Secure Distributed Source Coding with Side-Information

William Luh and Deepa Kundur

***Abstract*— This letter develops codes for the scenario in which users with correlated messages are to encipher and compress their messages without collaboration and without the use of cryptographic keys or other secret materials. We consider an eavesdropper that has access to an encoded message and in addition, some side-information in the form of uncoded symbols corresponding to the encoded message. Our codes are an extension of distributed source coding using syndromes (DISCUS) with the additional requirement of providing secrecy for the scenario described above. We state a secrecy condition that the subcodes of DISCUS must satisfy, and develop a general encoding algorithm meeting these conditions. We analyze the performance of the proposed code for the case of multiple eavesdropped messages.**

***Index Terms*— Security, distributed source coding, Reed-Solomon codes.**

I. Introduction and Background

THIS letter extends and generalizes the problem in which two nodes with correlated messages wish to encipher and source code their messages without collaboration and without the use of cryptographic keys. An eavesdropper with access to only one encoded message learns as little as possible about that message, whereas a *joint* decoder with all encoded messages can decode the messages without error.[1] In [2] we derived the capacity region, i.e. the set of all possible source coding rate pairs and equivocation rate (our measure of secrecy) pairs. The capacity region showed that simply applying Slepian-Wolf (SW) encoding simultaneously achieves the optimal equivocation rates.

Although SW encoding alone suffices to achieve the optimal equivocation rates (a result of the capacity region), and therefore any distributed source code may be applied, in practice the resulting secrecy is mediocre, giving the eavesdropper too much information concerning the message without the eavesdropper having to do any work. Thus in [2] we defined a different measure of secrecy that not only resolves the shortcomings of simply using equivocation as the measure of secrecy, but also accounts for the possible scenario in which the eavesdropper has access to uncoded symbols from the message in addition to intercepting the corresponding encoded message (similar threat model as [3]).

In [2] we gave a simple two-user example to demonstrate that the latter strong definition of secrecy is plausible. In this letter we derive general codes for any number of users, as well as analyze the case when the eavesdropper has access to multiple encoded messages. While the exploration of the capacity of various wiretap channel models has recently received much attention, practical coding for these models are emerging [3]–[7].

II. System Model

We define the problem for m users who have random messages (column vectors of elements from Galois field $GF(q)$) $U_1^k, \ldots, U_m^k \in (GF(q))^k$ such that they are marginally uniformly distributed. On the other hand, the collection of *all* m messages obeys the correlation model

$$w(U_1^k + \cdots + U_m^k) \leq t, \tag{1}$$

where $w(\cdot)$ is the Hamming weight, and addition is over $GF(q)$ [8]. The m users are to separately and linearly encode (jointly encipher and source code) their realizations $u_1^k, \ldots, u_m^k$, resulting in Galois vectors $x_1^{n_1}, \ldots, x_m^{n_m}$, respectively, via the relationship

$$x_i^{n_i} \quad = \quad \mathbf{H}_i u_i^k \tag{2}$$

for $i = 1, \ldots, m$ such that $\mathbf{H}_i$s are $n_i \times k$ matrices.[2] Given all $x_i^{n_i}$, $i = 1, \ldots, m$, the *joint* decoder is to reproduce all u_i^k, $i = 1, \ldots, m$ without error.

The eavesdropper is permitted to have at his disposal only one $x_i^{n_i}$; we later generalize this to include the case when the eavesdropper has multiple $x_i^{n_i}$. In addition the eavesdropper is also permitted to have α_i uncoded symbols from u_i^k if he intercepts $x_i^{n_i}$. These extra uncoded symbols (similar threat assumptions as in [3]) are the eavesdropper's side-information. The goal is to design an encoding and decoding scheme for the above system model, such that the eavesdropper cannot *uniquely solve* for any other symbols in u_i^k, nor any $u_j^k, j \neq i$ given that he has $x_i^{n_i}$ and the corresponding side-information. Formally, let $u_{i,j}$ be the j^{th} symbol of the i^{th} user's message. Then if the eavesdropper wishes to reveal $u_{i,j}$ (assuming he does not have $u_{i,j}$ as side-information already), the eavesdropper is faced with choosing $u_{i,j} \in \mathcal{S}$, such that $S \subseteq GF(q)$ and its cardinality is at least 2 (i.e. $|\mathcal{S}| \geq 2$), and all elements in $\mathcal{S}$ are equally likely from the eavesdropper's point of view. This level of secrecy is not unconditional, but may be satisfactory for certain applications, e.g. lightweight video encryption.

