

Lossless Secure Source Coding: Yamamoto's Setting

Shahab Asoodeh, Fady Alajaji, and Tamás Linder

Department of Mathematics and Statistics, Queen's University

{asoodehshahab, fady, linder}@mst.queensu.ca

Abstract—Given a private source of information, X^n and a public correlated source, Y^n , we study the problem of encoding the two-dimensional source (X^n, Y^n) into an index J such that a remote party, knowing J and some external side information Z^n , can losslessly recover Y^n while any eavesdropper knowing J and possibly a correlated side information E^n can retrieve very little information about X^n . We give general converse results for the amount of information about X^n that might be leaked in such systems and also achievability results that are optimal in some special cases.

Index Terms—Equivocation, information leakage, utility, privacy, lossless source coding with side information.

I. INTRODUCTION

Information-theoretic secrecy models concern a tradeoff between utility and privacy. Given a source Y^n , the goal is to transmit this source securely and reliably over a noiseless public channel which might be perfectly observed by a passive adversary. The utility is defined as the accuracy in the recovering of Y^n by a remote receiver and the privacy is defined as the uncertainty of the source given the message sent over the channel. However, in some cases, it may be desirable to define utility and privacy for two different sources, that is, we want the receiver to know Y^n with some level of accuracy while revealing very little information about a correlated source X^n , which we refer to as the private source.

To motivate this setting, consider the following example. Suppose Y denotes an attribute of a bank customer that a trusted advertising company would like to target and X denotes another, more sensitive, attribute of the customer. The bank has database (X^n, Y^n) corresponding to n different users. The company pays the bank to receive Y^n as accurately as possible. However, some governing laws prohibit the database X^n from being revealed too extensively over public communication channels. Consequently, the data given to the company must be chosen so that at most a prescribed amount of information is revealed about X^n over the communication channel while the recovery of Y^n by the company satisfies some level of quality.

Inspired by Yamamoto [1] where a lossy source coding problem is studied under a privacy constraint, we consider a secure lossless source coding model in which an encoder (Alice) encodes a two-dimensional source (X^n, Y^n) such that the receiver (Bob) is able to reconstruct Y^n correctly

with high probability and the leakage of information (the information obtained by an eavesdropper, Eve) about X^n is no more than $\Delta \geq 0$. It is clear that no non-trivial level of privacy can be obtained if no side information is available to Bob. Hence, we assume Bob has access to some correlated side information and after observing the channel output wants to recover Y^n with asymptotically vanishing error probability. We study this problem in terms of the compression rate and also the information leakage about X^n (or equivalently the equivocation between the compressed and the private data). We give converse results for different cases including when Bob has coded or uncoded side information, when Eve has uncoded side information, or when the private source, X^n , is hidden even from Alice.

When $X = Y$, the problem we consider here reduces to a well-known model which has been extensively studied, for example see [2]–[6]. In particular, Prabhakaran and Ramchandran [2] considered a similar secure lossless setting with $X = Y$ and Bob and Eve having correlated uncoded side information. They focused on the best achievable information leakage rate when the public channel has not rate limit. Gündüz et al. [3], [4] gave converse and achievability bounds for a similar setting for both compression rate and information leakage which do not necessarily match. Tandon et al. [6] considered a simpler case in which Eve has no side information, gave a single letter characterization of the optimal rates, and information leakage and showed that a simple coding scheme based on binning, similar to the one proposed by Wyner in [7], is indeed optimal with and without the privacy constraint. Our results recover all these results in the special case of $X = Y$.

The rest of this paper is organized as follows. In Section II, we formally define our problem and state an outer bound which is our main result. In Section III, we consider a more general model in which Eve has side information and present another outer bound. We then present a coding scheme which is shown to be optimal in some special cases. We complete the paper with some concluding remarks in Section IV.

