A Probabilistic Error-Correcting Scheme that Provides Partial Secrecy

Scott Decatur Oded Goldreich Dana Ron

August 25, 2019

Abstract

In the course of research in Computational Learning Theory, we found ourselves in need of an error-correcting encoding scheme for which few bits in the codeword yield no information about the plain message. Being unaware of a previous solution, we came-up with the scheme presented here.

Clearly, a scheme as described above cannot be deterministic. Thus, we introduce a probabilistic coding scheme which, in addition to the standard coding theoretic requirements, has the feature that any constant fraction of the bits in the (randomized) codeword yields no information about the message being encoded. This coding scheme is also used to obtain efficient constructions for the *Wire-Tap Channel* Problem.

Appeared (under the title "A Probabilistic Error-Correcting Scheme") as record 1997/005 of the *IACR Cryptology ePrint Archive*, 1997. In the current revision, the introduction was intentionally left intact, but the exposition of the main result (esp., its proof) was elaborated and made more reader-friendly.

1 Original Introduction (dated April 1997)

We believe that the following problem may be relevant to research in Cryptography:

> Provide an error-correcting encoding scheme for which few bits in the codeword yield no information about the plain message.

Certainly, no deterministic encoding may satisfy this requirement, and so we are bound to seek probabilistic error-correcting encoding schemes. Specifically, in addition to the standard coding theoretic requirements (i.e., of correcting upto a certain threshold number of errors), we require that obtaining less than a threshold number of bits in the (randomized) codeword yield no information about the message being encoded.

Below we present such a probabilistic encoding scheme. In particular, the scheme can (always) correct a certain constant fraction of errors, and has the property that fewer than a certain constant fraction of bits (in the codeword) yield no information about the encoded message. Thus, using this encoding scheme over an insecure channel tampered by an adversary who can read and modify (only) a constant fraction of the transmitted bits, we establish correct and private communication between the legitimate end-points.

The new coding scheme is also used to obtain *efficient constructions* for the *Wire-Tap Channel* Problem (cf., [9]). Related work has been pointed out to us recently by Claude Crépeau. These

include [4, 7, 1, 3]. In particular, the seemingly stronger version of the problem, considered in this work, was introduced by Csiszár and Körner [4]. Maurer has shown that this version of the problem can be reduced to the original one by using bi-directional communiaction [7]. Crépeau (private comm., April 1997) has informed us that, using the techniques in [1, 3], one may obtain an alternative efficient solution to the Wire-Tap Channel Problem again by using bi-directional communiaction.[1]

Our own motivation to study the problem had to do with Computational Learning Theory. Indeed, the solution was introduced and used in our work on *computational sample complexity* [5].

2 Main Result

We focus on good error correcting codes (and encoding schemes), which are codes of constant rate and constant relative distance. Recall that a standard (binary) (error-correcting) code of rate $\rho > 0$ and relative distance $\delta > 0$ is a mapping $C : \{0,1\}^* \to \{0,1\}^*$ that satisfies $|C(x)| = |x|/\rho$ and $\min_{x \neq y : |x| = |y|}\{\mathrm{wt(C(x) \oplus C(y))}\} \geq \delta \cdot \mathrm{|C(x)|}$, where $\mathrm{wt(z)} \stackrel{\mathrm{def}}{=} \mathrm{|\{i \in [|z|] : z_i = 1\}|}$ is the Hamming weight of z and $\alpha \oplus \beta$ denotes the bit-by-bit exclusive-or of the strings α and β.

We are interested in good codes that have efficient encoding and decoding algorithms, where the latter are applicable to error rates below $\delta/2$. That is, for some constant $\eta \in (0, \delta/2]$, we may hope to have a decoder such that for every x and $e \in \{0,1\}^{|C(x)|}$ of Hamming weight smaller than $\eta \cdot |C(x)|$, given a corrupted codewords $G(x) \oplus e$, recovers the original message x.

The non-standard (for coding theory) aspect that we consider here is *partial secrecy*. Specifically, for some constant $\varepsilon > 0$, any ε fraction of the bits of the codeword should yield no information about the original message. Obviously, this is not possible with an actual code, and so we settle for probabilistic encoding schemes as implicitly defined next.

Theorem 1 (a probabilistic error correction scheme with partial privacy): *There exist constants $\rho, \eta, \varepsilon > 0$ and a pair of probabilistic polynomial-time algorithms, denoted (E, D), such that*

1. Constant Rate*: $|E(x)| = |x|/\rho$, for all $x \in \{0,1\}^*$.*

2. Error Correction*: for every $x \in \{0,1\}^*$ and every $e \in \{0,1\}^{|E(x)|}$ such that* $\mathrm{wt(e)} \leq \eta \cdot \mathrm{|E(x)|}$, *it holds that*

$$\mathrm{Prob}(D(E(x) \oplus e) = x) \;=\; 1.$$

   Furthermore, Algorithm D is deterministic.

