Secure Data Storage Resilient Against Compromised Users via an Access Structure

Hassan ZivariFard and Rémi A. Chou
Department of Electrical Engineering and Computer Science
Wichita State University, Wichita, KS
E-mails:{hassan.zivarifard, remi.chou}@wichita.edu

***Abstract*—Consider a source and multiple users who observe the independent and identically distributed (i.i.d.) copies of correlated Gaussian random variables. The source wishes to compress and store its observation in a public database such that (i) authorized sets of users can reconstruct the source with some distortion level, and (ii) information leakage to non-authorized sets of colluding users is minimized. In other words, the recovery of the data is restricted to a predefined access structure of the users. One of the main results of this paper is a closed-form characterization of the fundamental trade-off between source coding rate and the information leakage rate when any authorized set of users has "better" side information than any set of unauthorized users.**

I. Introduction

A solution for secure data storage resilient to compromised users is distributed storage via traditional cryptographic solutions, i.e., traditional secret sharing [1], [2]. The idea behind such a solution is to avoid having a single point of entry that could compromise the private data in its entirety in the case of a security breach.

In this paper, we make three main modifications of the original secret sharing problem. First, we do not assume that secure channels are available between the users and the source, because secure channels come with a cost in practice, instead, we solely rely on public communication via a public database. Second, we consider that the users have side information about the source. While this consideration is not relevant in the original secret sharing problem, it becomes relevant in a data storage context. Not accounting for the fact that the users can have side information raises the following two issues that cannot been addressed with standard results for traditional secret sharing: (i) it leads to overestimating the security guarantees of the protocol, and (ii) it leads to inefficiency in terms of data storage size. Third, in our proposed setting, we consider a lossy reconstruction constraint, instead of a lossless reconstruction constraint in traditional secret sharing.

Two distinct bodies of work are related to our model. The first one is the literature on secret sharing, e.g., [1]–[3], which addresses the presence of an access structure for secure distributed data storage, and the second one is the literature on secure source coding, e.g., [4]–[10], which allows dealing with the presence of side information at the users. The problem studied in this paper proposes to simultaneously study the problems of having an access structure and considering that side information is available at the users.

Of particular relevance to this paper, [4], [5] have established the first characterization of the rate at which an encoder may compress a source such that an authorized user be able to recover the source in a lossless manner while guaranteeing a minimum information leakage from an unauthorized user who observes the encoded source. This problem is generalized in [6] to a scenario, in which the authorized user may recover the compressed source with some predefined distortion. Specifically, [6] characterized the optimal tradeoff between the rate, the desired distortion, and the information leakage when both the authorized and unauthorized users observe different i.i.d. side information sequences that are correlated with the compressed source. The secure lossy compression of a vector Gaussian source when both the authorized user and the unauthorized user have vector Gaussian side information have been studied in [9] and the authors derived inner and outer bounds on the optimal trade-off between the rate, the desired distortion, and the information leakage. [7], [8] extended this problem to a more general case in which the fidelity of the communication to the authorized user is measured by a distortion metric and the secrecy performance of the system is also evaluated under a distortion metric. Other related works include [11]–[13], where a function of a source must be reconstructed in a lossless manner by the authorized sets of users and must be kept secret from unauthorized sets of users, who all own side information about the source.

In this paper, we consider an encoder that wants to compress a source in such a way that (i) only pre-defined sets of authorized users can reconstruct, up to a prescribed distortion level, the source by pooling their side information, and (ii) information leakage about the source to any other sets of colluding users is minimized. This problem subsumes the secure lossy compression of a scalar Gaussian source when both the authorized user and the unauthorized user have scalar Gaussian side information as well as the secure lossy compression of a scalar Gaussian source when both the authorized user and the unauthorized user have vector Gaussian side information.

The key technical challenge consists in converting the problem to a scalar problem using sufficient statistics and dealing with the complexity of the compound structure raised by the presence of multiple authorized and unauthorized sets of users.

II. Preliminaries

Define $[\![a:b]\!] \triangleq [\lfloor a \rfloor, \lceil b \rceil] \cap \mathbb{N}^+$. Random variables are denoted by capital letters and their realizations by lower case letters. Vectors are denoted by boldface letters, e.g., $\mathbf{X}$ denotes a random vector and $\mathbf{x}$ denotes a realization of a random vector. $X^n_{\sim i}$ denotes the vector X^n except X_i. Throughout the paper, log denotes the base 2 logarithm.

III. Problem Statement

Consider any zero-mean Gaussian memoryless source $\big(\mathcal{X} \times \boldsymbol{\mathcal{Y}}_{\mathcal{L}}, P_{X\mathbf{Y}_{\mathcal{L}}}\big)$ with non-singular covariance matrix, where $\mathcal{L} \triangleq [\![1:L]\!]$ and $\mathbf{Y}_{\mathcal{L}} \triangleq (Y_\ell)_{\ell\in\mathcal{L}}$. The source generates n i.i.d. samples $(X_i, \mathbf{Y}_{\mathcal{L},i})_{i\in[\![1:n]\!]}$. Without loss of generality, let $Y_{\ell,i}$, for $\ell \in \mathcal{L}$, have the following form [14, Theorem 3.5.2],

$$Y_{\ell,i} = h_\ell X_i + N_{\ell,i}, \tag{1}$$

where $h_\ell \in \mathbb{R}\backslash 0$ and $N_{\ell,i}$, for $\ell \in \mathcal{L}$, are independent zero-mean Gaussian random variables with variance $\sigma'^2_\ell \geq 0$. Then, by normalizing (1), it is sufficient to consider $Y_{\ell,i}$, for $\ell \in \mathcal{L}$, to have the following form,

$$Y_{\ell,i} = X_i + N_{\ell,i}, \tag{2}$$

where $N_{\ell,i}$, for $\ell \in \mathcal{L}$, are independent zero-mean Gaussian random variables with variance $\sigma^2_\ell \geq 0$. Let $\mathbb{A}$ be a set of subsets of $\mathcal{L}$ such that for any $\mathcal{S} \subseteq \mathcal{L}$, if $\mathcal{S}$ includes a set that belongs to $\mathbb{A}$, then we must have $\mathcal{S} \in \mathbb{A}$, i.e., $\mathbb{A}$ has a monotone access structure [15]. Also, let $\mathbb{B} \triangleq 2^{\mathcal{L}}\backslash\mathbb{A}$ be the set of all colluding subsets of users for which the information leakage about the compressed source $\mathcal{X}^n$ must be minimized (see Fig.1). Let $d : \mathcal{X} \times \mathbf{Y}_{\mathcal{A}} \to [\![0 : d_{\max}]\!]$ be a finite distortion measure, i.e., such that $0 \leq d_{\max} < \infty$.