II. YAMAMOTO'S LOSSLESS SOURCE CODING: CODED SIDE INFORMATION AT BOB

Yamamoto [1] considered a lossy source coding scheme with a privacy constraint at the legitimate decoder. This is contrasted with the typical information-theoretic secrecy models in which the privacy is defined as the uncertainty

of the source against a passive eavesdropper. In this model, having observed (X^n, Y^n) , the encoder $\varphi : \mathcal{X}^n \times \mathcal{Y}^n \rightarrow \{1, 2, \dots, 2^{nR}\}$, transmits a message to the decoder, $\psi : \{1, 2, \dots, 2^{nR}\} \rightarrow \hat{\mathcal{Y}}^n$, which is required to recover Y^n within some distortion D while revealing little information about X^n . More precisely, for a given distortion measure $d : \mathcal{Y} \times \hat{\mathcal{Y}} \rightarrow \mathbb{R}_+$, we require $\frac{1}{n} \sum \mathbb{E}[d(Y_i, \hat{Y}_i)] \leq D$ while the normalized uncertainty about X^n at the decoder is lower-bounded, i.e., $\frac{1}{n} H(X^n | \varphi(X^n, Y^n)) \geq E$ for a non-negative $E \leq H(X)$. This requirement is different from the privacy constraint usually considered in information-theoretic secrecy (e.g., [3], [8], [6], and [5]), in that here the utility and privacy are measured with respect to two different sources Y and X , respectively. In this sense, X and Y correspond to the private and public sources, respectively. The correlation between X and Y makes the utility and privacy constraints contradicting.

We study a similar model as Yamamoto's but for *lossless* compression. Clearly, if no side information is available to the decoder, then the eavesdropper can obtain as much information about X^n as the legitimate decoder and hence only trivial levels of privacy can be achieved when lossless compression of Y is required. We, therefore, assume that side information is provided at the decoder, as depicted in Fig. 1.

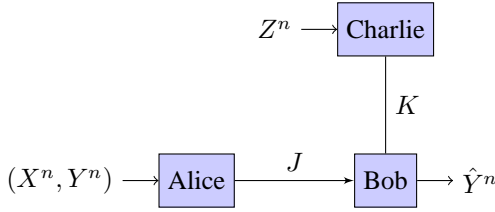


Fig. 1. Yamamoto's lossless source coding.

A $(2^{nR_A}, 2^{nR_C}, n)$ code for private lossless compression in this setup is composed of two encoding functions at Alice and Charlie, respectively, $f_A : \mathcal{X}^n \times \mathcal{Y}^n \rightarrow \{1, 2, \dots, 2^{nR_A}\}$ and $f_C : \mathcal{Z}^n \rightarrow \{1, 2, \dots, 2^{nR_C}\}$, and a decoder at Bob, $f_B : \{1, 2, \dots, 2^{nR_A}\} \times \{1, 2, \dots, 2^{nR_C}\} \rightarrow \hat{\mathcal{Y}}^n$, where (X^n, Y^n, Z^n) are n independent and identically distributed (i.i.d.) copies of (X, Y, Z) with joint distribution $P(x, y, z)$. We assume that both encoders communicate to Bob over noiseless channels; however, the channel between Alice and Bob is subject to eavesdropping and hence a passive party can have access to the message J transmitted over this channel. A triple $(R_A, R_C, \Delta) \in \mathbb{R}_+^3$ is said to be achievable if for any $\varepsilon > 0$, there exists a $(2^{nR_A}, 2^{nR_C}, n)$ code for n large enough such that

$$\Pr(f_B(J, K) \neq Y^n) < \varepsilon, \quad (1)$$

$$\frac{1}{n} H(X^n | J) \geq \Delta - \varepsilon, \quad (2)$$

where $J := f_A(X^n, Y^n)$ and $K := f_C(Z^n)$. We denote the set of all achievable triples (R_A, R_C, Δ) by \mathcal{R} . One special case of interest is when J contains absolutely no information about the private source, that is, when J is independent of X^n , which is called perfect privacy.

We note that for a special case of $X = Y$, inner and outer bounds on the achievable region were initially presented in [4, Theorem 3.1], although these bounds do not match in general. Tight bounds were then given in [6, Theorem 1] whose achievability resembles the binning scheme proposed by Wyner [7] for standard source coding with coded side information at the decoder. This therefore shows that the privacy constraint (2) does not change the optimal scheme.

Theorem 1. For any achievable triple $(R_A, R_C, \Delta) \in \mathcal{R}$ we have

$$\begin{aligned} R_A &\geq H(Y|V), \\ R_C &\geq I(Z; V), \\ \Delta &\leq I(X, Y; V) + H(X|U) - H(Y|U), \end{aligned}$$

for some auxiliary random variables $V \in \mathcal{V}$ and $U \in \mathcal{U}$ such that $P(x, y, z, u, v) = P(x, y, z)P(v|z)P(u|x, y)$ with $|\mathcal{U}| \leq |\mathcal{X}| \times |\mathcal{Y}| + 1$ and $|\mathcal{V}| \leq |\mathcal{Z}| + 2$.