3. Partial Secrecy*: A substring containing $\varepsilon \cdot |E(x)|$ bits of $E(x)$ yields no information on x. That is, for $I \subseteq [|\alpha|] = \{1, ..., |\alpha|\}$, let α_I denote the substring of α corresponding to the bits at locations in I* (i.e., for $I = \{i_1, i_2, ..., i_t\}$ such that $i_j < i_{j+1}$, it holds that $\alpha_I = \alpha_{i_1}\alpha_{i_2} \cdots \alpha_{i_t}$). *Then, for every $n \in \mathcal{N}$, $x, y \in \{0,1\}^n$, and $\varepsilon \cdot (n/\rho)$-subset $I \subseteq [n/\rho]$, it holds that $E(x)_I$ is distributed identically to $E(y)_I$; that is, for every $\alpha \in \{0,1\}^{|I|}$,*

$$\mathrm{Prob}[E(x)_I = \alpha] \;=\; \mathrm{Prob}[E(y)_I = \alpha].$$

   Furthermore, $E(x)_I$ is uniformly distributed over $\{0,1\}^{|I|}$.

[1] Added in revision: Note that, in contrast, our solution uses uni-directional communication. On the other hand, our solution holds only for a limited range of parameters; see discussion at the end of Section 3.

In addition, on input x, algorithm E uses $O(|x|)$ coin tosses.

Items 1 and 2 are standard requirements of coding theory, first met by Justesen [6]. What is non-standard in Theorem 1 is Item 3. Indeed, Item 3 is impossible if one insists that the encoding algorithm (i.e., E) is deterministic.

Proof: The key idea is to encode the information by first augmenting it with a sufficiently long random padding, and then encoding the result using a good error correcting code (i.e., one of constant rate and constant relative distance).

To demonstrate this idea, consider an $2n$-by-m matrix M defining a good (linear) error-correction code. That is, the string $z \in \{0,1\}^{2n}$ is encoded by $z \cdot M$. Further suppose that the submatrix defined by the last n rows of M and any $\varepsilon \cdot m$ of its columns is of full-rank (i.e., rank $\varepsilon \cdot m$). Then, we define the following probabilistic encoding, E, of strings of length n. To encode $x \in \{0,1\}^n$, we first select $y \in \{0,1\}^n$ uniformly at random, let $z = xy$ and output $E(x) = z \cdot M$.

Clearly, the error-correction features of M are inherited by E. To see that the secrecy requirement holds consider any sequence of $\varepsilon \cdot m$ bits in $E(x)$. The contents of these bit locations is the product of z by the corresponding columns in M; that is, $z \cdot M' = x \cdot A + y \cdot B$, where M' denotes the submatrix corresponding to these columns in M, and A (resp., B) is the matrix resulting by taking the first (resp., last) n rows of M'. By hypothesis B is full rank, which implies that $y \cdot B$ is uniformly distributed. Hence, $z \cdot M'$ is uniformly distributed (regardless of x).

We stress that the foregoing argument relies on the hypothesis that *the submatrix defined by the last n rows of M and any $\varepsilon \cdot m$ of its columns is of full-rank*. Let us call such a matrix nice. So what is missing is a construction of a good linear code that is generated by a nice matrix and has an efficient decoding algorithm. Such a construction can be obtained by mimicking Justesen's construction [6]. Basically, we construct inner and outer encoding schemes, which correspond to the inner and outer codes used in [6], and apply composition (and analyze it) analogously. The encoding schemes that we use satisfy the error correction and secrecy requirements of the theorem, and we show that the composed scheme satisfies these requirements too, which establishes the theorem.

Justesen's Code. Recall that Justesen's Code is obtained by composing two codes: An *outer* linear code over a large alphabet is composed with an *inner* binary linear code that is used to encode single symbols of the large alphabet. The outer code is the Reed-Solomon Code; that is, the n-bit long message is encoded by viewing it as the coefficients of a polynomial of degree $t-1$ over a field with $\approx 3t$ elements, where $n \approx t \log_2(3t)$, and letting the codeword consists of the values of this polynomial at all field elements. Using the Berlekamp-Welch Algorithm [2], one can efficiently retrieve the information from a codeword provided that at most t of the symbols (i.e., the values of the polynomial at t field elements) were corrupted.

