_{x_K} \sum_{y} \max_{x_S} \sum_{x_U} \mathbb{1}(\mathbf{Y} = \mathbf{EX}) \right) \\
&\overset{(c)}{=} \log_q \left(q^{-k-u} \sum_{x_K} \sum_{y} q^{u - r(\mathbf{U})} \right) \\
&\overset{(d)}{=} r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U}),
\end{aligned}$$

where (a) is due to the messages being independent, (b) is due to the code being deterministic, where $\mathbb{1}(\cdot)$ is the indicator function, (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 (10) 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

$$\tilde{\mathbf{E}} = [\tilde{\mathbf{K}} \quad \tilde{\mathbf{S}} \quad \tilde{\mathbf{U}}]. \tag{11}$$

Following a similar argument as in the achievability proof of the theorem, we can show that the leakage rate caused by the codeword $\mathbf{Y} = \tilde{\mathbf{E}}\mathbf{X}$ is

$$\mathcal{L}_{\tilde{\mathbf{E}}} = r([\tilde{\mathbf{S}} \quad \tilde{\mathbf{U}}]) - r(\tilde{\mathbf{U}}). \tag{12}$$

It remains to show that $L_{\tilde{\mathbf{E}}}$ is lower bounded by (10).

According to the proof of [2, Theorem 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, $\mathbf{B}[\tilde{\mathbf{S}} \quad \tilde{\mathbf{U}}] = [\mathbf{S} \quad \mathbf{U}]$. Also, there exists some matrix $\mathbf{D}$ such that $\tilde{\mathbf{U}} = [\tilde{\mathbf{S}} \quad \tilde{\mathbf{U}}]\mathbf{D}$, $\mathbf{U} = [\mathbf{S} \quad \mathbf{U}]\mathbf{D} = \mathbf{B}[\tilde{\mathbf{S}} \quad \tilde{\mathbf{U}}]\mathbf{D}$. Hence,

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

where (a) is due to Frobenius inequality [16]. By reorganizing the above result, we have $\mathcal{L}_{\tilde{\mathbf{E}}} = r([\tilde{\mathbf{S}} \quad \tilde{\mathbf{U}}]) - r(\tilde{\mathbf{U}}) \geq r([\mathbf{S} \quad \mathbf{U}]) - r(\mathbf{U})$, which, together with the fact that $\mathbf{M}$ is a fitting matrix, completes the proof. ∎

Remark 3: Following similar arguments as in the proof of Theorem 1, we can show that when the mutual information[1] $I(X_S; Y|X_K)$ 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. It seems that the rank $r(\mathbf{U})$ can be viewed as a rough measure of the level of protection against the adversary provided by the non-sensitive messages X_U.

Remark 4: Computing (10) for index coding instances with a large number of messages and receivers can be computationally challenging. When there is no non-sensitive messages, computing (10) reduces to computing the minrank value for the index coding instance over the finite field $\mathbb{F}_q$, which has been shown to be NP-complete [22].

The example below shows the efficacy of the scheme.

Example 1: Consider the problem Γ with $n = 5$ binary messages and $m = 5$ receivers 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, $K = \{5\}, S = \{1, 3\}, U = \{2, 4\}$. The fitting matrix

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{13}$$

achieves the broadcast rate of $\beta = r(\mathbf{M}) = 4$. It also gives

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

which is indeed the optimal leakage rate as we will show in Example 3 in the next section. The linear code given by (13) is optimal in both leakage and compression senses.

However, as we show by the simple two-receiver example below, it is not always possible to simultaneously achieve the optimal compression and leakage rates.

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

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

Assume for the adversary, $K = \emptyset, S = \{1, 2\}, U = \{3, 4\}$. 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, in the following we use Shannon-type inequalities [23, Chapter 14] to show that the compression rate of any index code that attains zero leakage is at least 2. Recall that we are considering binary uniform messages that are independent to each other. Thus, $t = 1$ and $H(X_J) = |J|, \forall J \subseteq [4]$. Consider any $(t, M, f, \mathbf{g})$ index code that achieves zero leakage, i.e., $L = I(X_{\{1,2\}}; Y) = 0$. We have $2 = H(X_{\{1,2\}}) = H(X_{\{1,2\}}|Y) = H(X_1|Y) + H(X_2|Y, X_1) = H(X_2|Y) + H(X_1|Y, X_2)$, which implies that $H(X_2|Y, X_1) = H(X_1|Y, X_2) = 1$. Then, we have