Proof. First note that Bob is required to reconstruct Y^n losslessly given J and K , and thus by Fano's inequality we have

$$H(Y^n | J, K) \leq n\varepsilon_n, \quad (3)$$

where $\varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$.

We start by obtaining a lower bound for R_A as follows:

$$\begin{aligned} nR_A &\geq H(J) \geq H(J|K) \\ &= H(Y^n, J|K) - H(Y^n|J, K) \\ &\stackrel{(a)}{\geq} H(Y^n, J|K) - n\varepsilon_n \\ &\geq H(Y^n|K) - n\varepsilon_n \\ &= \sum_{i=1}^n H(Y_i|Y^{i-1}, K) - n\varepsilon_n \\ &\geq \sum_{i=1}^n H(Y_i|Y^{i-1}, X^{i-1}, K) - n\varepsilon_n \\ &\stackrel{(b)}{=} \sum_{i=1}^n H(Y_i|V_i) - n\varepsilon_n \\ &\stackrel{(c)}{=} H(Y_Q|V_Q, Q) - n\varepsilon_n \\ &\stackrel{(d)}{=} nH(Y|V) - n\varepsilon_n \end{aligned}$$

where (a) follows from (3), and (b) is due to the definition $V_i := (Y^{i-1}, X^{i-1}, K)$. In (c) we have introduced a time-sharing random variable Q which is distributed uniformly over $\{1, 2, \dots, n\}$ and is independent of (X^n, Y^n, Z^n) . In (d) we have defined $V := (V_Q, Q)$ and used the fact that Y_Q has the distribution of Y and hence we can replace Y_Q with Y .

Next we obtain a lower bound on R_C :

$$\begin{aligned} nR_C &\geq H(K) = I(Z^n; K) = \sum_{i=1}^n I(Z_i; K|Z^{i-1}) \\ &\stackrel{(a)}{=} \sum_{i=1}^n I(Z_i; K, Z^{i-1}) \end{aligned}$$

$$\begin{aligned}
&\stackrel{(b)}{=} \sum_{i=1}^n I(Z_i; K, Z^{i-1}, X^{i-1}, Y^{i-1}) \\
&\geq \sum_{i=1}^n I(Z_i; K, X^{i-1}, Y^{i-1}) = nI(Z_Q; V_Q, Q) \\
&= nI(Z; V)
\end{aligned}$$

where (a) is due to the fact that Z_i is independent of Z^{i-1} for each i and (b) follows from the Markov chain relation $Z_i \text{---} (K, Z^{i-1}) \text{---} (Y^{i-1}, X^{i-1})$.

We now upper bound the equivocation that any asymptotically lossless scheme produces. First we show the following identity which expresses $H(X^n|J)$ in terms of $H(Y^n|J)$ and some auxiliary terms:

$$H(X^n|J) - H(Y^n|J) = \sum_{i=1}^n [H(X_i|U_i) - H(Y_i|U_i)], \quad (4)$$

where $U_i := (X_{i+1}^n, Y^{i-1}, J)$. We will prove a general version of this identity later in Lemma 1.

The equivocation can then be upper bounded as

$$\begin{aligned}
n(\Delta - \varepsilon) &\leq H(X^n|J) \\
&\stackrel{(a)}{=} H(Y^n|J) + \sum_{i=1}^n [H(X_i|U_i) - H(Y_i|U_i)] \\
&= H(Y^n|K, J) + I(Y^n; K|J) \\
&\quad + \sum_{i=1}^n [H(X_i|U_i) - H(Y_i|U_i)] \\
&\leq n\varepsilon_n + I(K; Y^n, X^n|J) \\
&\quad + \sum_{i=1}^n [H(X_i|U_i) - H(Y_i|U_i)] \\
&\stackrel{(b)}{\leq} n\varepsilon_n + I(K; X^n, Y^n) \\
&\quad + \sum_{i=1}^n [H(X_i|U_i) - H(Y_i|U_i)] \\
&= n\varepsilon_n + \sum_{i=1}^n I(K; X_i, Y_i|X^{i-1}, Y^{i-1}) \\
&\quad + \sum_{i=1}^n [H(X_i|U_i) - H(Y_i|U_i)] \\
&= n\varepsilon_n + \sum_{i=1}^n I(K, X^{i-1}, Y^{i-1}; X_i, Y_i) \\
&\quad + \sum_{i=1}^n [H(X_i|U_i) - H(Y_i|U_i)] \\
&= n\varepsilon_n + \sum_{i=1}^n I(V_i; X_i, Y_i) \\
&\quad + \sum_{i=1}^n [H(X_i|U_i) - H(Y_i|U_i)] \\
&= n\varepsilon_n + nI(V_Q; X_Q, Y_Q|Q) \\
&\quad + n[H(X_Q|U_Q, Q) - H(Y_Q|U_Q, Q)] \\
&\stackrel{(c)}{=} n\varepsilon_n + nI(V_Q, Q; X_Q, Y_Q)
\end{aligned}$$