Our outer encoding. We obtain a variation of this outer-code as follows: Given $x \in \{0,1\}^n$, we pick a minimal $t \in \mathcal{N}$ such that $2n < t \log_2(3t)$, and view x as a sequence of $\frac{t}{2}$ elements in GF$(3t)$.[2] We uniformly select $y \in \{0,1\}^n$ and view it as another sequence of $\frac{t}{2}$ elements in GF$(3t)$. We consider the degree $t-1$ polynomial defined by these t elements, where x corresponds to the high-order coefficients and y to the low-order ones. Clearly, we preserve the error-correcting features of the original outer code. Furthermore, any $t/2$ symbols of the codeword yield no information about x. To see this, note that the values of these $t/2$ locations are obtained by multiplying a t-by-$t/2$

[2] Here we assume that $3t$ is a prime power. Actually, we use the first power of 2 that is greater than $3t$. Clearly, this inaccuracy has a negligible effect on the construction.

Vandermonde with the coefficients of the polynomial. We can rewrite the product as the sum of two products the first being the product of a $t/2$-by-$t/2$ Vandermonde with the low order coefficients. Thus, a uniform distribution on these coefficients (represented by y) yields a uniformly distributed result (regardless of x). (In other words, the generating matrix of the corresponding linear code is nice.) Hence, we have obtained a randomized outer-encoding that satisfies both the error-correction and secrecy requirements of the theorem, but this is encoding is over the alphabet $\mathrm{GF}(3t)$.

Our inner encoding. Next, we obtain an analogue of the inner code used in Justesen's construction. Here, the aim is to encode information of length $\ell \stackrel{\mathrm{def}}{=} \log_2(3t)$ (i.e., the representation of an element in $\mathrm{GF}(3t)$) using codewords of length $O(\ell)$. Hence, we do not need an efficient decoding algorithm, since Maximum Likelihood Decoding via exhaustive search is affordable (because $2^\ell = O(t) = O(n)$). Furthermore, any code that can be specified by $O(\log n)$ many bits will do (since we can try and check all possibilities in poly(n)-time), which means that we can use a randomized argument provided that it utilizes only $O(\log n)$ random bits. For example, we may use a linear code specified by a (random) 2ℓ-by-4ℓ Toeplitz matrix.[3] Using a probabilistic argument one can show that, with positive probability, a random 2ℓ-by-4ℓ Toeplitz matrix is as required in the motivating discussion (i.e., it generates a good code and is nice (i.e., its rows generate a code of distance $\Omega(\ell)$ and the submatrix induced by any $\Omega(\ell)$ columns and the last ℓ rows is of full rank)).[4] In the rest of the discussion, we fix such a nice Toeplitz matrix. We shall use it to randomly encode ℓ-bit strings (i.e., elements of $\mathrm{GF}(3t)$) by applying the matrix to a random 2ℓ-bit long padding of the ℓ-bit long input. Hence, we obtain a randomized inner-code that satisfies both the error-correction and secrecy requirements of the theorem.

The composition. We now get to the final step in mimicking Justesen's construction: the composition of the two codes. That is, we have outer and inner encoding schemes that satisfy both the error-correction and secrecy requirements of the theorem, and we need to show that their composition satisfies these features too. Let us first spell out what this composition is.

Recall that we want to encode $x \in \{0,1\}^n$, which is viewed as $x \in \mathrm{GF}(3t)^{t/2}$, where $n \approx (t/2)\log_2(3t)$. Applying the outer encoding scheme, with randomization $y \in \{0,1\}^n$, we obtain a $3t$-long sequence over $\mathrm{GF}(3t)$, denoted $x_1, ..., x_{3t}$. (Specifically, the Reed-Solomon code is applied to the $2n$-bit long string xy, viewed as a t-long sequence over $\mathrm{GF}(3t)$, resulting in the sequence $(x_1, ..., x_{3t}) \in \mathrm{GF}(3t)^{3t}$.) Next, applying the inner encoding scheme to each of the x_i's, viewed as an ℓ-bit long string, we obatina a 3ℓ-long seqiuence of 4ℓ-bit inner codewords. That is, using the inner code (i.e., the Toeplitz matrix) and additional $3t$ random ℓ-bit strings, denoted $y_1, ..., y_{3t}$, we encode each of the above x_i's by a 4ℓ-bit long string that is the result the multiplying the Toeplitz matrix with the vector $x_i y_i$. Hence, letting M denote the fixed 2ℓ-by-4ℓ Toeplitz matrix and $C : \mathrm{GF}(3t)^t \to \mathrm{GF}(3t)^{3t}$ denote the Reed-Solomon code, we have $E(x) = (x_1 y_1 \cdot M, ..., x_{3t} y_{3t} \cdot M)$, where $(x_1, ..., x_{3t}) \leftarrow C(xy)$ and $y, y_1, ..., y_{3t}$ are uniformly and independently distributed in the relevant domains (i.e., $y \in \{0,1\}^n$ and $y_1, ..., y_{3t} \in \{0,1\}^\ell$).