$$\begin{aligned} &H(X_3|Y, X_{\{1,2\}}) \\ &= H(X_3|Y, X_2) - H(X_1|Y, X_2) + H(X_1|Y, X_{\{2,3\}}) \\ &\overset{(a)}{=} H(X_3|Y, X_2) - 1 \leq H(X_3) - 1 = 0, \end{aligned}$$

where (a) follows since that $H(X_1|Y, X_2) = 1$ and $H(X_1|Y, X_{\{2,3\}}) = 0$ due to the decoding requirement that receiver 1 must be able to recover X_1 from Y and $X_{\{2,3\}}$. As entropy is always non-negative, we have $H(X_3|Y, X_{\{1,2\}}) = 0$. Similarly, we can show that $H(X_4|Y, X_{\{1,2\}}) = 0$ and hence $H(X_{\{3,4\}}|Y, X_{\{1,2\}}) = 0$. Thus, we have

$$\begin{aligned} t^{-1} \log_q M &\geq H(Y) \\ &\geq I(Y; X_{\{3,4\}}|X_{\{1,2\}}) \\ &= H(X_{\{3,4\}}|X_{\{1,2\}}) - H(X_{\{3,4\}}|Y, X_{\{1,2\}}) \\ &= H(X_{\{3,4\}}) - H(X_{\{3,4\}}|Y, X_{\{1,2\}}) = 2. \end{aligned}$$

That is, the compression rate of any index coding attaining zero leakage must be at least 2. Therefore, $\beta = 1$ and $\mathcal{L}^* = 0$ can never be simultaneously achieved for this problem.

IV. A General Lower Bound on $\mathcal{L}^*$

We derive the following lower bound on the leakage rate that holds generally over any valid index codes.

[1]Note that mutual information between sensitive variables and codeword has been commonly used as a leakage metric in the literature [17]–[21].

Theorem 2: For any index coding problem Γ, we have

$$\mathcal{L}^*(\Gamma) \geq \beta(\tilde{\Gamma}(S \cup U)) - u, \tag{14}$$

where $\tilde{\Gamma}$ denotes the index coding problem constructed by adding an extra receiver, indexed by $m+1$, to the original problem, which knows side information $A_{m+1} = K \cup S$ and wants messages $W_{m+1} = U$.

Proof: Consider any valid $(t, M, f, \mathbf{g})$ index code. For any (x_K, y) value and any $J \subseteq S \cup U$, let

$$\mathcal{X}_J(x_K, y) = \{x_J \in \mathcal{X}_J : P_{Y, X_{K \cup J}}(y, x_{K \cup J}) > 0\}$$

denote the set of X_J values jointly possible with (x_K, y).

Now consider the problem $\tilde{\Gamma}$, which is constructed from Γ by adding an extra receiver $m+1$ with $A_{m+1} = K \cup S$ and $W_{m+1} = U$. Notice that by adding a receiver we are essentially adding more edges into the confusion graph (i.e., $E(\Gamma_t) \subseteq E(\tilde{\Gamma}_t)$). We first show the following inequality:

$$\max_{x_K, y} |\mathcal{X}_S(x_K, y)| \leq \alpha(\tilde{\Gamma}_t(S \cup U)), \tag{15}$$

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

We prove the inequality via contradiction. Assume there exists some (x_K, y) value such that $|\mathcal{X}_S(x_K, y)| > \alpha(\tilde{\Gamma}_t(S \cup U))$. Consider a subset $\mathcal{I}$ of the vertex set $\mathcal{X}_{S \cup U}(x_K, y)$ in the subgraph $\Gamma_t(S \cup U)$ such that for every $x_S \in \mathcal{X}_S(x_K, y)$, there is exactly one vertex in $v_{S \cup U} \in \mathcal{I}$ such that $v_S = x_S$. Thus $|\mathcal{I}| = |\mathcal{X}_S(x_K, y)|$. For a visualization of the construction of $\mathcal{I}$, see the schematic graph in Figure 3.

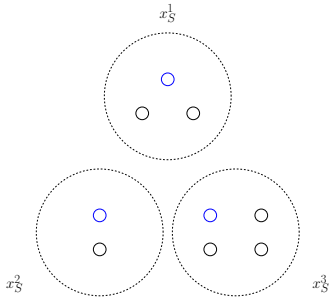

Figure 3. Vertices in $\mathcal{X}_{S \cup U}(x_K, y)$ are denoted by nodes in the figure. Note that the figure is for illustrative purpose only; there is no limit on the number of vertices in $\mathcal{X}_{S \cup U}(x_K, y)$. We partition the nodes into subgroups according to their x_S values, and each 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. For example, $\mathcal{I}$ can be constructed by the blue nodes, one from each subgroup. The number of distinct x_S values in $\mathcal{X}_{S \cup U}(x_K, y)$, $|\mathcal{X}_S(x_K, y)|$, is equal to the number of dashed circles in the graph, which is then equal to $|\mathcal{I}|$.

