ON SECURE DISTRIBUTED STORAGE UNDER DATA THEFT

Phillip A. Regalia[1,2]

[1]National Science Foundation
CISE Directorate
Arlington, VA 22230
pregalia@nsf.gov

Chin-Yu Lin[2]

[2]Dept. Electrical Engineering & Computer Science
Catholic University of America
Washington, DC 20064
13lin@cardinalstudents.cua.edu

ABSTRACT

Consider a message coded for storage in which a fraction of the stored data is stolen. Ideally, the data remaining should allow message recovery, while the stolen data should reveal no information on the message. This gives a twist on the erasure wiretap channel, in that "Bob" no longer has a clear channel from "Alice". We show how the storage capacity can, as in other multi-terminal coding problems, be approached using nested codes, and propose nested erasure codes using Krylov subspaces. These offer good performance and perfect secrecy, while integrating the nested code structure naturally.

Key words: secure distributed storage; perfect secrecy.

1. INTRODUCTION

Consider a message $\mathbf{m}$ comprising k bits, which is to be stored securely in the "cloud": The bits (after coding) will be physically stored in a distributed data center, ideally accessible from anywhere, and with two basic requirements:

1. **Data integrity**: If some of the data storage facilities go offline or suffer damage, the message $\mathbf{m}$ can still be recovered from those remaining.
2. **Data security**: If an adversary compromises a data center and acquires the contents of the storage facilities, she will gain no information on the message $\mathbf{m}$.

The first constraint nominally requires having multiple data centers distributed geographically. If a data center goes offline (due to, e.g., power failure, natural disaster, or sabotage), the remaining data centers remain active, allowing message recovery. The simplest example is *mirroring*, whereby the contents of one server are copied verbatim to another, corresponding to a spatial repetition code. While this ensures data integrity, it offers no security, since an adversary need only "pwn" a single data center to steal an arbitrary message.

The data security constraint is arguably more crucial. Data stored in the cloud should expectedly have some value: Financial information, proprietary secrets, or strategic or classified data, constitute information that adversaries may attempt to steal, through whatever means available. This consideration would suggest that data be encrypted. We adopt the viewpoint that cryptography proves an insufficient defense, *not* because an adversary could crack the cryptosystem, but rather could obtain the key through some weakness in key management.[1] As such, we focus instead on information-theoretic secrecy: Even if an adversary obtains the entire data (call it $\mathbf{Z}$) of one (or even a few) of the data centers, no information on any useful message will be leaked: $I(\mathbf{m};\mathbf{Z}) \to 0$.

Other scenarios may likewise be envisaged. An unscrupulous employee, for example, may walk off with a hard disk from a RAID array hoping to gain valuable data, or data communication between shared servers may be siphoned off from a man-in-the-middle attack. This leads us to consider a *partition channel*, as sketched in figure 1, in which Eve (the adversary) steals a fraction α of the data, and Bob (the ligitimate owner) retains the remaining fraction $1-\alpha$ of the data. Here Alice is a client who "entrusts" her data to the cloud. This is more complicated than the erasure wiretap channel (e.g., [4]–[6]), since Bob no longer has a clear channel from Alice.

2. STORAGE CAPACITY

As is well known (e.g., [7]–[9]), data integrity calls for a suitable erasure code. Formally, a k-bit message $\mathbf{m}$ is mapped to an n-bit code word $\mathbf{x}$, which is written for storage among the data centers. A portion $n\alpha$ of the bits are stolen by Eve (and designated by $\mathbf{z}$), while the remaining $n(1-\alpha)$ are retained by Bob (and designated by $\mathbf{y}$). Data integrity means that Bob can recover the message $\mathbf{m}$, while data security means that Eve infers negligible information on the message from her stolen bits. The largest attainable ratio k/n under these constraints is the storage capacity C. The secrecy capacity result of [10], [11] translates directly to:

Property 1 *The storage capacity C, subject to data integrity and data security constraints if up to a fraction α of the bits is stolen, is $C = 1-2\alpha$.*

[1]Social engineering schemes (e.g., [1]–[3]) exploit the human factor as the weakest link in the chain: Someone having access to a decryption key may inadvertently cough it up, or may be recruited by an adversary.