Manuscript received January 15, 2008. The associate editor coordinating the review of this letter and approving it for publication was V. Stanković.

The authors are with the Department of Electrical and Computer Engineering, Texas A&M University, 214 Zachry Engineering Center, College Station, Texas 77843 (e-mail: {luh, deepa}@ece.tamu.edu).

Digital Object Identifier 10.1109/LCOMM.2008.080070.

[1]This problem has for example applications in sensor networks in which energy-limited nodes must encode separately without collaboration [1]. Furthermore, node deployment in hostile envrionments makes cryptographic keys prone to capture or exposure and we thus avoid their use in our problem.

[2]Consideration of nonlinear codes are beyond the scope of this letter. In practice nonlinear codes may offer better protection against a wide assortment of cryptanalysis attacks. However the security of nonlinear codes are in general difficult to prove mathematically, while linear codes are amenable to analysis.

III. Results

Our codes belong to the class of distributed source coding using syndromes (DISCUS) [9]. In the DISCUS scheme, a supercode with the capability of correcting t-errors (the same t as in Eq. 1) is partitioned into m subcodes (where m is the number of users). The parity check matrices of these subcodes are then used in Eq. 2 to encode each of the user's messages, respectively.

A. Secrecy Condition on Subcodes

While the supercode requirement of being t-error-correctable aids decodability [9], in [2] we showed that secrecy (as defined above) can be achieved by imposing that the subcodes are maximum distance separable (MDS) codes; this is re-iterated explicitly in Theorem 1 along with restrictions on the quantity of side-information available to the eavesdropper.

Theorem 1: If $\alpha_i < k - n_i$ and the subcodes are each maximum distance separable, then the eavesdropper cannot solve for any symbols other than those given as side-information in u_i^k for each $i = 1, \ldots, m$.

B. Code Construction

We have conditions on the supercode (error correction capability equal to correlation [9]) and the subcodes (MDS from Theorem 1) for DISCUS with secrecy. However, simply choosing MDS subcodes will often result in unacceptable supercodes. Similarly, choosing an acceptable supercode and arbitrarily partitioning the supercode into subcodes will often result in non-MDS codes, e.g. in [2] we showed this is the case for the DISCUS codes in [8]. This section derives codes that satisfy both conditions.

Algorithm 1 provides a method of constructing DISCUS codes that are *both* decodable (zero errors), and secure in the sense developed in Section II.[3] An example for two users is given in [2].

Theorem 2: If $\mathbf{H}_i$, $i = 1, \ldots, m$ are selected using Algorithm 1 and the eavesdropper has side-information restricted to $\alpha_i < a_i$ (a_i from Algorithm 1), then the encoding scheme is secure and uniquely decodable.

Proof: Matrix $\mathbf{A}$ is the generator matrix of a Reed-Solomon code with minimum distance $d_{min} = k - (2s) + 1$. Defining $t \triangleq \left\lfloor \frac{k}{2} - s \right\rfloor$ satisfies the requirement $k \geq 2(s+t)$, and unique decodability is possible given the correlation model, Eq. 1, and given that the Slepian-Wolf (SW) constraints are satisfied.

Since $\mathbf{H}_i$ is a $(k - a_i) \times k$ matrix from Algorithm 1, the side-information constraint $\alpha_i < k - (k - a_i) = a_i$ follows from Theorem 1. Furthermore, since each symbol in a message is independent over $GF(k+1)$, the source coding rate given by Eq. 6 and constrained by Eq. 7 is simply the Slepian-Wolf theorem. Therefore we have proved that zero-error decodability is achieved given the above constraints are satisfied.

