Information Leakage in Index Coding

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)

2021 IEEE Information Theory Workshop (ITW) | 978-1-6654-0312-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ITW48936.2021.9611514

***Abstract*—We study the information leakage to a guessing adversary in index coding with a general message distribution. Under both vanishing-error and zero-error decoding assumptions, we develop lower and upper bounds on the optimal leakage rate, which are based on the broadcast rate of the subproblem induced by the set of messages the adversary tries to guess. When the messages are independent and uniformly distributed, the lower and upper bounds match, establishing an equivalence between the two rates.**

I. Introduction

Index coding [1], [2] studies the communication problem where a server broadcasts messages via a noiseless channel to multiple receivers with side information. Due to its simple yet fundamental model, index coding has been recognized as a canonical problem in network information theory, and is closely connected with many other problems such as network coding, distributed storage, and coded caching. Despite substantial progress achieved so far (see [3] and the references therein), the index coding problem remains open in general.

In secure index coding [4]–[7], the server must simultaneously satisfy the legitimate receivers' decoding requirements and protect the content of some messages from being obtained by an eavesdropping adversary. A variant of this setup puts security constraints on the receivers themselves against some messages [4], [8], [9]. Instead of protecting the messages, another variant of index coding has also been studied from a privacy-preserving perspective, where the goal is to limit the information that a receiver can infer about the *identities* of the requests of other receivers [10]. The privacy-utility tradeoff in a multi-terminal data publishing problem inspired by index coding was investigated in [11].

In this work, we study the information leakage to a guessing adversary in index coding, which, to the best of our knowledge, has not been considered in the literature. The adversary eavesdrops the broadcast codeword and tries to guess the message tuple via maximum likelihood estimation within a certain number of trials. Our aim is to characterize the information leakage to the adversary, which is defined as the ratio between the adversary's probability of successful guessing *after and before* observing the codeword [12]–[14]. For a visualization of the problem setup, see Figure 1.

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

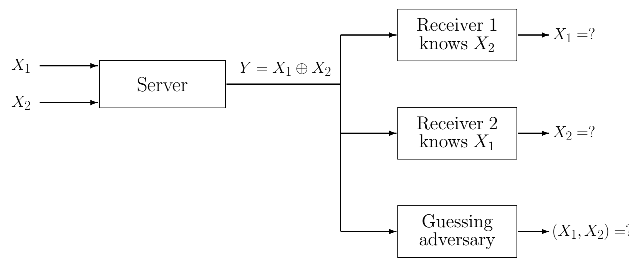

Figure 1. There are two correlated binary messages $X_{\{1,2\}} = (X_1, X_2)$ with distribution: $P_{X_{\{1,2\}}}(0,0) = 0.1$, $P_{X_{\{1,2\}}}(0,1) = 0.2$, $P_{X_{\{1,2\}}}(1,0) = 0.3$, and $P_{X_{\{1,2\}}}(1,1) = 0.4$. A binary codeword Y is generated by the server as $Y = X_1 \oplus X_2$ and broadcast to the receivers. Every receiver can decode its wanted message based on the codeword and its side information. An adversary eavesdrops the codeword Y and makes a single guess on $X_{\{1,2\}}$. If $Y = 0$, the adversary's guess will be $(1,1)$ as $P_{X_{\{1,2\}}|Y}(1,1|0) = 0.8 > P_{X_{\{1,2\}}|Y}(0,0|0) = 0.2 > P_{X_{\{1,2\}}|Y}(0,1|0) = P_{X_{\{1,2\}}|Y}(1,0|0) = 0$. Similarly, if $Y = 1$, the adversary's guess will be $(1,0)$.

Recently we have studied [15] information leakage to a guessing adversary in zero-error source coding defined by a family of confusion graphs [16]. While the index coding problem can also be characterized by a confusion graph family [17], the study of information leakage in index coding is intrinsically different from that of source coding in the following aspects. The most significant difference comes from the different internal structures within the two confusion graph families. More specifically, for the source coding model we considered [15], the relationship among the confusion graphs of different sequence lengths is characterized by the disjunctive product [18]. On the other hand, for the index coding problem, the relationship among the confusion graphs cannot be characterized by any previously defined graph product. Another difference is that while our previous work [15] requires zero-error decoding at the legitimate receiver assuming *worst-case* source distribution, this paper considers both zero-error and vanishing-error scenarios and assume a *general* message distribution. Furthermore, in this work we take into account the adversary's side information which can include any message in the system.

Our main contribution is developing lower and upper bounds (i.e., converse and achievability results) on the optimal information leakage rate, for both vanishing-error and zero-error scenarios. The converse bound is derived using graph-theoretic techniques based on the notion of confusion graphs for index coding [17]. The achievability result is established by constructing a deterministic coding scheme as a composite of the coding schemes for two subproblems, one induced by the messages the adversary knows as side information and the other induced by the messages the adversary does not know and thus tries to guess.