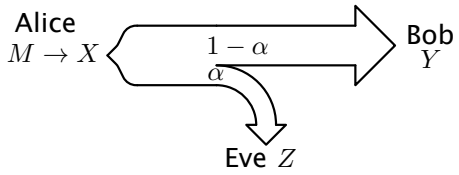

Fig. 1. Partition channel, in which Eve steals up to a fraction α of the bits.

The result is verified easily by calculating the difference in channel capacities connecting Alice to Bob and Alice to Eve, respectively[2]; Bob's erasure channel has capacity $1-\alpha$, while Eve's has capacity α, giving the difference $C = 1-2\alpha$.

3. CODE CONSTRUCTION AND PRIOR WORK

Let the message $\mathbf{m}$ comprise $\lfloor nC \rfloor = \lfloor n(1-2\alpha) \rfloor$ bits, and let X (resp., Z) denote the set of n-bit code words (resp., $n\alpha$-bit stolen portions). As in [4], [12], consider $2^{\lfloor nC \rfloor}$ code books $\{\mathcal{B}(\mathbf{m})\}$, one for each message realization $\mathbf{m}$. For a given message $\mathbf{m}$, the actual code word $\mathbf{x} \in X$ is randomly selected from $\mathcal{B}(\mathbf{m})$.

Each code book $\mathcal{B}(\mathbf{m})$ is *capacity saturating*[3] for Eve's channel (with capacity $C_E = \alpha$) provided $I(X;Z|M=\mathbf{m}) \geq n(\alpha-\delta)$ for a small constant δ. Therefore,

$$\begin{aligned} I(X;Z|M) &= \sum_{\mathbf{m}} \Pr(M=\mathbf{m})\, I(X;Z|M=\mathbf{m}) \\ &\geq n(\alpha-\delta) \quad \text{for sufficiently large } n. \end{aligned} \tag{1}$$

To show the link between capacity saturation and weak secrecy [16], expand $I(MX;Z)$ in two ways (as in [10], [4]):

$$\begin{aligned} I(MX;Z) &= I(M;Z) + I(X;Z|M) \\ &= I(X;Z) + I(M;Z|X) \end{aligned}$$

Now, the capacity of Eve's channel is given as $C_E = \alpha = \sup_{P_X(X)}[I(X;Z)/n]$ so that $I(X;Z)/n \leq \alpha$. And since $M \to X \to Z$ forms a Markov chain, we have $I(M;Z|X) = 0$ [17, §2.8]. We may thus isolate $I(M;Z)$ as

$$\begin{aligned} I(M;Z) &= I(X;Z) - I(X;Z|M) \\ &\leq n[\alpha - (\alpha-\delta)] = n\delta. \end{aligned}$$

Thus $I(M;Z)/n \leq \delta$, consistent with weak secrecy [16]. (For the finite-block length case we consider, strong secrecy becomes superfluous.) The rate R_E of each code book $\mathcal{B}(\mathbf{m})$ is *lower* bounded as $R_E = H(X|M)/n \geq I(X;Z|M)/n \geq (\alpha-\delta) = C_E - \delta$; hence the moniker *capacity saturating*.

A practical realization uses nested codes. Consider mapping the $\lfloor nC \rfloor$-bit message $\mathbf{m}$ to an n-bit word $\mathbf{x}$ as per

$$\begin{bmatrix} \mathbf{0} \\ \mathbf{m} \end{bmatrix} \equiv \underbrace{\begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_\Delta \end{bmatrix}}_{\mathbf{H}} \mathbf{x} \pmod 2 \tag{2}$$

Here $\mathbf{H}_1$ is a binary parity-check matrix of a *capacity-approaching* code for a channel with erasure probability α, meaning that if we erase up to a fraction α of bits from $\mathbf{x}$ (call the erased version $\mathbf{y}$), a decoding algorithm using $\mathbf{y}$ and $\mathbf{H}_1$ can recover $\mathbf{x}$ (with high enough probability), and thus also recover the message via $\mathbf{m} = \mathbf{H}_\Delta \mathbf{x} \bmod 2$. The null space of $\mathbf{H}_1$ gives the *fine code* [18]. If the size of $\mathbf{H}_1$ is $l \times n$, then the code rate is $R_B = 1-(l/n)$. As the capacity of a channel that erases a fraction α of the bits is $C_B = 1-\alpha$, the inequality $R_B < C_B$ imposes

$$(l/n) > \alpha. \tag{3}$$

Let now $\mathcal{B}(\mathbf{0})$ denote the nullspace of the parity-check matrix $\mathbf{H}$ in (2) (known commonly as the *coarse code* [18]):

$$\mathcal{B}(\mathbf{0}) = \{\mathbf{x} : \mathbf{H}\mathbf{x} \equiv \mathbf{0} \pmod 2\}.$$