Next we must check that the $\mathbf{H}_i$s are parity check matrices of MDS codes for all $i = 1, \ldots, m$ (to satisfy Theorem 1). In general, partitioning a MDS code may not result in MDS subcodes. However, the choice of starting with $\mathbf{A}$ in Algorithm 1 facilitates the generation of MDS subcodes as we show. The parity check polynomial corresponding to the generator matrix $\mathbf{A}_i$ is equal to

$$h_i(x) = \prod_{l=a_i^-}^{a_i^+} (x - \xi^{-l}) = \prod_{l=a_i^-}^{a_i^+} (x - \xi^{k-l}) = \prod_{l=k-a_i^+}^{k-a_i^-} (x - \xi^l) \tag{3}$$

where

$$a_i^- = \left(\sum_{j=1}^{i-1} a_j \right) + 1, \qquad a_i^+ = \sum_{j=1}^{i} a_j. \tag{4}$$

Therefore the generator polynomial corresponding to generator matrix $\mathbf{A}_i$ is given by

$$\begin{aligned} g_i(x) &= \frac{x^k - 1}{h_i(x)} = \prod_{l=1}^{k-a_i^+-1} (x - \xi^l) \prod_{l=k-a_i^-+1}^{k} (x - \xi^l) \\ &= \prod_{l=k-a_i^-+1}^{2k-a_i^+-1} (x - \xi^l) \end{aligned} \tag{5}$$

where the final equality follows since $\xi^{k+l} = \xi^l$. Since the generator polynomial has roots that are consecutive powers of the primitive element ξ, the code it generates is by definition Reed-Solomon, which is MDS. ■

Algorithm 1 Finding Secure Parity Check Matrices

Require: $\mathbf{H}_i$ for all $i = 1, \ldots, m$.

Ensure:

(i) Symbols for all messages are from $GF(k+1)$ where $k+1$ is a power of a prime number and $k \geq 2(s+t)$;

(ii) Eq. 1 is satisfied;

(iii) ξ is a primitive element in $GF(k+1)$ and

$$\mathbf{A} = \begin{pmatrix} 1 & \xi & \xi^2 & \ldots & \xi^{(k-1)} \\ 1 & \xi^2 & \xi^4 & \ldots & \xi^{2(k-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \xi^{2s} & \xi^{4s} & \ldots & \xi^{2s(k-1)} \end{pmatrix}$$

1: Partition the matrix $\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 \\ \hline \vdots \\ \hline \mathbf{A}_m \end{pmatrix}$ such that the $\mathbf{A}_i$s are $a_i \times k$ matrices, and

$$R_i \triangleq \frac{k - a_i}{k} \log_2(k+1) \tag{6}$$

for all $i = 1, \ldots, m$ satisfy

$$\sum_{i \in \mathcal{S}} R_i \geq H\left((U_j)_{j \in \mathcal{S}} \,|\, (U_j)_{j \in \{1,\ldots,m\} - \mathcal{S}} \right) \tag{7}$$

for all $\mathcal{S} \subseteq \{1, \ldots, m\}$. If this is not possible, then these rates cannot be used.

2: Select $\mathbf{H}_i$s such that $\mathbf{A}_i \mathbf{H}_i^{\mathrm{T}} = \mathbf{0}$ for all $i = 1, \ldots, m$.

C. Example

We illustrate a numerical example for two users. In Algorithm 1 let $s = 2, t = 5, k = 15$, so $k \geq 2(s+t)$ is

[3] As in the original DISCUS, the drawback is that these codes do not always exist for all parameters.

satisfied. Matrix $\mathbf{A}$ is then a 4×15 matrix. If we partition matrix $\mathbf{A}$ equally so that $a_1 = a_2 = 2$, then both users have the same rate of $R_1 = R_2 = \frac{k-a_i}{k}\log_2(k+1) = \frac{52}{15}$. Then note that $H(U_1, U_2) = (1 + t/k)\log_2(k+1) = \frac{16}{3}$, and $H(U_1|U_2) = H(U_2|U_1) = \frac{t}{k}\log_2(k+1) = \frac{4}{3}$, thus the SW constraints (Eq. 7) are satisfied as the reader may verify. To complete the example, the reader may derive parity check matrices $\mathbf{H}_1$ and $\mathbf{H}_2$, corresponding to generator matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, respectively. Finally Eq. 2 can be applied to get the encoded messages.