Definition 1. *A $(2^{nR}, n)$ source code for the memoryless source $\big(\mathcal{X} \times \boldsymbol{\mathcal{Y}}_{\mathcal{L}}, p_{X\mathbf{Y}_{\mathcal{L}}}\big)$ consists of*

- *An encoding function $f_E : x^n \mapsto m$, which assigns an index $m \in [\![1 : 2^{nR}]\!]$ to each $x^n \in \mathcal{X}^n$;*
- *Decoding functions $g_{\mathcal{A}} : m \times \boldsymbol{y}^n_{\mathcal{A}} \mapsto \hat{x}^n(\mathcal{A}) \cup \{\mathfrak{e}\}$, where $\mathcal{A} \in \mathbb{A}$ and $\hat{x}^n(\mathcal{A}) \in \mathcal{X}^n_{\mathcal{A}}$, which assign an estimate $\hat{x}^n(\mathcal{A})$ or an error $\mathfrak{e}$ to each $m \in [\![1 : 2^{nR}]\!]$ and $\boldsymbol{y}^n_{\mathcal{A}} \in \boldsymbol{\mathcal{Y}}^n_{\mathcal{A}}$.*

Definition 2. *A triple $(R, \Delta, D) \in \mathbb{R}^3_+$ is achievable if there exists a sequence of $(2^{nR}, n)$ source codes such that,*

$$\max_{\mathcal{A}\in\mathbb{A}} \limsup_{n\to\infty} \mathbb{E}\big[d\big(X^n, \hat{X}^n(\mathcal{A})\big)\big] \leq D, \tag{3a}$$

$$\max_{\mathcal{B}\in\mathbb{B}} \frac{1}{n} I(X^n; M, \boldsymbol{Y}^n_{\mathcal{B}}) \leq \Delta, \tag{3b}$$

where the distortion between sequences x^n and $\hat{x}^n(\mathcal{A})$ is defined by

$$d\big(x^n, \hat{x}^n(\mathcal{A})\big) \triangleq \frac{1}{n}\sum_{i=1}^{n} d\big(x_i, \hat{x}_i(\mathcal{A})\big). \tag{3c}$$

The set of all such achievable pairs is denoted by $\mathcal{R}(\mathbb{A})$ and is referred to as the rate-equivocation region. For a fixed $D > 0$, we define $\mathcal{R}(D, \mathbb{A}) \triangleq \{(R, \Delta) : (R, \Delta, D) \in \mathcal{R}(\mathbb{A})\}$.

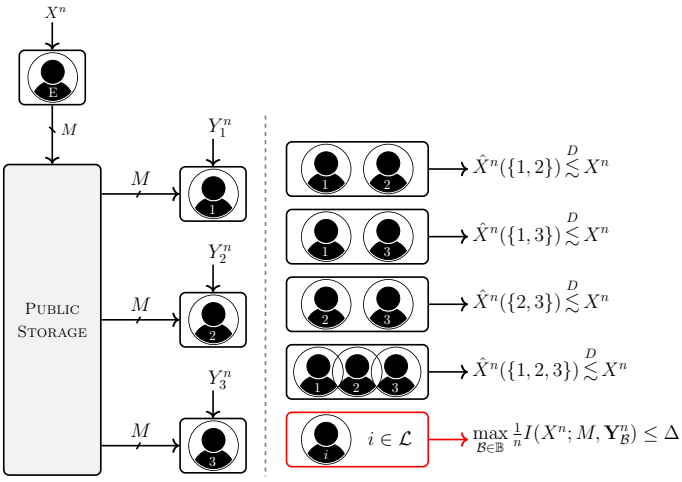

Fig. 1. Secure source coding with three users, i.e., $\mathcal{L} = \{1, 2, 3\}$, when any single user must not learn more than $n\Delta$ bits of information about the source X^n, i.e., we set $\mathbb{A} = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$, and $\mathbb{B} = \{\{1\}, \{2\}, \{3\}\}$. $\hat{X}^n(\{i, j\}) \overset{D}{\lesssim} X^n$, for $i, j \in \{1, 2, 3\}$ and $i \neq j$, means that the distortion between the reconstructed source by the users i and j together and the source sequence X^n must be less than D.

Here, the distortion of the reconstructed sequence $\big(\hat{X}_i(\mathcal{A})\big)^n_{i=1}$ defined in Definition 2 is measured by the mean square error as,

$$d\big(X_i, \hat{X}_i(\mathcal{A})\big) = \mathbb{E}\big[\big(X_i - \hat{X}_i(\mathcal{A})\big)^2\big] \leq nD. \tag{4}$$

Henceforth, we assume $0 \leq D \leq \sigma^2_{X|Y}$ where $\sigma^2_{X|Y}$ is the conditional variance of X given Y, $\sigma^2_{X|Y} = \mathbb{E}\big[\big(X - \mathbb{E}[X|Y]\big)^2\big|Y\big]$. Since the minimizer of the mean square error is the Minimum Mean-Square Error (MMSE) estimator, which is given by the conditional mean, we assume that the authorized terminals use this optimal estimator, which means that the authorized users select their reconstruction function as $\hat{X}_i(\mathcal{A}) = \mathbb{E}\big[X_i|\mathbf{Y}^n_{\mathcal{A}}, f(X^n)\big]$, for any $i \in [\![1 : n]\!]$.