$$\begin{aligned}
&+ n[H(X_Q|U_Q, Q) - H(Y_Q|U_Q, Q)] \\
&\stackrel{(d)}{=} n\varepsilon_n \\
&+ n[I(V; X, Y) + H(X|U) - H(Y|U)],
\end{aligned}$$

where (a) follows from (4), (b) follows from the Markov chain relation $J \text{---} (X^n, Y^n) \text{---} K$ and hence $I(X^n, Y^n; K|J) \leq I(X^n, Y^n; K)$, (c) is due to the fact that Q is independent of (X_Q, Y_Q) and in (d) we have introduced $U := (U_Q, Q)$.

We note that by definitions of U and V , the Markov chain conditions $(X, Y) \text{---} Z \text{---} V$ and $Z \text{---} (X, Y) \text{---} U$ are satisfied. The cardinality bounds given in the statement of the theorem can be proved using support lemma [9]. ■

Remark 1. As mentioned earlier, the special case $X = Y$ is studied in [6] where it is shown that for any achievable triple (R_A, R_C, Δ) , the optimal equivocation satisfies $\Delta \leq I(Y; V)$. We see that Theorem 1 yields the same result and thus gives a tight bound in this special case.

In practice, the private source X might not be directly available to Alice. In this case, her mapping is $f_A : \mathcal{Y}^n \rightarrow \{1, 2, \dots, 2^{nR_A}\}$ and the above theorem reduces to the following corollary.

Corollary 1. When the source X^n is not available to Alice, any achievable triple (R_A, R_C, Δ) satisfies

$$\begin{aligned}
R_A &\geq H(Y|V), \\
R_C &\geq I(Z; V), \\
\Delta &\leq I(Y; V) + H(X|U) - H(Y|U),
\end{aligned}$$

for some $U \in \mathcal{U}$ and $V \in \mathcal{V}$ such that $P(x, y, z, u, v) = P(x, y, z)P(v|z)P(u|y)$ and $|\mathcal{U}| \leq |\mathcal{Y}|+1$ and $|\mathcal{V}| \leq |\mathcal{Z}|+2$.

Proof. The proof follows easily from the proof of Theorem 1. In particular, introducing $V_i := (Y^{i-1}, K)$ and $U_i := (X_{i+1}^n, Y^{i-1}, J)$, we can follow easily the chain of inequalities given for the equivocation analysis with appropriate modifications. Since now $J = f_A(Y^n)$, we have $(X_i, Z_i) \text{---} Y_i \text{---} U_i$. ■

III. YAMAMOTO'S LOSSLESS SOURCE CODING: UNCODED SIDE INFORMATION AT EVE

We now turn our focus to the case where there is an eavesdropper, Eve, with perfect access to the channel from Alice to Bob and also side information E^n . Unlike in the last section, in this model the achievable (R_A, R_C, Δ) has not been fully characterized in the case of $X = Y$. However, Gündüz et al. [3] and Probhakaran and Ramchandran [2] showed that if $R_C > H(Z)$, that is uncoded side information is available at Bob, then (R_A, Δ) is an achievable pair if and only if $R_A \geq H(Y|Z)$ and $\Delta \leq \max[I(Y; Z|U) - I(Y; E|U)]$ where the maximization is taken over U that satisfies $Z \text{---} Y \text{---} U$, thus providing a full single-letter characterization of the achievable rate-equivocation region. In this section, we assume coded side information is available at Bob and Eve has uncoded side information E^n . As in