D. Multiple Eavesdropping

Algorithm 1 only allows the eavesdropper to have access to *one* encoded message with some corresponding side-information. We now analyze the above code for the case when the eavesdropper has access to multiple encoded messages.

Proposition 1: If the eavesdropper has access to μ encoded messages $\{x_{j_1}^{n_{j_1}}, \ldots, x_{j_\mu}^{n_{j_\mu}}\}$ for $1 < \mu < m$, and any amount of side-information, then the scheme of Algorithm 1 is *not* secure. If the eavesdropper has *no* side-information, then the scheme of Algorithm 1 is secure.[4]

Proof: Given $\{x_{j_1}^{n_{j_1}}, \ldots, x_{j_\mu}^{n_{j_\mu}}\}$, we can write

$$\begin{pmatrix} x_{j_1}^{n_{j_1}} \\ \hline \vdots \\ \hline x_{j_\mu}^{n_{j_\mu}} \end{pmatrix} = \left(\begin{array}{c|c|c} \mathbf{H}_{j_1} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \ddots & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} & \mathbf{H}_{j_\mu} \end{array}\right) \begin{pmatrix} u_{j_1}^{k} \\ \hline \vdots \\ \hline u_{j_\mu}^{k} \end{pmatrix}. \quad (8)$$

Eq. 8 will be conveniently denoted by

$$x = \mathbf{H}u. \quad (9)$$

It can easily be seen that $\mathbf{H}$ is no longer the parity check matrix of a MDS code, since its minimum distance is strictly less than the Singleton bound. This is proved by considering $\mathbf{H}$ as a generator matrix; thus the codewords generated by $\mathbf{H}$ are precisely the concatenation of codewords generated by $\mathbf{H}_{j_1}, \ldots, \mathbf{H}_{j_\mu}$. Therefore the codewords generated by $\mathbf{H}$ must have minimum distance

$$d_{min} = \min\{d_{min,j_1}, \ldots, d_{min,j_\mu}\} \quad (10)$$

where d_{min,j_i} is the minimum distance of the code corresponding to generator matrix $\mathbf{H}_{j_i}$. Therefore the code generated by $\mathbf{H}$ is not MDS, and neither is its dual, thus it is not secure when side-information is available.

However, when the eavesdropper does not have side-information, we show that $\mathbf{H}$ is secure. Now *if* there exists a $1 \times (n_{j_1} + \cdots + n_{j_\mu})$ row vector b_i such that

$$b_i x = b_i \mathbf{H} u = I_i^{\mu k} u \quad (11)$$

where now $I_i^{\mu k}$ is a $1 \times \mu k$ row vector with 1 in position i and 0 elsewhere, then the eavesdropper can solve for the i^{th} symbol in u. Eq. 11 is equivalent to $I_i^{\mu k} \in \text{rowspace}(\mathbf{H})$. Thus if we show that $I_i^{\mu k} \notin \text{rowspace}(\mathbf{H})$ for all $i = 1, \ldots, \mu k$, then the eavesdropper cannot solve for any symbols in u. First it is easy to see that the rows of $\mathbf{H}$ are independent, since the rows of each $\mathbf{H}_{j_1}, \ldots, \mathbf{H}_{j_\mu}$ are independent

[4]All the codes in this letter are also secure against the more general scenario in which the eavesdropper has access to α_i linear combinations of symbols of the uncoded message u_i^k.