IV. Main Results

The main result of this paper is a closed-form expression for the optimal trade-off between the compression rate and the leakage rate, which is provided in the following theorem.

Theorem 1. *For any access structure $(\mathbb{A}, \mathbb{B})$, when $\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\mathcal{A}^\star}\right) \leq \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\mathcal{B}^\star}\right)$,*

$$\mathcal{R}(D, \mathbb{A}) = \left\{ \begin{array}{l} (R, \Delta) : \\ R \geq \left[\frac{1}{2}\log\frac{\sigma^2_X}{D} - \frac{1}{2}\log\left(1 + \frac{\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\mathcal{A}^\star}\right)}{\sigma^{-2}_X}\right)\right]^+ \\ \Delta \geq \left[\frac{1}{2}\log\frac{\sigma^2_X}{D} - \frac{1}{2}\log\left(1 + \frac{\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\mathcal{A}^\star}\right)}{\sigma^{-2}_X}\right)\right]^+ \\ \qquad + \frac{1}{2}\log\left(1 + \frac{\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\mathcal{B}^\star}\right)}{\sigma^{-2}_X}\right) \end{array} \right\},$$

where $\mathcal{A}^\star \in \underset{\mathcal{A}\in\mathbb{A}}{\mathrm{argmin}}\{\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\mathcal{A}}\right)\}$, $\mathcal{B}^\star \in \underset{\mathcal{B}\in\mathbb{B}}{\mathrm{argmax}}\{\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\mathcal{B}}\right)\}$.

In Theorem 1, the term $\frac{1}{2}\log\frac{\sigma^2_X}{D}$ is the source coding capacity in the absence of side information [16, Theorem 3.6], the term $-\frac{1}{2}\log\left(1 + \frac{\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}_{\mathcal{A}^\star}\right)}{\sigma^{-2}_X}\right)$ is the gain provided

by the side information at the authorized users, the term $+\frac{1}{2}\log\left(1+\frac{\mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{B}^\star}^{-1}\right)}{\sigma_X^{-2}}\right)$ is the penalty rate coming from the side information at the unauthorized users.

V. Converse Proof of Theorem 1

We first consider the secure source coding problem as in [6, Section III-B], which consists of a source $\left((\mathcal{X},\mathcal{Y}_1,\mathcal{Y}_2),P_{XY_1Y_2}\right)$ with three outputs X^n at the encoder, Y_1^n at the legitimate terminal, and Y_2^n at the eavesdropper. The encoder wishes to encode its observed sequence in such a way that the legitimate receiver can reconstruct the source sequence X^n with distortion D, and the information leakage about X^n at the eavesdropper is minimized. Here, a (n,R)-code for source coding is defined by, an encoding function at the encoder $f:\mathcal{X}^n\to[\![1:2^{nR}]\!]$ and a decoding function $g:[\![1:2^{nR}]\!]\times\mathcal{Y}_1^n\to X^n$. In this problem, a tuple $(R,\Delta,D)\in\mathbb{R}_+^3$ is achievable if, for any $\epsilon>0$, there exists a $(n,R+\epsilon)$ code such that, $\mathbb{E}[d(x^n,\hat{X}(f(X^n),Y_1^n))]\le D+\epsilon$ and $\frac{1}{n}I(X^n;f(X^n),Y_2^n)\le\Delta-\epsilon$.

Now consider the secure source coding problem defined in Section III and the rate pair $(R,\Delta)\in\mathcal{R}(D,\mathbb{A})$. We first provide a general upper bound on the secure rate-distortion region of the problem defined above, and then show that this region reduces to the region in Theorem 1.

Theorem 2. *For every $(\mathbb{A},\mathbb{B})$, the region $\mathcal{R}(D,\mathbb{A})$ is included in $\bigcap_{(\mathcal{A},\mathcal{B})\in(\mathbb{A},\mathbb{B})}\mathcal{R}_G(\boldsymbol{Y}_\mathcal{A},\boldsymbol{Y}_\mathcal{B})$, where*

$$\mathcal{R}_G(\boldsymbol{Y}_\mathcal{A},\boldsymbol{Y}_\mathcal{B})\triangleq\bigcup_{\substack{U-V-X-(\boldsymbol{Y}_\mathcal{A},\boldsymbol{Y}_\mathcal{B})\\ \mathbb{E}\left[\sigma^2_{X|\boldsymbol{Y}_\mathcal{A},V}\right]\le D}}\left\{\begin{array}{l}(R,\Delta):\\ R>I(V;X|\boldsymbol{Y}_\mathcal{A})\\ \Delta>I(V;X)\\ \quad -I(V;\boldsymbol{Y}_\mathcal{A}|U)\\ \quad +I(X;\boldsymbol{Y}_\mathcal{B}|U)\end{array}\right\}.$$

The proof is similar to [6, Theorem 3] and is omitted for the sake of brevity.

In [6, Theorem 3], the equivocation is used as a measure of leakage but, since we consider continuous sources in our setting, to avoid a negative equivocation, we replace the equivocation with mutual information leakage (see Definition 2).

A. Conversion to a Scalar Problem

For every $(\mathcal{A},\mathcal{B})\in(\mathbb{A},\mathbb{B})$, the problem defined above can be seen as a secure source coding problem in which the encoder encodes the scalar Gaussian source X while the authorized and the unauthorized users observe a vector Gaussian source $\mathbf{Y}_\mathcal{A}$ and $\mathbf{Y}_\mathcal{B}$, respectively. Hence, the problem in (2) can be rewritten as follows,

$$\mathbf{Y}_\mathcal{A}=\mathbb{1}_\mathcal{A}X+\mathbf{N}_\mathcal{A},\quad \mathbf{Y}_\mathcal{B}=\mathbb{1}_\mathcal{B}X+\mathbf{N}_\mathcal{B},\tag{6}$$

where $\mathbf{N}_\mathcal{A}$ and $\mathbf{N}_\mathcal{B}$ are independent zero-mean Gaussian random vectors with covariance matrices $\mathbf{\Sigma}_\mathcal{A}\succ 0$ and $\mathbf{\Sigma}_\mathcal{B}\succ 0$, respectively, and $\mathbb{1}_\mathcal{A}$ is the all-ones vector with size $|\mathcal{A}|$. Also, $(\mathbf{N}_\mathcal{A},\mathbf{N}_\mathcal{B})$ is independent of X. To prove the converse part of Theorem 1, we use the following lemma from [17] to reduce the setting to a problem in which the encoder, the authorized users, and the unauthorized users observe a scalar Gaussian source.