From the secrecy constraint above, we desire that this give a capacity saturating code for an erasure channel with erasure probability $1-\alpha$. Similarly, let $\mathcal{B}(\mathbf{m})$ denote the coset

$$\mathcal{B}(\mathbf{m}) = \left\{ \mathbf{x} : \mathbf{H}\mathbf{x} \equiv \begin{bmatrix} \mathbf{0} \\ \mathbf{m} \end{bmatrix} \pmod 2 \right\},$$

where $\mathbf{H}$ is partitioned as in (2), giving a code book indexed by the message $\mathbf{m}$. Each code book $\mathcal{B}(\mathbf{m})$ has identical distance properties, since one differs from another by an offset.

If the dimensions of $\mathbf{H}$ are $j \times n$, then its rate is $R_E = 1-(j/n)$, and the inequality $R_E \geq C_E$ implies $j/n \leq 1-\alpha$. When combined with (3), the size of the message $\mathbf{m}$ in (2) can now be bounded as

$$\frac{\text{length}(\mathbf{m})}{n} = \frac{j-l}{n} < 1-2\alpha\,,$$

consistent with Property 1.

4. KRYLOV SUBSPACE CODES

Let $\mathbf{H}$ be a parity-check matrix whose columns form a Krylov sequence [19, §9.1.1]:

$$\mathbf{H} = [\,\mathbf{b} \quad \mathbf{A}\mathbf{b} \quad \mathbf{A}^2\mathbf{b} \quad \mathbf{A}^3\mathbf{b} \quad \cdots \quad \mathbf{A}^{n-1}\mathbf{b}\,]. \qquad (j \times n)$$

Here $\mathbf{A}$ is a $j \times j$ binary matrix (with $j < n$) and $\mathbf{b}$ is a $j \times 1$ binary vector, and each product $\mathbf{A}^\ell\mathbf{b}$ (with $1 \leq \ell \leq n-1$)

[2]The result from [11] properly asserts the secrecy capacity to be

$$C = \max_{U \to X \to (Y,Z)} [I(U;Y) - I(U;Z)]$$

where U is an auxiliary random variable that forms a Markov chain with X and (Y,Z). If the eavesdropper's channel is degraded, meaning that $I(U;Y) \geq I(U;Z)$ for all Markov chains $U \to X \to (Y,Z)$, this simplifies to $C = \max_{P(x)}[I(X;Y) - I(X;Z)]$ with $P(X)$ the input distribution. If this same distribution maximizes both $I(X;Y)$ and $I(X;Z)$ (applicable here since both channels are symmetric, and hence maximized by a uniform input distribution), the maximized difference of mutual information terms becomes the difference of channel capacities.

[3]Some authors offer here instead a *capacity approaching* code, which would reverse the inequality (1) and thus belie a finite storage capacity; for connections to channel resolvability [13], [14], see [15].

is calculated modulo-2. We assume that $\mathbf{H}$ has full rank j, which is equivalent to the pair $(\mathbf{A}, \mathbf{b})$ being completely controllable in linear system parlance (e.g., [20]). If, in addition, $\mathbf{A}$ is invertible $\bmod 2$, then controllability of $(\mathbf{A}, \mathbf{b})$ implies that any successive j columns $\mathbf{A}^\ell\mathbf{b}$, $\mathbf{A}^{\ell+1}\mathbf{b}$, …, $\mathbf{A}^{\ell+j-1}\mathbf{b}$ remain linearly independent, as per a "good" erasure code [21].