Moreover, we show that when the messages are uniformly distributed and independent of each other (as in most existing works for index coding), the lower and upper bounds developed match. This establishes an equivalence between the optimal leakage rate of the problem and the optimal compression rate of the subproblem induced by the messages the adversary tries to guess.

II. System Model and Problem Formulation

Notation: For any $a \in \mathbb{Z}^+$, $[a] \doteq \{1, 2, \cdots, a\}$. For any discrete random variable Z with probability distribution P_Z, we denote its alphabet by $\mathcal{Z}$ with realizations $z \in \mathcal{Z}$.

There are n discrete memoryless stationary messages (sources), $X_i, i \in [n]$, of some common finite alphabet $\mathcal{X}$. For any $S \subseteq [n]$, set $X_S \doteq (X_i, i \in S)$, $x_S \doteq (x_i, i \in S)$, and $\mathcal{X}_S \doteq \mathcal{X}^{|S|}$. 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, $X_\emptyset = x_\emptyset = \mathcal{X}_\emptyset = \emptyset$. We consider an arbitrary, but fixed distribution $P_{X_{[n]}}$ on $\mathcal{X}_{[n]}$, assuming without loss of generality that it has full support[1].

There is a server containing all messages. It encodes the tuple of n message sequences $X^t_{[n]} = (X^t_i, i \in [n])$ according to some (possibly randomized) encoding function f to some codeword Y that takes values in the code alphabet $\mathcal{Y} = \{1, 2, \dots, M\}$. Each message sequence $X^t_i = (X_{i,1}, X_{i,2}, \dots, X_{i,t})$ is of length t symbols. The server then transmits the codeword to n receivers via a noiseless broadcast channel of normalized unit capacity. Let $P_{Y, X^t_{[n]}}$ denote the joint distribution of the message sequence tuple $X^t_{[n]}$ and the codeword Y. For any $S \subseteq [n]$, we define the following notation for message sequence tuples.

- $\mathcal{X}^t_S = \mathcal{X}^{t|S|}$;
- $X^t_S = (X^t_i, i \in S) = (X_{S,1}, X_{S,2}, \dots, X_{S,t})$, where $X_{S,j} = (X_{i,j}, i \in S)$ for every $j \in [t]$. Note that $X_{i,j}$ denotes the j-th symbol of message sequence X^t_i.
- Similarly, $x^t_S = (x^t_i, i \in S) = (x_{S,j}, j \in [t])$, where $x_{S,j} = (x_{i,j}, i \in S)$ for every $j \in [t]$.

Also, as the messages are memoryless, for any $x^t_{[n]} = (x_{[n],1}, x_{[n],2}, \dots, x_{[n],t})$, $P_{X^t_{[n]}}(x^t_{[n]}) = \prod_{j \in [t]} P_{X_{[n]}}(x_{[n],j})$.

On the receiver side, we assume that receiver $i \in [n]$ wishes to obtain message X^t_i and knows $X^t_{A_i}$ as side information for some $A_i \subseteq [n] \setminus \{i\}$.

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

- One stochastic encoder $f : \mathcal{X}^{nt} \to \{1, 2, \dots, M\}$ at the server that maps each message sequence tuple $x^t_{[n]} \in \mathcal{X}^{nt}$ to a codeword $y \in \{1, 2, \dots, M\}$, and
- n deterministic decoders $\mathbf{g} = (g_i, i \in [n])$, one for each receiver $i \in [n]$, such that $g_i : \{1, 2, \dots, M\} \times \mathcal{X}^{t|A_i|} \to \mathcal{X}^t$ maps the codeword y and the side information $x^t_{A_i}$ to some estimated sequence $\hat{x}^t_i$.

For any $\epsilon > 0$, we say a $(t, M, f, \mathbf{g})$ index code is *valid* (with respect to ϵ) if and only if (iff) the average probability of error satisfies $\mathrm{P_e} \doteq \mathrm{P}\{(\hat{X}_{[n]}) \neq (X_{[n]})\} \leq \epsilon$. We say a compression rate R is achievable iff for every $\epsilon > 0$, there exists a valid $(t, M, f, \mathbf{g})$ code such that $R \geq (\log M)/t$.

The *optimal* compression rate $\mathcal{R}$, also referred to as the *broadcast rate*, can be defined as

$$\mathcal{R} = \lim_{\epsilon \to 0} \lim_{t \to \infty} \inf_{\text{valid } (t, M, f, \mathbf{g}) \text{ code}} \frac{\log M}{t}. \quad (1)$$

We say a $(t, M, f, \mathbf{g})$ code is valid with respect to zero-error decoding iff the average probability of error is zero. The zero-error broadcast rate ρ can then be defined as

$$\rho \doteq \lim_{t \to \infty} \inf_{\substack{\text{valid } (t, M, f, \mathbf{g}) \text{ code w.r.t.} \\ \text{zero-error decoding}}} \frac{\log M}{t}. \quad (2)$$

Clearly, by definition, we always have $\mathcal{R} \leq \rho$.