Lemma 1. *([17, Lemma 3.1]) Consider the channel*

$$\boldsymbol{Y}=\boldsymbol{h}X+\boldsymbol{N},\tag{7}$$

where $\boldsymbol{N}$ is a Gaussian noise with zero mean and covariance matrix $\mathbf{\Sigma}$ and $\boldsymbol{h}\in\mathbb{R}^n$. A sufficient statistic to correctly determine X from $\boldsymbol{Y}$ is the following scalar

$$\tilde{Y}=\boldsymbol{h}^\intercal\mathbf{\Sigma}^{-1}\boldsymbol{Y}.\tag{8}$$

According to Lemma 1 the sufficient statistics to correctly determine X from $\mathbf{Y}_\mathcal{A}$ and $\mathbf{Y}_\mathcal{B}$ in (6) are the following scalars,

$$\tilde{Y}_\mathcal{A}=\mathbb{1}_\mathcal{A}^\intercal\mathbf{\Sigma}_\mathcal{A}^{-1}\mathbf{Y}_\mathcal{A},\quad \tilde{Y}_\mathcal{B}=\mathbb{1}_\mathcal{B}^\intercal\mathbf{\Sigma}_\mathcal{B}^{-1}\mathbf{Y}_\mathcal{B}.\tag{9}$$

Therefore, the channel described in (6) is equivalent to the following scalar Gaussian channel,

$$\tilde{Y}_\mathcal{A}=h_\mathcal{A}X+\tilde{N}_\mathcal{A},\quad \tilde{Y}_\mathcal{B}=h_\mathcal{B}X+\tilde{N}_\mathcal{B},\tag{10a}$$

where

$$h_\mathcal{A}=\mathbb{1}_\mathcal{A}^\intercal\mathbf{\Sigma}_\mathcal{A}^{-1}\mathbb{1}_\mathcal{A}=\mathrm{tr}\left(\mathbf{\Sigma}_\mathcal{A}^{-1}\right),\tag{10b}$$
$$h_\mathcal{B}=\mathbb{1}_\mathcal{B}^\intercal\mathbf{\Sigma}_\mathcal{B}^{-1}\mathbb{1}_\mathcal{B}=\mathrm{tr}\left(\mathbf{\Sigma}_\mathcal{B}^{-1}\right),\tag{10c}$$
$$\tilde{N}_\mathcal{A}=\mathbb{1}_\mathcal{A}^\intercal\mathbf{\Sigma}_\mathcal{A}^{-1}\mathbf{N}_\mathcal{A},\tag{10d}$$
$$\tilde{N}_\mathcal{B}=\mathbb{1}_\mathcal{B}^\intercal\mathbf{\Sigma}_\mathcal{B}^{-1}\mathbf{N}_\mathcal{B},\tag{10e}$$

where (10b) and (10c) follow since $\mathbf{\Sigma}_\mathcal{A}$ and $\mathbf{\Sigma}_\mathcal{B}$ are diagonal matrices. Now we have,

$$V^n-X^n-\mathbf{Y}_\mathcal{A}^n-\tilde{Y}_\mathcal{A}^n,\ V^n-X^n-\tilde{Y}_\mathcal{A}^n-\mathbf{Y}_\mathcal{A}^n,\tag{11a}$$
$$V^n-X^n-\mathbf{Y}_\mathcal{B}^n-\tilde{Y}_\mathcal{B}^n,\ V^n-X^n-\tilde{Y}_\mathcal{B}^n-\mathbf{Y}_\mathcal{B}^n,\tag{11b}$$

where

- the Markov chains in (11) follow since V^n is a function of X^n;
- the Markov chain in (11a) follows since V^n is a function of X^n and from $X-\tilde{Y}_\mathcal{A}-\mathbf{Y}_\mathcal{A}$ [18, Section 2.9];
- the Markov chain in (11b) follow by the same argument as the arguments for (11a).

Using (12) and the Markov chains in (11) one can show that the sufficient statistics in (10) preserve the distortion constraint and the leakage constraint in Definition 2, i.e.,

$$\mathbb{E}\left[d\left(X^n,\hat{X}^n(M,\mathbf{Y}_\mathcal{A}^n)\right)\right]\le D\Leftrightarrow$$
$$\mathbb{E}\left[d\left(X^n,\hat{X}_\mathcal{A}^n(M,\tilde{Y}_\mathcal{A}^n)\right)\right]\le D,\tag{12a}$$
$$I(X^n;M,\mathbf{Y}_\mathcal{B}^n)\le\Delta\Leftrightarrow I(X^n;M,\tilde{Y}_\mathcal{B}^n)\le\Delta.\tag{12b}$$

The proof of (12) follows from the Markov chains in (11) and the definition of the rate distortion which can be expressed as follows,

$$D\ge\frac{1}{n}\sum_{i=1}^n\mathbb{E}\Big[\Big(X_i-\mathbb{E}[X_i|V^n,\mathbf{Y}_\mathcal{A}^n]\Big)^2\Big].\tag{13}$$

Note that, we can write,

$$\tilde{Y}_\mathcal{B}=\frac{h_\mathcal{B}}{h_\mathcal{A}}\tilde{Y}_\mathcal{A}+N',\tag{14}$$

where $N' \sim \mathcal{N}\left(0, \mathrm{tr}(\mathbf{\Sigma}_{\mathcal{B}}^{-1})\left(1-\frac{\mathrm{tr}(\mathbf{\Sigma}_{\mathcal{B}}^{-1})}{\mathrm{tr}(\mathbf{\Sigma}_{\mathcal{A}}^{-1})}\right)\right)$. Therefore, when $\mathrm{tr}(\mathbf{\Sigma}_{\mathcal{B}}^{-1}) \leq \mathrm{tr}(\mathbf{\Sigma}_{\mathcal{A}}^{-1})$, without loss of generality we can convert the problem to the case where the unauthorized users side information is stochastically degraded with respect to the authorized users side information, i.e., $X - \tilde{Y}_{\mathcal{A}} - \tilde{Y}_{\mathcal{B}}$.