$$\begin{aligned}
0 &\stackrel{(a)}{=} \sum_{i=1}^n I(Y_i, E_i; X_{i+1}^n, E_{i+1}^n | J, Y^{i-1}, E^{i-1}) - I(Y^{i-1}, E^{i-1}; X_i, E_i | J, X_{i+1}^n, E_{i+1}^n) \\
&= H(Y^n, E^n | J) - H(X^n, E^n | J) - \sum_{i=1}^n [H(Y_i, E_i | X_{i+1}^n, Y^{i-1}, E^{i-1}, J) - H(X_i, E_i | X_{i+1}^n, Y^{i-1}, E^{i-1}, J)] \\
&= H(Y^n | E^n, J) - H(X^n | E^n, J) - \sum_{i=1}^n [H(Y_i | E_i, X_{i+1}^n, Y^{i-1}, E^{i-1}, J) - H(X_i | E_i, X_{i+1}^n, Y^{i-1}, E^{i-1}, J)] \\
&\stackrel{(b)}{=} H(Y^n | E^n, J) - H(X^n | E^n, J) - \sum_{i=1}^n [H(Y_i | E_i, U_i) - H(X_i | E_i, U_i)] \tag{5}
\end{aligned}$$

[6], we assume that the Eve's side information E^n forms the Markov chain $X^n \text{---} Y^n \text{---} E^n$.

A. A Converse Result

We consider the model depicted in Fig. 2 in which Eve has access to side information E^n which satisfies $E^n \rightarrow Y^n \rightarrow X^n$.

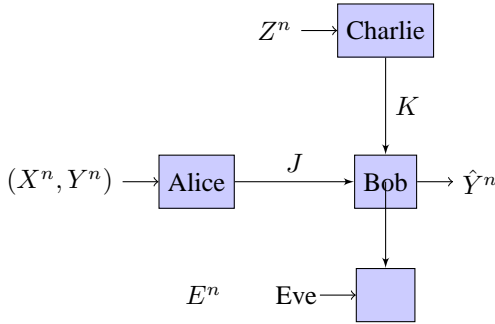


Fig. 2. Yamamoto's lossless source coding with eavesdropper having side information.

The achievable (R_A, R_C, Δ) in this model is defined similarly as before with the utility constraint (1) and the privacy constraint

$$\frac{1}{n} H(X^n | E^n, J) \geq \Delta - \varepsilon. \tag{6}$$

Before we get to an outer bound for the achievable region of this model, we need to state the following lemma which is a generalization of identity (4) that we used in the proof of Theorem 1.

Lemma 1. *Let (J, X^n, Y^n, E^n) be jointly distributed according to $P(j, x^n, y^n, e^n)$. Then we can write:*

$$\begin{aligned}
&H(X^n | E^n, J) - H(Y^n | E^n, J) \\
&= \sum_{i=1}^n [H(X_i | E_i, U_i) - H(Y_i | E_i, U_i)]
\end{aligned}$$

where $U_i := (X_{i+1}^n, Y^{i-1}, E^{i-1}, J)$ for each $1 \leq i \leq n$ and $E^{-i} := (E_{i-1}, E_{i+1}^n)$.

Proof. The proof is presented in (5), where (a) follows from Ciszár sum identity [10, page 25], in (b) we used the definition of U_i . ■

Theorem 2. *The set of all achievable triples (R_A, R_C, Δ) for this model when Eve is provided with side information E^n and $E^n \text{---} Y^n \text{---} X^n$, satisfies*

$$\begin{aligned}
R_A &\geq H(Y|V), \\
R_C &\geq I(Z; V), \\
\Delta &\leq I(X, Y; V) - I(X, Y; E|U) \\
&\quad + H(X|E, U) - H(Y|E, U),
\end{aligned}$$

for some U and V which form $(Z, E) \text{---} (X, Y) \text{---} U$ and $(X, Y, E) \text{---} Z \text{---} V$.

Proof. The lower bounds for both R_A and R_C follow along the same lines as in the proof of Theorem 1. We shall show the upper bound for the equivocation. We note that since Bob is required to reconstruct Y^n losslessly, Fano's inequality implies that

$$H(Y^n | J, K) \leq n\varepsilon_n \tag{7}$$

for $\varepsilon_n \rightarrow 0$ as $n \rightarrow \infty$. As before, let $J = f_A(X^n, Y^n)$ and $K = f_C(Z^n)$.