The side information availability at receivers for a specific index coding instance can be represented by a sequence $(i|j \in A_i), i \in [n]$. Alternatively, it can be characterized by a family of *confusion graphs*, $(\Gamma_t, t \in \mathbb{Z}^+)$ [17]. For a given sequence length t, the confusion graph Γ_t is an undirected graph defined on the message sequence tuple alphabet $\mathcal{X}^t_{[n]}$. That is, $V(\Gamma_t) = \mathcal{X}^t_{[n]}$. Vertex $x^t_{[n]}$ in Γ corresponds to the realization $x^t_{[n]}$. Any two different vertices $x^t_{[n]}, z^t_{[n]}$ are adjacent in Γ_t iff $x^t_i \neq z^t_i$ and $x^t_{A_i} = z^t_{A_i}$ for some receiver $i \in [n]$. We call any pair of vertices satisfying this condition *confusable*. Hence, $E(\Gamma_t) = \{\{x^t_{[n]}, z^t_{[n]}\} : x^t_i \neq z^t_i \text{ and } x^t_{A_i} = z^t_{A_i} \text{ for some } i \in [n]\}$.

For correct decoding at all receivers, any two realizations $x^t_{[n]}, z^t_{[n]}$ can be mapped to the same codeword y with nonzero probabilities iff they are not confusable [17]. 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 [18]).

Consider any set $S \subseteq [n]$. The subproblem induced by S is jointly characterized by the distribution P_{X_S} and the sequence $(i|A_i \cap S), i \in S$. Let $\Gamma_t(S)$ denote the confusion graph of sequence length t of the subproblem induced by S. Let $\mathcal{R}(S)$ and $\rho(S)$ denote the broadcast rate and zero-error broadcast rate of the subproblem induced by S, respectively.

Preliminaries on $\mathcal{R}$ and ρ: Consider any index coding problem characterized by confusion graphs $(\Gamma_t, t \in \mathbb{Z}^+)$ and distribution $P_{X_{[n]}}$. We start with the following lemma.

Lemma 1: To characterize the broadcast rate $\mathcal{R}$ and zero-error broadcast rate ρ, it suffices to only consider index codes with deterministic encoding function f.

The above lemma can be simply proved by showing that given any valid index code with a stochastic encoding function, one can construct another valid code with a deterministic encoding function and same or smaller compression rate.

[1] While a common assumption in most index coding literatures is that the messages are independent and uniformly distributed, here we consider the more general case with arbitrary joint distribution for the messages.

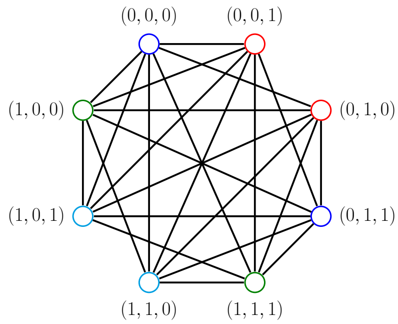

Figure 2. The confusion graph Γ_1 with $t = 1$ for the 3-message index coding instance $(1|-)$, $(2|3)$, $(3|2)$. Note that, for example, $x_{[n]} = (0, 0, 0)$ and $z_{[n]} = (0, 0, 1)$ are confusable because $x_3 = 0 \neq z_3 = 1$ and $x_{A_3} = x_2 = 0 = z_2 = z_{A_3}$. Suppose $(0, 0, 0)$ and $(0, 0, 1)$ are mapped to the same codeword y with certain nonzero probabilities. Then upon receiving this y, receiver 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 to the fractional chromatic number, both of which equal to 4. We have drawn an optimal coloring scheme with 4 colors in the graph.

Most existing results in the literature on the optimal compression rate (vanishing or zero error) of index coding were established assuming deterministic encoding functions. Lemma 1 indicates that those results can be directly applied to characterizing $\mathcal{R}$ and ρ.

Since we are considering fixed-length codes (rather than variable-length codes), the zero-error broadcast rate ρ does not depend on $P_{X_{[n]}}$ and can be characterized solely by the confusion graphs $(\Gamma_t, t \in \mathbb{Z}^+)$ [17] as

$$\rho = \lim_{t\to\infty} \frac{1}{t} \log \chi(\Gamma_t) \overset{(a)}{=} \lim_{t\to\infty} \frac{1}{t} \log \chi_{\mathrm{f}}(\Gamma_t), \tag{3}$$

where $\chi(\cdot)$ and $\chi_{\mathrm{f}}(\cdot)$ respectively denote the chromatic number and fractional chromatic number of a graph, and the proof of (a) can be found in [3, Section 3.2].

It has been shown [19] that, with the messages $X_{[n]}$ being uniformly distributed and independent of each other, the vanishing-error broadcast rate $\mathcal{R}$ equals to the zero-error broadcast rate ρ. Such equivalence does not hold for a general distribution $P_{X_{[n]}}$ as it has been shown in [20] that the (vanishing-error) broadcast rate $\mathcal{R}$ can be strictly smaller than its zero-error counterpart ρ.