B. When Authorized Users Have Better Side Information

We study the case in which, for any access structure $(\mathcal{A},\mathcal{B}) \in (\mathbb{A},\mathbb{B})$, $\mathrm{tr}(\mathbf{\Sigma}_{\mathcal{B}}^{-1}) \leq \mathrm{tr}(\mathbf{\Sigma}_{\mathcal{A}}^{-1})$, so that the unauthorized users side information is stochastically degraded with respect to the side information of the authorized users. As discussed, by preserving the marginal distributions $P_{\tilde{Y}_{\mathcal{A}}|X}$ and $P_{\tilde{Y}_{\mathcal{B}}|X}$, we can transform the problem to a problem such that $U - V - X - \tilde{Y}_{\mathcal{A}} - \tilde{Y}_{\mathcal{B}}$, hence

$$\begin{aligned}
&I(V;X) - I(V;\tilde{Y}_{\mathcal{A}}|U) + I(X;\tilde{Y}_{\mathcal{B}}|U)\\
&\overset{(a)}{=} I(V;X) - I(V;\tilde{Y}_{\mathcal{A}}) + I(U;\tilde{Y}_{\mathcal{A}}) + I(X;\tilde{Y}_{\mathcal{B}}) - I(U;\tilde{Y}_{\mathcal{B}})\\
&\overset{(b)}{\geq} I(V;X) - I(V;\tilde{Y}_{\mathcal{A}}) + I(X;\tilde{Y}_{\mathcal{B}})\\
&\overset{(c)}{=} I(V;X|\tilde{Y}_{\mathcal{A}}) + I(X;\tilde{Y}_{\mathcal{B}}),
\end{aligned} \tag{15}$$

where (a), (b), and (c) follow since $U - V - X - \tilde{Y}_{\mathcal{A}} - \tilde{Y}_{\mathcal{B}}$. This implies that the region in Theorem 2 is included in the following region

$$\bigcap_{(\mathcal{A},\mathcal{B})\in(\mathbb{A},\mathbb{B})} \bigcup_{\substack{V-X-\tilde{Y}_{\mathcal{A}}-\tilde{Y}_{\mathcal{B}}\\ \mathbb{E}\left[\sigma^2_{X|\tilde{Y}_{\mathcal{A}},V}\right]\leq D}} \left\{\begin{array}{l}(R,\Delta):\\ R > I(V;X|\tilde{Y}_{\mathcal{A}})\\ \Delta > I(V;X|\tilde{Y}_{\mathcal{A}}) + I(X;\tilde{Y}_{\mathcal{B}})\end{array}\right\}. \tag{16}$$

Since the source is Gaussian the term $I(X;\tilde{Y}_{\mathcal{B}})$ is fixed, and we know that the term $I(V;X|\tilde{Y}_{\mathcal{A}}) = \mathbb{h}(X|\tilde{Y}_{\mathcal{A}}) - \mathbb{h}(X|\tilde{Y}_{\mathcal{A}},V)$ is minimized by joint Gaussian $(V,X,\tilde{Y}_{\mathcal{A}})$. Also, optimizing the rate and the leakage constraints in (16) separately results in a larger region, i.e., an outer bound. As a result, the region in (16) is included in the following region,

$$\bigcap_{(\mathcal{A},\mathcal{B})\in(\mathbb{A},\mathbb{B})} \left\{\begin{array}{l}(R,\Delta):\\ R > \min\limits_{\sigma^2_{X|\tilde{Y}_{\mathcal{A}},V}\leq D} \frac{1}{2}\log\frac{\sigma^2_{X|\tilde{Y}_{\mathcal{A}}}}{\sigma^2_{X|\tilde{Y}_{\mathcal{A}},V}}\\ \Delta > \min\limits_{\sigma^2_{X|\tilde{Y}_{\mathcal{A}},V}\leq D} \left[\frac{1}{2}\log\frac{\sigma^2_{X|\tilde{Y}_{\mathcal{A}}}}{\sigma^2_{X|\tilde{Y}_{\mathcal{A}},V}} + \frac{1}{2}\log\frac{\sigma^2_X}{\sigma^2_{X|\tilde{Y}_{\mathcal{B}}}}\right]\end{array}\right\}. \tag{17}$$

From the monotonicity of the log function the region above is included in,

$$\bigcap_{(\mathcal{A},\mathcal{B})\in(\mathbb{A},\mathbb{B})} \left\{\begin{array}{l}(R,\Delta):\\ R > \frac{1}{2}\log\frac{\sigma^2_{X|\tilde{Y}_{\mathcal{A}}}}{D}\\ \Delta > \frac{1}{2}\log\frac{\sigma^2_{X|\tilde{Y}_{\mathcal{A}}}}{D} + \frac{1}{2}\log\frac{\sigma^2_X}{\sigma^2_{X|\tilde{Y}_{\mathcal{B}}}}\end{array}\right\}. \tag{18}$$