The upper bound for the equivocation is obtained in (8) shown on top of the next page where (a) follows from Lemma 1 and (b) is due to (7). Since $K \text{---} (X^n, Y^n) \text{---} J$ and $E^n \text{---} Y^n \text{---} J$, we have $I(X^n, Y^n; K|J) \leq I(X^n, Y^n; K)$ and $I(Y^n; E^n|J) = I(Y^n; E^n) - I(E^n; J)$ and hence (c) follows. We again used the Markov chain relation $E^n \text{---} Y^n \text{---} X^n$ in (d). The definition $V_i := (K, X^{i-1}, Y^{i-1})$ and the fact that $I(E_i; J, E^{i-1}) \leq I(E_i; U_i)$ are used in (e). Note that since $U_i \text{---} (X_i, Y_i) \text{---} E_i$ we have in (f) that $I(X_i, Y_i; E_i|U_i) = I(X_i, Y_i; E_i) - I(E_i; U_i)$. The proof completes by introduction of a time sharing random variable Q uniformly distributed over $\{1, 2, \dots, n\}$ and independent of (X^n, Y^n, Z^n, E^n) and letting $X = X_Q, Y = Y_Q, E = E_Q, V = (V_Q, Q)$ and $U = (U_Q, Q)$. ■

Remark 2. Setting $E^n = \emptyset$ and thus removing the eavesdropper's side information, Theorem 2 yields $\Delta \leq I(X, Y; V) + H(X|U) - H(Y|U)$ and hence Theorem 2 subsumes Theorem 1.

In the simple case of $X = Y$, the optimal scheme when coded side information is available at Bob and $E^n = \emptyset$ is proposed in [6] which is shown to resemble the binning

$$\begin{aligned}
H(X^n|E^n, J) &\stackrel{(a)}{=} H(Y^n|E^n, J) + \sum_{i=1}^n [H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&= H(Y^n|J, K) + I(Y^n; K|J) - I(Y^n; E^n|J) + \sum_{i=1}^n [H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\stackrel{(b)}{\leq} n\varepsilon_n + I(X^n, Y^n; K|J) - I(Y^n; E^n|J) + \sum_{i=1}^n [H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\stackrel{(c)}{\leq} n\varepsilon_n + I(X^n, Y^n; K) - I(Y^n; E^n) + I(E^n; J) + \sum_{i=1}^n [H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\stackrel{(d)}{=} n\varepsilon_n + \sum_{i=1}^n [I(X_i, Y_i; K, X^{i-1}, Y^{i-1}) - I(Y_i, X_i; E_i) + I(E_i; J, E^{i-1}) \\
&\quad + H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\stackrel{(e)}{\leq} n\varepsilon_n + \sum_{i=1}^n [I(X_i, Y_i; V_i) - I(Y_i, X_i; E_i) + I(E_i; U_i) + H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\stackrel{(f)}{=} n\varepsilon_n + \sum_{i=1}^n [I(X_i, Y_i; V_i) - I(Y_i, X_i; E_i|U_i) + H(X_i|E_i, U_i) - H(Y_i|E_i, U_i)] \\
&\stackrel{(g)}{=} n\varepsilon_n + I(X_Q, Y_Q; V_Q, Q) - I(Y_Q, X_Q; E_Q|U_Q, Q) + H(X_Q|E_Q, U_Q, Q) - H(Y_Q|E_Q, U_Q, Q) \quad (8)
\end{aligned}$$

scheme of Wyner in [7]. Although, a tight bound for the equivocation when E^n is available is not yet known, Theorem 2, specialized to $X = Y$, implies

$$\Delta \leq I(Y; V) - I(Y; E|U),$$

for auxiliary random variables U and V which form Markov chains $V \text{---} Z \text{---} (Y, E)$ and $U \text{---} Y \text{---} (Z, E)$.

B. A Coding Scheme When Bob Has Uncoded Side Information

As a special case, we consider the case where Alice does not see the private source and also $R_C > H(Z)$ (i.e., Bob has uncoded side information). In this case, Theorem 2 implies that the best achievable equivocation is upper bounded by

$$\max[I(Y; Z) - I(Y; E|U) + H(X|E, U) - H(Y|E, U)],$$

where the maximization is taken over U which forms the Markov chain relation $U \text{---} Y \text{---} (Z, E, X)$. In the following we give a simple coding scheme which incurs a smaller equivocation and is thus suboptimal. In fact, if the above maximization results in a U which is independent of Z , then the following coding scheme is optimal. On the other hand, if the maximization results in a U which is constant, then it implies that Slepian-Wolf binning is optimal, because if Alice uses Slepian-Wolf binning then the equivocation is equal to $H(X|E) - H(Y|Z)$, as observed in [2].