Now suppose for some arbitrary $I_i^{\mu k}$, column i runs through $\mathbf{H}_r$ in $\mathbf{H}$. Therefore $I_i^{\mu k}$ is not in the rowspace consisting of all rows in $\mathbf{H}$ with matrices $\mathbf{H}_l$, $l \neq r$, since the i^{th} element in each of these $\mathbf{H}_l$ is 0, while it is 1 in $I_i^{\mu k}$.

Thus we only have to check that $I_i^{\mu k}$ is not in the rowspace consisting of the rows in $\mathbf{H}$ with matrix $\mathbf{H}_r$, i.e.

$$I_i^{\mu k} \notin \text{rowspace}\left(\begin{array}{c|c|c|c|c|c|c} \mathbf{0} & \cdots & \mathbf{0} & \mathbf{H}_r & \mathbf{0} & \cdots & \mathbf{0} \end{array}\right). \quad (12)$$

Truncating $\mathbf{H}$ to yield $\mathbf{H}_r$, and similarly truncating $I_i^{\mu k}$ to yield the corresponding $1 \times k$ row vector I_l, $l \in \{1, \ldots, k\}$ proves Eq. 12 since $I_l \notin \text{rowspace}(\mathbf{H}_r)$ from the proof of Algorithm 1.

There is one final technical point that we must prove. We have been assuming that all u realizations in Eq. 9 are equally likely. This is true so long as $\mu < m$ (where m is the total number of users). If $\mu = m$, i.e. if the eavesdropper has access to *all* encoded messages, then by the correlation model (see Eq. 1), some u realizations (in Eq. 9) are impossible. On the other hand if $\mu = m - 1$, then there is no constraint on $\{u_{j_1}^k, \ldots, u_{j_\mu}^k\}$; only the m^{th} $u_{j_{\mu+1}}^k$ would have to be chosen so that all m messages satisfy Eq. 1. Thus as long as the eavesdropper has fewer than m encoded messages, all u's (in Eq. 9) are equally likely to the eavesdropper. ■

IV. Conclusion

In this letter we generalized DISCUS with secrecy to multiple users. We showed our general code is secure (viz. Section II) when the eavesdropper has one encoded message with corresponding side-information, and when the eavesdropper has multiple encoded messages without side-information.

References

[1] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Mag.*, vol. 21, pp. 80–94, Sept. 2004.

[2] W. Luh and D. Kundur, "Separate enciphering of correlated messages for confidentiality in distributed networks," in *Proc. IEEE Globecom*, Washington, DC, Nov. 2007.

[3] Y. Luo, C. Mitrpant, A. J. Han Vinck, and K. Chen, "Some new characters on the wire-tap channel of type II," *IEEE Trans. Inform. Theory*, vol. 51, no. 3, pp. 1222–1229, Mar. 2005.

[4] M. Bloch, J. Barros, M. R. D. Rodrigues, and S. W. McLaughlin, "Wireless information-theoretic security," *IEEE Trans. Inform. Theory*, 2008, accepted.

[5] R. Liu, Y. Liang, H. V. Poor, and P. Spasojevic, "Secure nested codes for type II wiretap channels," in *Proc. IEEE Information Theory Workshop on Frontiers in Coding Theory*, Lake Tahoe, CA, Sept. 2007.

[6] A. Thangaraj, S. Dihidar, A. R. Calderbank, S. W. McLaughlin, and J.-M. Merolla, "Applications of LDPC codes to the wiretap channel," *IEEE Trans. Inform. Theory*, vol. 53, no. 8, pp. 2933–2945, Aug. 2007.

[7] C. Ye and P. Narayan, "Secret key and private key constructions for simple multiterminal source models," in *Proc. IEEE International Symposium on Information Theory*, Sept. 2005.

[8] V. Stanković, A. Liveris, Z. Xiong, and C. Georghiades, "Design of Slepian-Wolf codes by channel code partitioning," in *Proc. Data Compression Conference*, Snowbird, UT, Mar. 2004.

[9] S. Pradhan and K. Ramchandran, "Distributed source coding: Symmetric rates and applications to sensor networks," in *Proc. DCC'00*, Snowbird, UT, Mar. 2000.