Now, we have

$$\sigma^2_{X|\tilde{Y}_{\mathcal{A}}} = \sigma^2_X - \frac{\sigma^2_{X,\tilde{Y}_{\mathcal{A}}}}{\sigma^2_{\tilde{Y}_{\mathcal{A}}}} \overset{(a)}{=} \sigma^2_X - \frac{h^2_{\mathcal{A}}\sigma^4_X}{h^2_{\mathcal{A}}\sigma^2_X + \mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{A}}^{-1}\right)}$$

$$\overset{(b)}{=} \frac{\sigma^2_X}{\mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{A}}^{-1}\right)\sigma^2_X + 1}, \tag{19a}$$

where (a) follows by calculating $\sigma^2_{X,\tilde{Y}_{\mathcal{A}}}$ and $\sigma^2_{\tilde{Y}_{\mathcal{A}}}$ from (10); (b) follows since from (10) we have $h_{\mathcal{A}} = \mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{A}}^{-1}\right)$. Similarly, we have

$$\sigma^2_{X|\tilde{Y}_{\mathcal{B}}} = \frac{\sigma^2_X}{\mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{B}}^{-1}\right)\sigma^2_X + 1}. \tag{19b}$$

Substituting (19) in (18) and since the arguments of the log functions are decreasing in $\mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{A}}^{-1}\right)$ and increasing in $\mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{B}}^{-1}\right)$ we can compute the intersection in (18) and rewrite the region in (18) as follows,

$$\left\{\begin{array}{l}(R,\Delta):\\ R > \frac{1}{2}\log\frac{\sigma^2_X}{D\left(\mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{A}^\star}^{-1}\right)\sigma^2_X+1\right)}\\ \Delta > \frac{1}{2}\log\frac{\sigma^2_X}{D\left(\mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{A}^\star}^{-1}\right)\sigma^2_X+1\right)}\\ \quad + \frac{1}{2}\log\left(\mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{B}^\star}^{-1}\right)\sigma^2_X + 1\right)\end{array}\right\}, \tag{20}$$

where $\mathcal{A}^\star \in \underset{\mathcal{A}\in\mathbb{A}}{\mathrm{argmin}}\{\mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{A}}^{-1}\right)\}$ and $\mathcal{B}^\star \in \underset{\mathcal{B}\in\mathbb{B}}{\mathrm{argmax}}\{\mathrm{tr}\left(\mathbf{\Sigma}_{\mathcal{B}}^{-1}\right)\}$.

VI. Achievability Proof of Theorem 1

A. Discrete Alphabet

To prove the achievability of Theorem 1 we first provide an achievable rate region for the discrete alphabet and then extend this region to the continuous alphabet.

Theorem 3. *For any access structure $\mathbb{A}$ and $\mathbb{B}$, triple $(R,\Delta,D) \in \mathbb{R}^3_+$ is achievable if,*

$$\bigcup_{U-V-X-\mathbf{Y}_{\mathcal{L}}} \left\{\begin{array}{l}R > \max\limits_{\mathcal{A}\in\mathbb{A}}\{I(V;X|\mathbf{Y}_{\mathcal{A}})\}\\ \Delta > \max\limits_{(\mathcal{A},\mathcal{B})\in(\mathbb{A},\mathbb{B})} \big\{I(V;X)\\ \qquad -I(V;\mathbf{Y}_{\mathcal{A}}|U) + I(X;\mathbf{Y}_{\mathcal{B}}|U)\big\}\\ D > \max\limits_{\mathcal{A}\in\mathbb{A}} \mathbb{E}\big[d\big(X,\hat{X}_{\mathcal{A}}(V,\mathbf{Y}_{\mathcal{A}})\big)\big]\end{array}\right\} \tag{21}$$

The proof of Theorem 3 is similar to the Proof of [6, Theorem 3] and is omitted for the sake of brevity.

B. Continuous Alphabet

Now, we show that the region in Theorem 3 reduces to the region in Theorem 1 when the sources are Gaussian, as described in Section III, by quantizing the output of the Gaussian source $P_{X\mathbf{Y}_{\mathcal{L}}}$. The main problem of the quantization is that it can result in underestimating the information that the unauthorized users sets can learn about the source. However, we will show that we can overcome this problem if the quantization is fine enough. We now present the following lemma, which helps to extend the region in Theorem 3 to the continuous case by using quantization.

Lemma 2 ([18]–[20])**.** *Let X and Y be two real-valued random variables with distributions P_X and P_Y, respectively. Let $\mathcal{C}_{\Phi_1} = \{C_i\}_{i\in\mathcal{I}}$ and $\mathcal{K}_{\Phi_2} = \{K_j\}_{j\in\mathcal{J}}$ be two partitions of the real line for X and Y, respectively, such that for any*

$i \in \mathcal{I}$, $P_X[C_i] = \Phi_1$ *and for any* $j \in \mathcal{J}$, $P_Y[K_j] = \Phi_2$, *where* $\Phi_1 > 0$ *and* $\Phi_2 > 0$. *We denote the quantized versions of* X *and* Y *with respect to the partitions* $\mathcal{C}_{\Phi_1}$ *and* $\mathcal{K}_{\Phi_2}$ *by* X_{Φ_1} *and* Y_{Φ_2}, *respectively. Then, we have*

$$I(X;Y) = \lim_{\Phi_1,\Phi_2 \to 0} I(X_{\Phi_1}; Y_{\Phi_2}). \tag{22}$$

Note that, a quantization $\mathbf{Y}^n_{\mathcal{B}}$, $\mathcal{B} \in \mathbb{B}$ can lead to underestimation of $I(X^n; M, \mathbf{Y}^n_{\mathcal{B}})$. Next, we show that quantization does not affect the security constraint in Definition 2.