Theorem 3. *When X^n is not given to Alice and Bob observes side information Z^n , then (R_A, Δ) which satisfies*

$$\begin{aligned}
R_A &\geq H(Y|Z), \\
\Delta &\leq I(Y; Z|U) - I(Y; E|U)
\end{aligned}$$

$$+ H(X|E, U) - H(Y|E, U),$$

is achievable where the auxiliary random variable U forms the Markov chain $(X, Z, E) \text{---} Y \text{---} U$.

Proof. Our scheme is similar to the ones proposed in [3] and [11]. Given Y^n , we generate $2^{n(I(Y;U)+\varepsilon)}$ independent codewords of length n , $U^n(w)$, $w \in \{1, 2, \dots, 2^{n(I(Y;U)+\varepsilon)}\}$ according to $\prod_{i=1}^n P(u_i)$. We then uniformly bin all the U^n sequences into $2^{n(I(Y;U)-I(U;Z))}$ bins. Let $B(i)$ be the indices assigned to bin i . There are approximately $2^{nI(U;Z)}$ indices in each bin. We also uniformly bin Y^n sequences into $2^{n(H(Y|U,Z)+\varepsilon)}$ bins and let $C(k)$ be the set of sequences Y^n in bin k . Alice adopts a two-part encoding scheme. Given Y^n , Alice, in the first part, looks for a codeword $U^n(w)$ such that $(Y^n, U^n(w)) \in \mathcal{A}_{YU}^n$, where \mathcal{A}_{YU}^n denotes the set of all strongly typical $(y^n, u^n) \in \mathcal{Y}^n \times \mathcal{U}^n$ with respect to the distribution $P(y, u)$. She then reveals the bin index J_1 such that $w \in B(J_1)$. In the second part, she reveals J_2 such that $Y^n \in C(J_2)$.

Given J_1, J_2 and Z^n , Bob can find, with high probability, $U^n(w)$ such that $w \in B(J_1)$ and $(U^n(w), Z^n) \in \mathcal{A}_{ZU}^n$. It is then clear from the Slepian-Wolf theorem that Bob can recover Y^n with high probability given $U^n(w)$, Z^n , and J_2 .

The rate of this encoder is clearly equal to $H(Y|U, Z) + I(Y; U) - I(U; Z) = H(Y|Z)$.

The equivocation for this scheme can be found as

$$\begin{aligned}
&H(X^n|J_1, J_2, E^n) \\
&= H(X^n|J_1, E^n) - I(X^n; J_2|J_1, E^n) \\
&\geq H(X^n|U^n, E^n) - H(J_2)
\end{aligned}$$

$$\begin{aligned}
H(X^n|U^n, E^n) &= \sum_{(u^n, e^n) \in \mathcal{U}^n \times \mathcal{E}^n} P(u^n, e^n) H(X^n|U^n = u^n, E^n = e^n) \\
&\geq \sum_{(u^n, e^n) \in \mathcal{T}_{U, E}^n} P(u^n, e^n) H(X^n|U^n = u^n, E^n = e^n) \\
&= \sum_{(u^n, e^n) \in \mathcal{T}_{U, E}^n} P(u^n, e^n) \left[- \sum_{x^n \in \mathcal{X}^n} P(x^n|u^n, e^n) \log(P(x^n|u^n, e^n)) \right] \\
&\geq \sum_{(u^n, e^n) \in \mathcal{T}_{U, E}^n} P(u^n, e^n) \left[- \sum_{x^n \in \mathcal{T}_{X|u^n, e^n}^n} P(x^n|u^n, e^n) \log(P(x^n|u^n, e^n)) \right] \\
&\stackrel{(c)}{\geq} n(H(Y|U, E) - \delta_n) \sum_{(u^n, e^n) \in \mathcal{T}_{U, E}^n} P(u^n, e^n) \left[\sum_{x^n \in \mathcal{T}_{X|u^n, e^n}^n} P(x^n|u^n, e^n) \right] \\
&= n(H(Y|U, E) - \delta_n) \sum_{(u^n, e^n) \in \mathcal{T}_{U, E}^n} P(u^n, e^n) \left[\Pr\{(u^n, e^n, X^n) \in \mathcal{T}_{X|u^n, e^n}^n\} \right] \\
&\stackrel{(d)}{\geq} n(H(Y|U, E) - \delta_n)(1 - \delta'_n)
\end{aligned} \tag{9}$$