Leakage to a guessing adversary:

We assume the adversary knows messages X_P and tries to guess the remaining messages X_Q, where $Q = [n] \setminus P$, via maximum likelihood estimation within a number of trials. In other words, the adversary generates a list of certain size of guesses, and is satisfied iff the true message sequence is in the list. We characterize the number of guesses the adversary can make by a function of sequence length, $c : \mathbb{Z}^+ \to \mathbb{Z}^+$, namely, the guessing capability function. We assume $c(t)$ to be non-decreasing and upper-bounded[2] by $\alpha(\Gamma_t(Q))$, where $\alpha(\cdot)$ denotes the independence number of a graph.

Consider any valid $(t, M, f, \mathbf{g})$ index code. Before eavesdropping the codeword y, the expected probability of the adversary successfully guessing x_Q^t with $c(t)$ number of guesses is

$$\mathrm{P_s}(X_P^t) = \mathbb{E}_{X_P^t}\left[\max_{K\subseteq\mathcal{X}_Q^t:|K|\le c(t)} \sum_{x_Q^t\in K} P_{X_Q^t|X_P^t}(x_Q^t|X_P^t)\right],$$

and the expected successful guessing probability after observing y is

$$\mathrm{P_s}(X_P^t, Y) = \mathbb{E}_{Y,X_P^t}\left[\max_{\substack{K\subseteq\mathcal{X}_Q^t:\\|K|\le c(t)}} \sum_{x_Q^t\in K} P_{X_Q^t|Y,X_P^t}(x_Q^t|Y, X_P^t)\right].$$

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_Q *after and before* observing the transmitted codeword y. That is,

$$L \doteq \log \frac{\mathrm{P_s}(X_P^t, Y)}{\mathrm{P_s}(X_P^t)}. \tag{4}$$

The (optimal) leakage rate can then be defined as

$$\mathcal{L} \doteq \lim_{\epsilon\to 0} \lim_{t\to\infty} t^{-1} \inf_{\text{valid } (t,M,f,\mathbf{g}) \text{ code}} L. \tag{5}$$

Remark 1: It can be readily verified that the leakage metric L is always non-negative. When $c(t) = 1$ (i.e., the adversary only makes a single guess after each observation), L reduces to the min-entropy leakage [12]. When $c(t) = 1$ and the messages are uniformly distributed, L is equal to the maximal leakage [14] and the maximum min-entropy leakage [13].

If we require zero-error decoding at receivers, the zero-error (optimal) leakage rate λ can be similarly defined as

$$\lambda \doteq \lim_{t\to\infty} t^{-1} \inf_{\substack{\text{valid } (t,M,f,\mathbf{g}) \text{ code w.r.t.}\\ \text{zero-error decoding}}} L. \tag{6}$$

By definition, we always have $\mathcal{L} \le \lambda$.

III. Information Leakage in Index Coding

A. Leakage Under A General Message Distribution

Consider any index coding problem $(i|j \in A_i), i \in [n]$ with confusion graphs $(\Gamma_t, t \in \mathbb{Z}^+)$ and distribution $P_{X_{[n]}}$. Our main result is the following theorem.

Theorem 1: For the vanishing-error leakage rate $\mathcal{L}$, we have

$$\rho(Q) - |Q| + \log \frac{1}{\sum_{x_P} \max_{x_Q} P_{X_{[n]}}(x_{[n]})} \le \mathcal{L} \le \mathcal{R}(Q). \tag{7}$$

For the zero-error leakage rate λ, we have

$$\rho(Q) - |Q| + \log \frac{1}{\sum_{x_P} \max_{x_Q} P_{X_{[n]}}(x_{[n]})} \le \lambda \le \rho(Q). \tag{8}$$

In the following, we prove the lower and upper bounds in (7). As for (8), the lower bound follows directly from the lower bound in (7) and the fact that $\mathcal{L} \le \lambda$, and the upper bound can be shown using similar techniques to the proof of the upper bound in (7).

Proof of the lower bound in (7)*:* Consider any $\epsilon > 0$ and any valid $(t, M, f, \mathbf{g})$ index code for which $\mathrm{P_e} \le \epsilon$.

Consider any codeword $y \in \mathcal{Y}$ and any realization $x_P^t \in \mathcal{X}_P^t$. Let $G_{\mathcal{X}_Q^t}(y, x_P^t)$ denote the collection of realizations

[2]It can be verified that if for some t we have $c(t) > \alpha(\Gamma_t(Q))$, then the probability of the adversary successfully guessing x_Q^t after observing y is at least $1 - \mathrm{P_e}$, which tends to 1 as ϵ tends to 0, making the problem trivial.