Lemma 3. *If the quantization* $X^n_{\Phi_1}$ *of* X^n *and* $\boldsymbol{Y}^n_{\mathcal{B},\Phi_2}$ *of* $\boldsymbol{Y}^n_{\mathcal{B}}$, *for every* $\mathcal{B} \in \mathbb{B}$, *is fine enough, then for every* $\epsilon > 0$,

$$\max_{\mathcal{B}\in\mathbb{B}} I(X^n; M, \boldsymbol{Y}^n_{\mathcal{B}}) \leq \max_{\mathcal{B}\in\mathbb{B}} I(X^n_{\Phi_1}; M, \boldsymbol{Y}^n_{\mathcal{B},\Phi_2}) + \delta. \tag{23}$$

Proof. For any $\epsilon > 0$, and for any $\mathcal{B} \in \mathbb{B}$, we have

$$\begin{aligned} I(X^n; M, \mathbf{Y}^n_{\mathcal{B}}) &\leq \left| I(X^n; M, \mathbf{Y}^n_{\mathcal{B}}) - I(X^n_{\Phi_1}; M, \mathbf{Y}^n_{\mathcal{B},\Phi_2}) \right| \\ &\quad + I(X^n_{\Phi_1}; M, \mathbf{Y}^n_{\mathcal{B},\Phi_2}) \qquad (24) \\ &\leq \max_{\mathcal{B}\in\mathbb{B}} \left| I(X^n; M, \mathbf{Y}^n_{\mathcal{B}}) - I(X^n_{\Phi_1}; M, \mathbf{Y}^n_{\mathcal{B},\Phi_2}) \right| \\ &\quad + \max_{\mathcal{B}\in\mathbb{B}} I(X^n_{\Phi_1}; M, \mathbf{Y}^n_{\mathcal{B},\Phi_2}) \qquad (25) \\ &\overset{(a)}{\leq} \delta + \max_{\mathcal{B}\in\mathbb{B}} I(X^n_{\Phi_1}; M, \mathbf{Y}^n_{\mathcal{B},\Phi_2}), \qquad (26) \end{aligned}$$

where (a) follows from Lemma 2 when the quantization $\mathbf{Y}^n_{\mathcal{B},\Phi_2}$ is fine enough, for any $\mathcal{B} \in \mathbb{B}$. Note that (26) is valid for any $\mathcal{B} \in \mathbb{B}$, therefore (26) results to the bound in Lemma 3. □

Considering Lemma 3 and choosing the quantization parameter Φ small enough one can show that the constraints in Theorem 3 for the continuous case reduces to,

$$R > \max_{\mathcal{A}\in\mathbb{A}} \big\{ I(V; X|\mathbf{Y}_{\mathcal{A}}) \big\}, \tag{27a}$$

$$\Delta > \max_{(\mathcal{A},\mathcal{B})\in(\mathbb{A},\mathbb{B})} \big\{ I(V;X) - I(V;\mathbf{Y}_{\mathcal{A}}|U) + I(X;\mathbf{Y}_{\mathcal{B}}|U) \big\}. \tag{27b}$$

Next, similar to what we have done in Section V-A, we convert the problem to a scalar problem by using sufficient statistics described in (10). When $\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}^\star}\right) \leq \mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{B}^\star}\right)$, where $\mathcal{A}^\star \in \underset{\mathcal{A}\in\mathbb{A}}{\mathrm{argmin}}\{\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}}\right)\}$, $\mathcal{B}^\star \in \underset{\mathcal{B}\in\mathbb{B}}{\mathrm{argmax}}\{\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{B}}\right)\}$, we choose the auxiliary random variable $U = \emptyset$ and choose the auxiliary random variable V to be jointly Gaussian random variable.

1) Conversion to a Scalar Problem via Sufficient Statistics: Since the side information $\mathbf{Y}_{\mathcal{A}}$ and $\mathbf{Y}_{\mathcal{B}}$ are vectors and we aim to find the relationship between $\sigma^2_{X|V}$ and $\sigma^2_{X|\mathbf{Y}_{\mathcal{A}},V}$, it is easier to work with scalar random variables and use sufficient statistics to evaluate the mutual information expressions in the achievable rate region provided in (27). When $\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}^\star}\right) \leq \mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{B}^\star}\right)$, we choose the auxiliary random variable $U = \emptyset$ and choose the auxiliary random variable V to be a jointly Gaussian random variable with X. Next, we show that using the sufficient statistics does not change the achievable rate region in Theorem 3. Similar to [18, Section 2.9] one can show that we have,

$$U - V - X - \tilde{Y}_{\mathcal{A}} - \mathbf{Y}_{\mathcal{A}},\ U - V - X - \mathbf{Y}_{\mathcal{A}} - \tilde{Y}_{\mathcal{A}}, \tag{28a}$$

$$U - V - X - \tilde{Y}_{\mathcal{B}} - \mathbf{Y}_{\mathcal{B}},\ U - V - X - \mathbf{Y}_{\mathcal{B}} - \tilde{Y}_{\mathcal{B}}. \tag{28b}$$

where $\tilde{Y}_{\mathcal{A}}$ and $\tilde{Y}_{\mathcal{B}}$ are defined in (10a). Hence,

$$I(V; X|\mathbf{Y}_{\mathcal{A}}) \overset{(a)}{=} I(V; X|\mathbf{Y}_{\mathcal{A}}, \tilde{Y}_{\mathcal{A}}) \overset{(b)}{=} I(V; X|\tilde{Y}_{\mathcal{A}}), \tag{29}$$

where (a) follows from (28a) and (b) follows from (28a). Similarly, one can also show that

$$I(V;\mathbf{Y}_{\mathcal{A}}|U) = I(V;\tilde{Y}_{\mathcal{A}}|U),\ I(V;\mathbf{Y}_{\mathcal{B}}|U) = I(V;\tilde{Y}_{\mathcal{B}}|U).$$