Clearly, E preserves the error-correcting features of Justesen's construction [6], and the rate

[3] A Toeplitz matrix, $T = (t_{i,j})$, satisfies $t_{i,j} = t_{i+1,j+1}$, for every i, j.

[4] The proof uses the fact that any (non-zero) linear combination of rows (or columns) in a random Toeplitz matrix is uniformly distributed. The first condition is proved by observing that the probability that a non-zero combination of the rows of the 2ℓ-by-4ℓ matrix has Hamming weight smaller than ℓ' is upper-bounded by $(2^{2\ell}-1)\cdot\sum_{i=0}^{\ell'-1}\binom{4\ell}{i}\cdot 2^{-4\ell}$, which is $o(1)$ for some $\ell' = \Omega(\ell)$. The second condition is proved by observing that the probability that there exist ℓ'' columns that yield a submatrix (of the last ℓ rows) that is not full rank is upper-bounded by $\binom{4\ell}{\ell''}\cdot(2^{\ell''}-1)\cdot 2^{-\ell}$, which is $o(1)$ for some $\ell'' = \Omega(\ell)$.

also remains constant (although cut by a factor of 4). The secrecy condition is proved analogously to the way in which the error correction feature is established in [6], where the analogy is between revealed codeword bits and corrupted codeword bits. Specifically, x_i remains secret if few bits in it are revealed, whereas the relatively few x_i's that cannot be guaranteed to remain secret do not harm the secrecy of x. Details follow.

Establishing the secrecy of E. We consider the partition of the codeword into consecutive 4ℓ-bit long subsequences corresponding to the codewords of the inner code. Given a set I of locations (as in the secrecy requirement), we consider the relative locations in each subsequence, denoting the induced locations in the i^{th} subsequence by I_i. We classify the subsequences into two categories depending on whether or not the size of the induced I_i is above the secrecy threshold for the inner-encoding. By a counting argument, only a small fraction of the subsequences have I_i's above the threshold.

For the typical (i.e., relatively small) I_i's, we use the secrecy feature of the inner-encoding, and infer that no information is revealed about the corresponding x_i's. Hence, the only information about x that may be present in $E(x)$ is present in the non-typical subsequences (i.e., those associated with large I_i's). Using the secrecy feature of the outer-encoding, we conclude that these few subsequences (or even the corresponding x_i's themselves) yield no information about x. The secrecy condition of the composed encoding follows. ■

3 An Efficient Wire-Tap Channel Encoding Scheme

The *Wire-Tap Channel* Problem, introduced by Wyner [9], generalized the standard setting of a Binary Symmetric Channel. Recall that a Binary Symmetric Channel with crossover probability p, denoted $\mathtt{BSC}_p$, is a randomized process which represents transmission over a noisy channel in which each bit is flipped with probability p (independently of the rest). Thus, for a string $\alpha \in \{0,1\}^n$, the random variable $\mathtt{BSC}_p(\alpha)$ equals $\beta \in \{0,1\}^n$ with probability $p^d \cdot (1-p)^{n-d}$, where d is the Hamming distance between α and β (i.e., the number of bits on which they differ). In the *Wire-Tap Channel Problem* there are two (independent) noisy channels from the sender: one representing the transmission to the legitimate receiver, and the other representing information obtained by an adversary tapping the legitimate transmission line and incurring some noise as well. In Wyner's work [9] the wire-tap channel introduces additional noise on top of the legitimate channel (and so may be thought of as taking place at the receiver's side). Here we consider a seemingly more difficult setting (introduced in [4]) in which the wire-tap channel is applied to the original packet being transmitted (and so may be thought of as taking place at the sender's side).

Wyner studied the information theoretic facet of the problem [9], analogously to Shannon's pioneering work on communication [8]. Below we consider the computational aspect of the problem for the special case of very noisy tapping-channel.