x_Q^t such that (y, x_P^t, x_Q^t) has nonzero probability, and for the event that $x_{[n]}^t = (x_P^t, x_Q^t)$ is the true message sequence tuple realization and y is the codeword realization, every receiver can correctly decode its requested message. That is,

$$G_{\mathcal{X}_Q^t}(y, x_P^t) = \{x_Q^t \in \mathcal{X}_Q^t : g_i(y, x_{A_i}^t) = x_i^t, \forall i \in [n]\}$$

Then, we have

$$\sum_{y, x_P^t} \sum_{x_Q^t \in G_{\mathcal{X}_Q^t}(y, x_P^t)} P_{Y, X_{[n]}^t}(y, x_P^t, x_Q^t) = 1 - \mathrm{P_e} \geq 1 - \epsilon. \tag{9}$$

We also have

$$|G_{\mathcal{X}_Q^t}(y, x_P^t)| \leq \alpha(\Gamma_t(Q)), \tag{10}$$

which can be shown by contradiction as follows. Assume there exists two different $x_{[n]}^t, z_{[n]}^t \in \mathcal{X}_{[n]}^t$, such that $x_P^t = z_P^t$, $x_Q^t \in G_{\mathcal{X}_Q^t}(y, x_P^t)$, $z_Q^t \in G_{\mathcal{X}_Q^t}(y, x_P^t)$, and x_Q^t and z_Q^t are adjacent (i.e., confusable) in $\Gamma_t(Q)$. Hence, there exists some receiver $i \in Q$ such that $x_i^t \neq z_i^t$ and $x_{A_i \cap Q}^t = z_{A_i \cap Q}^t$. Then considering $x_{[n]}^t$ and $z_{[n]}^t$, since they have the same realizations for messages in P, we have $x_{A_i}^t = (x_{A_i \cap P}^t, x_{A_i \cap Q}^t) = (z_{A_i \cap P}^t, z_{A_i \cap Q}^t) = z_{A_i}^t$. From the perspective of receiver i, upon receiving codeword y and observing side information $x_{A_i}^t = z_{A_i}^t$, it cannot tell whether the true sequence for message i is x_i^t or z_i^t. Therefore, with the transmitted codeword being y, either $x_{[n]}^t$ or $z_{[n]}^t$ being the true realization will lead to an erroneous decoding at receiver i, which contradicts the assumption that both x_Q^t and z_Q^t belong to $G_{\mathcal{X}_Q^t}(y, x_P^t)$. Therefore, any realizations $x_Q^t, z_Q^t \in G_{\mathcal{X}_Q^t}(y, x_P^t)$ must be not confusable and thus not adjacent to each other in $\Gamma_t(Q)$. In other words, the vertex subset $G_{\mathcal{X}_Q^t}(y, x_P^t) \subseteq V(\Gamma_t(Q))$ must be an independent set in $\Gamma_t(Q)$ and thus its cardinality is upper bounded by the independence number of $\Gamma_t(Q)$.

We lower bound $\mathrm{P_s}(X_P^t, Y)$, i.e., the adversary's expected successful guessing probability after observing Y, as

$$\begin{aligned}
&\sum_{y, x_P^t} \max_{K \subseteq \mathcal{X}_Q^t : |K| \leq c(t)} \sum_{x_Q^t \in K} P_{Y, X_{[n]}^t}(y, x_{[n]}^t) \\
&\geq \sum_{y, x_P^t} \max_{K \subseteq G_{\mathcal{X}_Q^t}(y, x_P^t) : |K| \leq c(t)} \sum_{x_Q^t \in K} P_{Y, X_{[n]}^t}(y, x_{[n]}^t) \\
&\geq \sum_{y, x_P^t} \frac{1}{|K \subseteq G_{\mathcal{X}_Q^t}(y, x_P^t) : |K| = c(t)^-|} \sum_{K \subseteq G_{\mathcal{X}_Q^t}(y, x_P^t) : |K| = c(t)^-} \sum_{x_Q^t \in K} P_{Y, X_{[n]}^t}(y, x_{[n]}^t) \\
&\overset{(a)}{=} \sum_{y, x_P^t} \frac{\binom{|G_{\mathcal{X}_Q^t}(y, x_P^t) - 1|}{c(t)^- - 1} \sum_{x_Q^t \in G_{\mathcal{X}_Q^t}(y, x_P^t)} P_{Y, X_{[n]}^t}(y, x_{[n]}^t)}{\binom{|G_{\mathcal{X}_Q^t}(y, x_P^t)|}{c(t)^-}} \\
&= \sum_{y, x_P^t} \frac{c(t)^-}{|G_{\mathcal{X}_Q^t}(y, x_P^t)|} \sum_{x_Q^t \in G_{\mathcal{X}_Q^t}(y, x_P^t)} P_{Y, X_{[n]}^t}(y, x_{[n]}^t) \\
&\overset{(b)}{\geq} \frac{c(t)(1-\epsilon)}{\alpha(\Gamma_t(Q))},
\end{aligned} \tag{11}$$