By the definition of the confusion graph, $\mathcal{X}_{S \cup U}(x_K, y)$ is an independent set in the induced subgraph $\Gamma_t(S \cup U)$. Since $\mathcal{I} \subseteq \mathcal{X}_{S \cup U}(x_K, y)$, $\mathcal{I}$ is also an independent set in $\Gamma_t(S \cup U)$. Thus, any two vertices in $\mathcal{I}$ are not confusable at any receiver $i \in S \cup U$ of the subproblem $\Gamma(S \cup U)$. Note that the extra receiver $m+1$ 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 receiver $m+1$ either. In conclusion, any two vertices in $\mathcal{I}$ are not confusable at any receivers 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 $|\mathcal{I}| \leq \alpha(\tilde{\Gamma}_t(S \cup U))$, which contradicts with the assumption that $|\mathcal{X}_S(x_K, y)| = |\mathcal{I}| > \alpha(\tilde{\Gamma}_t(S \cup U))$.

Having proved (15), we have

$$\begin{aligned}
&\sum_{x_K, y} \max_{x_S} P_{Y, X_{K,S}}(y, x_{K,S}) \\
&\overset{(a)}{\geq} \sum_{x_K, y} \frac{1}{|\mathcal{X}_S(x_K, y)|} \sum_{x_S \in \mathcal{X}_S(x_K, y)} P_{Y, X_{K,S}}(y, x_{K,S}) \\
&\overset{(b)}{\geq} \frac{1}{\alpha(\tilde{\Gamma}_t(S \cup U))} \sum_{x_K, y} \sum_{x_S \in \mathcal{X}_S(x_K, y)} P_{Y, X_{K,S}}(y, x_{K,S}) \\
&= \frac{1}{\alpha(\tilde{\Gamma}_t(S \cup U))},
\end{aligned} \tag{16}$$

where (a) follows since the maximum is larger than the average, and (b) is due to (15). Finally, we have

$$\begin{aligned}
&\mathcal{L}^*(\Gamma) \\
&= \lim_{t \to \infty} \frac{1}{t} \min_{(t, M, f, \mathbf{g}) \text{ codes}} \log_q \sum_{x_K, y} \max_{x_S} P_{Y, X_K | X_S}(y, x_K | x_S) \\
&\overset{(c)}{\geq} \lim_{t \to \infty} \frac{1}{t} \log_q \frac{|\mathcal{X}|^{ts}}{\alpha(\tilde{\Gamma}_t(S \cup U))} \\
&\overset{(d)}{=} \lim_{t \to \infty} \frac{1}{t} \log_q \frac{|V(\tilde{\Gamma}_t(S \cup U))| \cdot |\mathcal{X}|^{-tu}}{\alpha(\tilde{\Gamma}_t(S \cup U))} \\
&\overset{(e)}{=} \lim_{t \to \infty} \frac{1}{t} \log_q \chi_{\mathrm{f}}(\tilde{\Gamma}_t(S \cup U)) - u \\
&\overset{(f)}{=} \beta(\tilde{\Gamma}(S \cup U)) - u,
\end{aligned}$$

where (c) follows from (16) and the fact that all x_S are equally likely, (d) follows from $|V(\tilde{\Gamma}_t(S \cup U))| = |\mathcal{X}|^{t(s+u)}$, (e) follows since confusion graphs are vertex-transitive [15], and (f) follows from (2). ∎

Remark 5: While in the current work $\tilde{\Gamma}$ only appears as a part of the computable expression of the converse result, the deeper relationship between Γ and $\tilde{\Gamma}$ in both compression and secrecy senses is worth investigating.

Remark 6: It can be shown using the generalized maximal acyclic induced subgraph (MAIS) bound [2] that the lower bound in Theorem 2 is always non-negative. In particular, when there are no non-sensitive messages in the system (i.e., $U = \emptyset$), the extra receiver in $\tilde{\Gamma}(S \cup U)$ becomes trivial as it knows all the messages in the system and wants nothing. In such case, the lower bound in Theorem 2 reduces to that in Corollary 1 in our previous work [6], and is always tight.