$$\begin{aligned}
&\stackrel{(a)}{\geq} H(X^n|U^n, E^n) - nH(Y|U, Z) \\
&\stackrel{(b)}{\geq} n[H(X|U, E) - H(Y|U, Z)] \\
&= n[H(X|E, U) - H(Y|E, U)] \\
&\quad + I(Y; Z|U) - I(Y; E|U),
\end{aligned}$$

where (a) follow from the fact that J_2 is a random variable over a set of size $2^{nH(Y|U, Z)}$ and (b) is proved in (9) where $\mathcal{T}_{U, E}^n$ denotes the set of typical sequences (u^n, e^n) and (c) is due to the property of typical sequences; in particular for typical x^n sequence with respect to $P(x^n|u^n, e^n)$ for $(u^n, e^n) \in \mathcal{T}_{U, E}^n$ we have $P(x^n|u^n, e^n) \leq 2^{-(n(H(X|U, E) - \delta(n)))}$ for $\delta_n \rightarrow 0$ as $n \rightarrow \infty$. We invoked Markov lemma [10, Lemma 12.1] in (d) to conclude that for the Markov chain relation $(X, E) \text{---} Y \text{---} U$ we have $(x^n, y^n, e^n, u^n) \in \mathcal{T}_{X, Y, E, U}^n$ and hence $\Pr\{(u^n, e^n, X^n) \in \mathcal{T}_{U, E, X}^n\} > 1 - \delta'_n$ for each pair $(u^n, e^n) \in (u^n, e^n) \in \mathcal{T}_{U, E}^n$ and $\delta'_n \rightarrow 0$ as $n \rightarrow \infty$. ■

IV. CONCLUDING REMARKS

Having combined the idea of compression of private and public sources of Yamamoto [1] with secure source coding problem (e.g. [3], [6] and [2]), we introduced a lossless source coding problem in which, given a two-dimensional source (X^n, Y^n) , the encoder must compress the source into an index J with rate R_A such that the receiver recovers Y^n losslessly and simultaneously reveals only little information about X^n . This model differs from typical information-theoretic secrecy models in that the utility and privacy constraints are defined for two different sources and thus provides a more general utility-equivocation tradeoff.

We gave converse results for compression rates and also the information leakage rate (or equivocation) which reduce to known results in the special case of $X = Y$. In particular,

with this simplifying assumption, Theorem 1 and Theorem 3 reduce to [6, Theorem 1] and [3, Corollary 3.2].

However, it is not clear at the moment that the bounds are tight in general. Constructing an achievability scheme for the most general case (i.e., the setting of Theorem 2) is the subject of our future studies.

REFERENCES

- [1] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver or wiretappers," *IEEE Trans. Inf. Theory*, vol. 29, no. 6, pp. 918–923, Nov. 1983.
- [2] V. Prabhakaran and K. Ramchandran, "On secure distributed source coding," in *IEEE Inf. Theory Workshop (ITW)*, Sept. 2007, pp. 442–447.
- [3] D. Gündüz, E. Erkip, and H. Poor, "Secure lossless compression with side information," in *Proc. IEEE Inf. Theory Workshop*, May 2008, pp. 169–173.
- [4] —, "Lossless compression with security constraints," in *IEEE Int. Sym. on Inf. Theory (ISIT)*, July 2008, pp. 111–115.
- [5] 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, June 2013.
- [6] R. Tandon, S. Ulukus, and K. Ramchandran, "Secure source coding with a helper," *IEEE Trans. Inf. Theory*, vol. 59, no. 4, pp. 2178–2187, April 2013.
- [7] A. Wyner, "On source coding with side information at the decoder," *IEEE Trans. Inf. Theory*, vol. 21, no. 3, pp. 294–300, May 1975.
- [8] E. Ekrem and S. Ulukus, "Secure lossy source coding with side information," in *Proc. Annual Allerton Conference on Communication, Control, and Computing*, Sept. 2011, pp. 1098–1105.
- [9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
- [10] Y. H. Kim and A. E. Gamal, *Network Information Theory*. Cambridge University press, 2012.
- [11] C. Schieler and P. Cuff, "Secrecy is cheap if the adversary must reconstruct," in *Proc. IEEE Int. Symp. on Inf. Theory (ISIT)*, July 2012, pp. 66–70.