where $c(t)^- = \min\{c(t), |G_{\mathcal{X}_Q^t}(y, x_P^t)|\}$, and

- (a) follows from the fact that each $x_Q^t \in G_{\mathcal{X}_Q^t}(y, x_P^t)$ appears in exactly $\binom{|G_{\mathcal{X}_Q^t}(y, x_P^t) - 1|}{c(t)^- - 1}$ subsets of $G_{\mathcal{X}_Q^t}(y, x_P^t)$ of size $c(t)^-$,
- (b) follows from (9), (10), and that if $c(t) \leq |G_{\mathcal{X}_Q^t}(y, x_P^t)|$, then $\frac{c(t)^-}{|G_{\mathcal{X}_Q^t}(y, x_P^t)|} = \frac{c(t)}{|G_{\mathcal{X}_Q^t}(y, x_P^t)|} \geq \frac{c(t)}{\alpha(\Gamma_t(Q))}$, otherwise we have $c(t) > |G_{\mathcal{X}_Q^t}(y, x_P^t)|$ and thus $\frac{c(t)^-}{|G_{\mathcal{X}_Q^t}(y, x_P^t)|} = 1 \geq \frac{c(t)}{\alpha(\Gamma_t(Q))}$, where the inequality is due to the assumption that $c(t) \leq \alpha(\Gamma_t(Q))$.

For bounding $\mathrm{P_s}(X_P^t)$, consider any two disjoint subsets $A, B \subseteq [n]$. Note that any realization x_A^t can be explicitly denoted as $(x_{A,1}, x_{A,2}, \ldots, x_{A,t})$. We have

$$\begin{aligned}
&\sum_{x_A^t} \max_{x_B^t} P_{X_{A \cup B}^t}(x_A^t, x_B^t) \\
&= \sum_{x_A^t} \max_{x_B^t} \prod_{j \in [t]} P_{X_{A \cup B}}(x_{A,j}, x_{B,j}) \\
&= \Big(\sum_{x_A} \max_{x_B} P_{X_{A \cup B}}(x_A, x_B) \Big)^t,
\end{aligned} \tag{12}$$

where the last equality can be shown via induction.

Based on (11) and (12), we have

$$\begin{aligned}
\mathcal{L} &\geq \lim_{\epsilon \to 0} \lim_{t \to \infty} \frac{1}{t} \log \frac{\frac{c(t)(1-\epsilon)}{\alpha(\Gamma_t(Q))}}{\sum_{x_P^t} \max_{K \subseteq \mathcal{X}_Q^t : |K| \leq c(t)} \sum_{x_Q^t \in K} P_{X_{[n]}^t}(x_{[n]}^t)} \\
&\geq \lim_{\epsilon \to 0} \lim_{t \to \infty} \frac{1}{t} \log \frac{\frac{c(t)(1-\epsilon)}{\alpha(\Gamma_t(Q))}}{c(t) \cdot \sum_{x_P^t} \max_{x_Q^t} P_{X_{[n]}^t}(x_{[n]}^t)} \\
&= \lim_{\epsilon \to 0} \lim_{t \to \infty} \frac{1}{t} \log \frac{\frac{(1-\epsilon)|V(\Gamma_t(Q))|}{\alpha(\Gamma_t(Q))} \cdot \frac{1}{|V(\Gamma_t(Q))|}}{(\sum_{x_P} \max_{x_Q} P_{X_{[n]}}(x_{[n]}))^t} \\
&\overset{(c)}{=} \lim_{t \to \infty} \frac{1}{t} \log \chi_{\mathrm{f}}(\Gamma_t(Q)) + \log \frac{|\mathcal{X}|^{-|Q|}}{\sum_{x_P} \max_{x_Q} P_{X_{[n]}}(x_{[n]})} \\
&\overset{(d)}{=} \rho(Q) - |Q| + \log \frac{1}{\sum_{x_P} \max_{x_Q} P_{X_{[n]}}(x_{[n]})},
\end{aligned}$$

where (c) follows from the fact that for any vertex-transitive graph G, $\chi_{\mathrm{f}}(G) = |V(G)|/\alpha(G)$ [18, Proposition 3.1.1], and that any confusion graph for index coding is vertex-transitive [3, Section 11.4], and (d) follows from (3). ■

Proof of the upper bound in (7): Consider any decoding error $\epsilon > 0$. Construct a deterministic encoding function f that maps messages $X_{[n]}^t$ to codeword $Y = (Y_1, Y_2)$ according to the following rules.

1) Codeword Y_1 is generated from X_P^t according to some deterministic encoding function $f_1 : \mathcal{X}^{t|P|} \to \{1, 2, \ldots, |\mathcal{Y}_1|\}$ such that there exist some decoding functions $g_i, i \in P$ allowing zero-error decoding for all receivers $i \in P$ and that $t^{-1} \log |\mathcal{Y}_1| = \rho(P)$.
2) Codeword Y_2 is generated from X_Q^t according to some deterministic encoding function $f_2 : \mathcal{X}^{t|Q|} \to \{1, 2, \ldots, |\mathcal{Y}_2|\}$ such that there exist some decoding functions $g_i, i \in Q$ allowing ϵ-error decoding for all receivers $i \in Q$ and that $t^{-1} \log |\mathcal{Y}_2| = \mathcal{R}(Q)$.