Theorem 2 can sometimes yield tight lower bounds on $\mathcal{L}^*$.

Example 3: Consider the problem Γ in Example 1. Theorem 2 gives $\mathcal{L}^*(\Gamma) \geq \beta(\tilde{\Gamma}(\{1,2,3,4\})) - |\{2,4\}| = 3 - 2 = 1$, where $\beta(\tilde{\Gamma}(\{1,2,3,4\})) = 3$ can be proved using the MAIS bound [2] and the cycle covering scheme [24], [25]. The lower bound on $\mathcal{L}^*(\Gamma)$ above matches the achievability result in Example 1 and thus establishes $\mathcal{L}^*(\Gamma)$.

Remark 7: It can be shown utilizing existing results on broadcast rates [26, Theorem 1.1] that no constant gap exists between the upper and lower bounds in Theorems 1 and 2.

REFERENCES

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

[2] Z. Bar-Yossef, Y. Birk, T. Jayram, and T. Kol, "Index coding with side information," *IEEE Trans. Inf. Theory*, vol. 57, pp. 1479–1494, 2011.

[3] G. Smith, "On the foundations of quantitative information flow," in *International Conference on Foundations of Software Science and Computational Structures*. Springer, 2009, pp. 288–302.

[4] C. Braun, K. Chatzikokolakis, and C. Palamidessi, "Quantitative notions of leakage for one-try attacks," *Electronic Notes in Theoretical Computer Science*, vol. 249, pp. 75–91, 2009.

[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, 2019.

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

[7] 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, 2012.

[8] L. Ong, B. N. Vellambi, P. L. Yeoh, J. Kliewer, and J. Yuan, "Secure index coding: Existence and construction," in *Proc. IEEE Int. Symp. on Information Theory (ISIT)*, Barcelona, Spain, 2016, pp. 2834–2838.

[9] 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. on Information Theory (ISIT)*, Vail, CO, 2018, pp. 841–845.

[10] M. M. Mojahedian, M. R. Aref, and A. Gohari, "Perfectly secure index coding," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7382–7395, 2017.

[11] Y. Liu, Y.-H. Kim, B. Vellambi, and P. Sadeghi, "On the capacity region for secure index coding," in *Proc. IEEE Information Theory Workshop (ITW)*, Guanzhou, China, Nov. 2018.

[12] N. Alon, E. Lubetzky, U. Stav, A. Weinstein, and A. Hassidim, "Broadcasting with side information," in *49th Annu. IEEE Symp. on Foundations of Computer Science (FOCS)*, Oct. 2008, pp. 823–832.

[13] F. Arbabjolfaei and Y.-H. Kim, "Fundamentals of index coding," *Foundations and Trends® in Communications and Information Theory*, vol. 14, no. 3-4, pp. 163–346, 2018.

[14] M. Langberg and M. Effros, "Network coding: Is zero error always possible?" in *Proc. 49th Ann. Allerton Conf. Comm. Control Comput.*, 2011, pp. 1478–1485.

[15] E. R. Scheinerman and D. H. Ullman, *Fractional Graph Theory: A Rational Approach to the Theory of Graphs*. Courier Corporation, 2011.

[16] O. Baksalary and G. Trenkler, "On k-potent matrices," *The Electronic Journal of Linear Algebra*, vol. 26, pp. 446–470, 2013.

[17] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949.

[18] 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, 1994.

[19] C. Schieler and P. Cuff, "Rate-distortion theory for secrecy systems," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7584–7605, 2014.

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

[21] 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.

[22] S. Y. El Rouayheb, M. A. R. Chaudhry, and A. Sprintson, "On the minimum number of transmissions in single-hop wireless coding networks," in *Proc. IEEE Information Theory Workshop (ITW)*, 2007, pp. 120–125.

[23] R. W. Yeung, *Information Theory and Network Coding*. Springer Science & Business Media, 2008.

[24] M. A. R. Chaudhry, Z. Asad, A. Sprintson, and M. Langberg, "On the complementary index coding problem," in *Proc. IEEE Int. Symp. on Information Theory (ISIT)*, 2011, pp. 244–248.

[25] M. J. Neely, A. S. Tehrani, and Z. Zhang, "Dynamic index coding for wireless broadcast networks," *IEEE Trans. Inf. Theory*, vol. 59, no. 11, pp. 7525–7540, 2013.

[26] E. Lubetzky and U. Stav, "Nonlinear index coding outperforming the linear optimum," *IEEE Trans. Inf. Theory*, vol. 55, no. 8, pp. 3544–3551, 2009.