Theorem 2 (efficient wire-tap channel encoding): *Let (E,D) be a coding scheme as in Theorem 1 and let $\mathtt{BSC}_p(\alpha)$ be a random process which represents the transmission of a string α over a Binary Symmetric Channel with crossover probability p.*[5] *Then:*

1. Error Correction*: Decoding succeeds with overwhelmingly high probability. That is, for every $x \in \{0,1\}^*$,*

$$\text{Prob}[D(\mathtt{BSC}_{\frac{p}{2}}(E(x))) = x] \;=\; 1 - \exp(-\Omega(|x|)).$$

[5] Recall that the *crossover probability* is the probability that a bit is complemented in the transmission process.

2. Secrecy*: The wire-tapper gains no significant information. That is, for every* $x \in \{0,1\}^*$

$$\sum_{\alpha \in \{0,1\}^{|E(x)|}} \left| \text{Prob}[\mathtt{BSC}_{\frac{1}{2}-\frac{\varepsilon}{4}}(E(x)) = \alpha] \;-\; 2^{-|E(x)|} \right|$$

is exponentially vanishing in $|x|$*.*

Proof: Let η and ε be the constants associated with the error-correction and secrecy guarantees of E. Item 1 follows by observing that, with overwhelmingly high probability, the channel complements less than a η fraction of the bits of the codeword. Item 2 follows by representing $\mathtt{BSC}_{(1-\gamma)/2}(\alpha)$ as a two-stage process: In the first stage each bit of α is *set* (to its current value) with probability γ, independently of the other bits. In the second stage each bit which was not set in the first stage, is assigned a uniformly chosen value in $\{0,1\}$. Next, letting $\gamma = \varepsilon/2$, we observe that, with overwhelmingly high probability, at most $2 \cdot \gamma |E(x)| = \varepsilon \cdot |E(x)|$ bits were set in the first stage. Suppose we are in this case. Then, applying Item 3 of Theorem 1, the bits set in Stage 1 are uniformly distributed regardless of x, and due to Stage 2 the unset bits are also random. ■

Discussion: As mentioned above, the setting considered in Theorem 2 is actually due to Csiszár and Körner [4]. Clearly, a solution cannot exist unless the channel of Item 1 is more reliable than the one of Item 2. A special case of the results in [4] is that a solution always exists when the channel of Item 1 is more reliable than the one of Item 2. However, the latter result is non-constructive. In contrast, the result of Theorem 2 is constructive and efficient, but it requires a significant gap between the reliability of the two channels. In particular, the crossover probability of the channel in Item 1 (denoted $\frac{\eta}{2}$) is typically very small (i.e., of the order of 0.01); whereas the crossover probability of the channel in Item 2 (denoted $\frac{1}{2} - \frac{\varepsilon}{4}$) is typically very close to 1/2 (i.e., of the order of 0.49).

Crépeau (private comm., April 1997) has informed us that alternative solutions, which utilize bi-directional communication, may be obtained by using the techniques in [7, 1, 3]. We stress that when using bi-directional communication one can cope with an arbitrary pair of channels (and specifically the channel in the secrecy condition may be more reliable than the channel in the error-correcting condition) – see [7].

Acknowledgments

We are grateful to Moni Naor and Ronny Roth for helpful discussions. We also wish to thank Claude Crépeau for pointing out and explaining to us some related work (i.e., [4, 7, 1, 3]).

References

[1] C.H. Bennett, G. Brassard, C. Crépeau, and U. Maurer. Generalized Privacy Amplification. *IEEE Transaction on Information Theory*, Vol. 41 (6), pages 1915-1923. 1995.

[2] E. Berlekamp and L. Welch. Error correction of algebraic block codes. US Patent 4,633,470, 1986.

[3] C. Cachin and U.M. Maurer. Linking Information Reconciliation and Privacy Amplification, *Journal of Cryptology*, Vol. 10 (2), pages 97–110, 1997.

[4] I. Csiszár and J. Körner. Broadcast channels with confidential messages. *IEEE Transactions on Info. Theory*, Vol. 24, pages 339–348, 1978.

[5] S. Decatur, O. Goldreich and D. Ron. Computational Sample Complexity. In *10th COLT*, pages 130–142, 1997. (Later appeared in *SIAM Journal on Computing*, Vol. 29 (3), pages 854–879, 1999.)

[6] J. Justesen. A class of constructive asymptotically good algebraic codes. *IEEE Trans. Inform. Theory*, 18: 652–656, 1972.

[7] U.M. Maurer. Perfect Cryptographic Security from partially independent channels. In *23rd STOC*, 1991, pp. 561–571.

[8] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, Vol. 27, pages 379–423 and 623–656, 1948.

[9] A. D. Wyner. The wire-tap channel. *Bell System Technical Journal*, Vol. 54 (8), pages 1355–1387, 1975.