Such encoding functions f_1 and f_2 exist for sufficiently large t. We further verify that the coding scheme described above leads to an average probability of error $\mathrm{P_e}$ no more than ϵ and thus is valid. Note that f_1 and f_2 are all deterministic. Define $B_{\mathcal{X}_Q^t} = \{x_Q^t$: there exists some $i \in Q$ such that $g_i(f_2(x_Q^t)) \neq x_i^t\}$. That is, $B_{\mathcal{X}_Q^t}$ denotes the set of x_Q^t for which there is at least one receiver $i \in Q$ that decodes erroneously. We have

$$\epsilon \geq \sum_{x_Q^t \in B_{\mathcal{X}_Q^t}} P_{X_Q^t}(x_Q^t). \tag{13}$$

Hence, we have

$$\begin{aligned}\mathrm{P_e} &= \sum_{x_P^t} \sum_{x_Q^t \in B_{\mathcal{X}_Q^t}} P_{X_{[n]}^t}(x_{[n]}^t) \\ &= \sum_{x_Q^t \in B_{\mathcal{X}_Q^t}} P_{X_Q^t}(x_Q^t) \sum_{x_P^t} P_{X_P^t|X_Q^t}(x_P^t|x_Q^t) \\ &= \sum_{x_Q^t \in B_{\mathcal{X}_Q^t}} P_{X_Q^t}(x_Q^t) \cdot 1 \leq \epsilon,\end{aligned}$$

where the last inequality follows from (13). Now we have shown that the proposed coding scheme is valid.

The optimal leakage rate is upper bounded by the rate of the information leakage of the proposed coding scheme as ϵ goes to 0. Let $P_{Y,X_{[n]}^t}$ denote the joint distribution of $Y = (Y_1, Y_2)$ and $X_{[n]}^t$ according to the proposed coding scheme. For any $x_P^t \in \mathcal{X}_P^t$ and $y_2 \in \mathcal{Y}_2$, define

$$\mathcal{X}_Q^t(x_P^t, y_2) = \{x_Q^t \in \mathcal{X}_Q^t : P_{Y_2,X_{[n]}^t}(y_2, x_P^t, x_Q^t) > 0\},$$

Then we have

$$\begin{aligned}\mathcal{L} &\leq \lim_{\epsilon\to 0}\lim_{t\to\infty} \frac{1}{t} \log \frac{\sum\limits_{\substack{x_P^t,\\ y_1,y_2}} \max\limits_{\substack{K\subseteq\mathcal{X}_Q^t:\\ |K|\leq c(t)}} \sum\limits_{x_Q^t\in K} P_{Y,X_{[n]}^t}(y_1,y_2,x_{[n]}^t)}{\sum\limits_{x_P^t} \max\limits_{\substack{K\subseteq\mathcal{X}_Q^t:\\ |K|\leq c(t)}} \sum_{x_Q^t\in K} P_{X_{[n]}^t}(x_{[n]}^t)} \\ &\overset{(a)}{=} \lim_{\epsilon\to 0}\lim_{t\to\infty} \frac{1}{t} \log \frac{\sum\limits_{x_P^t,y_2} \max\limits_{\substack{K\subseteq\mathcal{X}_Q^t(x_P^t,y_2):\\ |K|\leq c(t)}} \sum\limits_{x_Q^t\in K} P_{X_{[n]}^t}(x_{[n]}^t)}{\sum\limits_{x_P^t} \max\limits_{\substack{K\subseteq\mathcal{X}_Q^t:\\ |K|\leq c(t)}} \sum\limits_{x_Q^t\in K} P_{X_{[n]}^t}(x_{[n]}^t)} \\ &\overset{(b)}{\leq} \lim_{\epsilon\to 0}\lim_{t\to\infty} \frac{1}{t} \log \frac{|\mathcal{Y}_2| \cdot (\sum\limits_{x_P^t} \max\limits_{\substack{K\subseteq\mathcal{X}_Q^t:\\ |K|\leq c(t)}} \sum\limits_{x_Q^t\in K} P_{X_{[n]}^t}(x_{[n]}^t))}{\sum\limits_{x_P^t} \max\limits_{\substack{K\subseteq\mathcal{X}_Q^t:\\ |K|\leq c(t)}} \sum\limits_{x_Q^t\in K} P_{X_{[n]}^t}(x_{[n]}^t)} \\ &= \lim_{\epsilon\to 0}\lim_{t\to\infty} \frac{1}{t} \log |\mathcal{Y}_2| = \mathcal{R}(Q),\end{aligned}$$

where (a) follows from the definition of $\mathcal{X}_Q^t(x_P^t, y_2)$ and the fact that Y_1 is a deterministic function of X_P^t and Y_2 is a deterministic function of X_Q^t, and (b) follows from $\mathcal{X}_Q^t(x_P^t, y_2) \subseteq \mathcal{X}_Q^t$. ■

Remark 2: An interesting observation is that the bounds in Theorem 1 is independent of the guessing capability function $c(t)$. Whether $\mathcal{L}$ and λ depend on $c(t)$ remains unclear.