Since, when the source is Gaussian, we use the mean square error to measure the distortion of the reconstructed sequence and MMSE as the estimator, the distortion constraint in Theorem 3 reduces to $\sigma^2_{X|\mathbf{Y}_{\mathcal{A}},V} \leq D$. Considering the Markov chain $V - X - \tilde{Y}_{\mathcal{A}}$, $h_{\mathcal{A}} = \tilde{\sigma}^2_{\mathcal{A}} = \mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}}\right)$, and $h_{\mathcal{B}} = \tilde{\sigma}^2_{\mathcal{B}} = \mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{B}}\right)$ as showed in (10), we rewrite the achievable rate region in Theorem 3 as the union over the random variables V such that $V - X - \mathbf{Y}_{\mathcal{L}}$ of

$$\left\{ \begin{array}{l} (R,\Delta): \\ R > \max\limits_{\mathcal{A}\in\mathbb{A}} \left\{ \frac{1}{2} \log \frac{\sigma^2_X \left(\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}}\right)\sigma^2_{X|V}+1\right)}{\sigma^2_{X|V}\left(\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}}\right)\sigma^2_X+1\right)} \right\} \\ \Delta > \max\limits_{(\mathcal{A},\mathcal{B}\in(\mathbb{A},\mathbb{B}))} \left\{ \frac{1}{2} \log \frac{\sigma^2_X \left(\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}}\right)\sigma^2_{X|V}+1\right)}{\sigma^2_{X|V}\left(\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}}\right)\sigma^2_X+1\right)} \right. \\ \qquad \left. + \frac{1}{2} \log \left(\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{B}}\right) \sigma^2_X + 1 \right) \right\} \\ D \geq \max\limits_{\mathcal{A}\in\mathbb{A}} \left\{ \frac{\sigma^2_{X|V}}{\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}}\right)\sigma^2_{X|V}+1} \right\} \end{array} \right\}. \tag{30}$$

Next, since the argument of the log functions in the region above is decreasing in $\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}}\right)$ and increasing in $\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{B}}\right)$ we can solve the maximization over $(\mathcal{A},\mathcal{B})$ in the above region with $\mathcal{A}^\star \in \underset{\mathcal{A}\in\mathbb{A}}{\mathrm{argmin}}\{\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}}\right)\}$ and $\mathcal{B}^\star \in \underset{\mathcal{B}\in\mathbb{B}}{\mathrm{argmax}}\{\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{B}}\right)\}$. Therefore, choosing V such that $D = \sigma^2_{X|\tilde{Y}_{\mathcal{A}^\star},V}$, and considering (19), we can rewrite (30) as,

$$\left\{ \begin{array}{l} (R,\Delta): \\ R > \frac{1}{2} \log \frac{\sigma^2_X}{\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}^\star}\right)\sigma^2_X D + D} \\ \Delta > \frac{1}{2} \log \frac{\sigma^2_X}{\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{A}^\star}\right)\sigma^2_X D + D} + \frac{1}{2} \log \left(\mathrm{tr}\left(\mathbf{\Sigma}^{-1}_{\mathcal{B}^\star}\right) \sigma^2_X + 1 \right) \end{array} \right\}. \tag{31}$$

VII. Conclusion

We studied Gaussian secure lossy source coding in the presence of an access structure. When any authorized set of users has "better" side information than any set of unauthorized users, we derived the optimal trade-off between the source coding rate and the information leakage rate. When this is not the case, we also derived the optimal trade-off between the rate and the leakage, but this result is not reported here due to space constraint [21].

References

[1] G. R. Blakley, "Safeguarding cryptographic keys," in *Proc. AFIPS 79th Nat. Comput. Conf.*, New York, NY, USA, Jun. 1979, pp. 313–317.

[2] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, Nov. 1979.

[3] A. Beimel, "Secret-sharing schemes: A survey," in *International conference on coding and cryptology*. Springer, 2011, pp. 11–46.

[4] D. Gündüz, E. Erkip, and H. V. Poor, "Secure lossless compression with side information," in *Proc. IEEE Info. Theory Workshop (ITW)*, Porto, Portugal, May 2008, pp. 169–173.

[5] ——, "Lossless compression with security constraints," in *Proc. IEEE Int. Symp. on Info. Theory (ISIT)*, Toronto, Canada, Jul. 2008, pp. 111–115.

[6] J. Villard and P. Piantanida, "Secure multiterminal source coding with side information at the eavesdropper," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3668–3692, Jun. 2013.

[7] E. C. Song, P. Cuff, and H. V. Poor, "A rate-distortion based secrecy system with side information at the decoders," in *Proc. 52th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sep. 2014, pp. 755–762.

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

[9] E. Ekrem and S. Ulukus, "Secure lossy transmission of vector Gaussian sources," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5466–5487, Sep. 2013.

[10] R. Tandon, S. Ulukus, and K. Ramchandran, "Secure source coding with a helper," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2178–2187, Apr. 2013.

[11] R. Vidhi, R. A. Chou, and H. M. Kwon, "Information-theoretic secret sharing from correlated Gaussian random variables and public communication," *IEEE Trans. Inf. Theory*, vol. 68, no. 1, pp. 549–559, Jan. 2022.

[12] I. Csiszár and P. Narayan, "Capacity of a shared secret key," in *Proc. IEEE Int. Symp. on Info. Theory (ISIT)*, Austin, TX, USA, Jun. 2010, pp. 2593–2596.

[13] R. A. Chou, "Secret sharing over a public channel from correlated random variables," in *Proc. IEEE Int. Symp. on Info. Theory (ISIT)*, Vail, CO, USA, Jun. 2018, pp. 991–995.

[14] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[15] J. Benaloh and J. Leichter, "Generalized secret sharing and monotone functions," in *Proc. Conf. Theory Appl. Cryptogr.*, New York, NY, USA, Feb. 1988, pp. 27–35.

[16] A. El Gamal and Y.-H. Kim, *Network Information Theory*, 1st ed. Cambridge, U.K: Cambridge University Press, 2012.

[17] P. Parada and R. Blahut, "Secrecy capacity of SIMO and slow fading channels," in *Proc. IEEE Int. Symp. on Info. Theory (ISIT)*, Adelaide, SA, Australia, Sep. 2005, pp. 2152–2155.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. John Wiley & Sons, Inc., 2001.

[19] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. Holden-Day Inc., 1964.

[20] R. Fano, *Transmission of Information: A Statistical Theory of Communications*. Cambridge, MA, USA: MIT Press, 1961.

[21] H. ZivariFard and R. A. Chou, "Secure source coding resilient against compromised users via an access structure," *Submitted to IEEE Transactions on Information Theory*, Jun. 2022.