The upper and lower bounds in Theorem 1 do not match in general, as shown in the following example.

Consider the 4-message index coding problem $(1|4), (2|3), (3|2), (4|1)$, where the messages are binary and independent of each other with $P_{X_1}(0) = 1/4$ and $P_{X_1}(1) = 3/4$, and X_2, X_3, and X_4 all follow a uniform distribution. Consider an adversary knowing $X_P = X_4$ as side information, and thus $Q = \{1, 2, 3\}$. The broadcast rate for the subproblem induced by Q has been previously found [20] to be $\mathcal{R}(Q) = 3 - \frac{3}{4}\log 3$. By Theorem 1, the leakage rate $\mathcal{L}$ is upper bounded by $\mathcal{R}(Q)$, and lower bounded as

$$\begin{aligned}\mathcal{L} \geq& \rho(Q) - |Q| + \log \frac{1}{\sum_{x_P} \max_{x_Q} P_{X_{[n]}}(x_{[n]})} \\ =& 2 - 3 + \log \frac{1}{3/32 + 3/32} = 3 - \log 3.\end{aligned}$$

Note that $\rho(Q) = 2$ can be easily verified (for example, see [3, Section 8.6]). For the zero-error leakage rate λ, by (8) in Theorem 1, we have

$$3 - \log 3 \leq \lambda \leq 2.$$

B. Leakage Under A Uniform Message Distribution

In most existing works for index coding, the messages $X_{[n]}$ are assumed to be uniformly distributed and thus independent of each other. In such cases, Theorem 1 simplifies to the following corollary.

Corollary 1: If $P_{X_{[n]}}$ follows a uniform distribution, then

$$\mathcal{L} = \lambda = \mathcal{R}(Q) = \rho(Q). \tag{14}$$

Proof: We have

$$\begin{aligned}&\rho(Q) - |Q| + \log \frac{1}{\sum_{x_P} \max_{x_Q} P_{X_{[n]}}(x_{[n]})} \\ &= \rho(Q) - |Q| + \log \frac{1}{|\mathcal{X}|^{t|P|} \cdot (1/|\mathcal{X}|^{tn})} \\ &= \rho(Q) = \mathcal{R}(Q),\end{aligned}$$

where the last equality follows from the fact that the vanishing-error and zero-error broadcast rates are equal when messages are uniformly distributed [19]. Combining Theorem 1 and the above result yields (14). ■

Remark 3: Even though we have established the equivalence between the leakage and broadcast rates under uniform message distribution, a computable single-letter characterization of the value in (14) is unknown. Nevertheless, the equivalence between the leakage and broadcast rates means that the extensive results on the broadcast rate of index coding established in the literature (such as single-letter lower and upper bounds, explicit characterization for special cases, and structural properties) can be directly used to determine or bound the leakage rate.

Remark 4: As the leakage rate in (14) can be achieved by the proposed coding scheme in the achievability proof of Theorem 1, for any index coding instance with uniform message distribution satisfying $\mathcal{R} = \mathcal{R}(P) + \mathcal{R}(Q)$ (or equivalently, $\rho = \rho(P) + \rho(Q)$), we know that the broadcast rate and leakage rate can be simultaneously achieved by some deterministic index code.

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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] L. Ong, B. N. Vellambi, J. Kliewer, and P. L. Yeoh, "A code and rate equivalence between secure network and index coding," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 106–120, 2021.

[8] Y. Liu, P. Sadeghi, N. Aboutorab, and A. Sharififar, "Secure index coding with security constraints on receivers," in *Proc. IEEE Int. Symp. on Information Theory and its Applications (ISITA)*, Oct. 2020.

[9] V. Narayanan, J. Ravi, V. K. Mishra, B. K. Dey, N. Karamchandani, and V. M. Prabhakaran, "Private index coding," *arXiv preprint arXiv:2006.00257*, 2020.

[10] 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, 2019.

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

[12] 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.

[13] 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.

[14] 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.

[15] Y. Liu, L. Ong, S. Johnson, J. Kliewer, P. Sadeghi, and P. L. Yeoh, "Information leakage in zero-error source coding: A graph-theoretic perspective," in *Proc. IEEE Int. Symp. on Information Theory (ISIT)*, Melbourne, Australia, 2021.

[16] J. Körner, "Coding of an information source having ambiguous alphabet and the entropy of graphs," in *6th Prague Conference on Information Theory*, 1973, pp. 411–425.

[17] 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.

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

[19] 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.

[20] S. Miyake and J. Muramatsu, "Index coding over correlated sources," in *Proc. IEEE Int. Symp. on Network Coding (NetCod)*, Sydney, Australia, 2015, pp. 36–40.