Suppose now $\mathbf{A}$ and $\mathbf{b}$ are partitioned according to

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{A}_{12} & \mathbf{A}_{22} \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}, \tag{4}$$

with $\mathbf{A}_{11}$ of dimensions $l \times l$ and $\mathbf{b}_1$ of dimensions $l \times 1$. The parity-check matrix $\mathbf{H}$ then naturally partitions as $\mathbf{H} = \left[\begin{smallmatrix} \mathbf{H}_1 \\ \mathbf{H}_\Delta \end{smallmatrix}\right]$ in which

$$\mathbf{H}_1 = [\, \mathbf{b}_1 \quad \mathbf{A}_{11}\mathbf{b}_1 \quad \mathbf{A}_{11}^2\mathbf{b}_1 \quad \cdots \quad \mathbf{A}_{11}^{n-1}\mathbf{b}_1 \,]. \qquad (l \times n)$$

This likewise assumes a Krylov sequence structure, and thus should also provide a "good" erasure code provided $\mathbf{A}_{11}$ is invertible and $(\mathbf{A}_{11}, \mathbf{b}_1)$ is controllable.

One may check that $\mathbf{H}$ will generically be a "medium density" parity-check matrix, since the columns will have about half their entries equal to 1, for random choices of $\mathbf{A}$ and $\mathbf{b}$. As such, message-passing decoding will not perform well with such parity-check matrices. Instead, linear algebra decoding works well for reasonable block lengths n; see §5.

We recall that the Cayley-Hamilton theorem [20] asserts that a square matrix satisfies its own characteristic equation. That is, given the characteristic polynomial $p(\lambda)$ defined as

$$\begin{aligned} p(\lambda) &= \det(\lambda\mathbf{I} - \mathbf{A}) \\ &\equiv \lambda^j + a_{j-1}\lambda^{j-1} + \cdots + a_2\lambda^2 + a_1\lambda + a_0, \end{aligned}$$

with each $a_i \in \{0, 1\}$ (by finite field conventions [22]), the matrix function obtained by replacing λ with $\mathbf{A}$ (see [23], [19, Ch. 11]) gives a null result:

$$p(\mathbf{A}) = \mathbf{A}^j + a_{j-1}\mathbf{A}^{j-1} + \cdots + a_1\mathbf{A} + a_0\mathbf{I} \equiv \mathbf{0} \pmod 2.$$

To deduce the code words $\{\mathbf{x} : \mathbf{H}\mathbf{x} \equiv \mathbf{0} \pmod 2\}$, the coefficients $\{a_k\}$ hold the key, since the relation

$$p(\mathbf{A}) = \sum_{k=0}^{j} a_k\mathbf{A}^k \equiv \mathbf{0} \pmod 2$$

clearly implies that

$$\sum_{k=0}^{j} a_k\mathbf{A}^k\mathbf{b} \equiv \mathbf{0}, \quad \text{or} \quad [\mathbf{b} \;\; \mathbf{A}\mathbf{b} \;\; \mathbf{A}^2\mathbf{b} \;\cdots\; \mathbf{A}^j\mathbf{b}] \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_j \end{bmatrix} \equiv \mathbf{0}.$$

This shows which combinations of the rows of $\mathbf{H}$ sum to zero. This relation also implies that, for any integer $\ell > 0$,

$$\sum_{k=0}^{j} a_k\mathbf{A}^{k+\ell}\mathbf{b} = \mathbf{A}^\ell \underbrace{\sum_{k=0}^{j} a_k\mathbf{A}^k\mathbf{b}}_{\mathbf{0}} \equiv \mathbf{0}.$$

We can combine these relations into matrix form as

$$\underbrace{[\mathbf{b} \quad \mathbf{A}\mathbf{b} \quad \cdots \quad \mathbf{A}^{n-1}\mathbf{b}]}_{\mathbf{H}\ (j \times n)} \underbrace{\begin{bmatrix} a_0 & 0 & \cdots & 0 \\ a_1 & a_0 & \ddots & \vdots \\ a_2 & a_1 & \ddots & 0 \\ \vdots & \ddots & \ddots & a_0 \\ a_j & a_{j-1} & \ddots & \vdots \\ 0 & a_j & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{j-1} \\ 0 & \cdots & 0 & a_j \end{bmatrix}}_{\mathbf{G}\ [n \times (n-j)]} \equiv \mathbf{0}$$

to identify the generator matrix $\mathbf{G}$ for the coarse code. Thus any $\mathbf{x}$ satisfying $\mathbf{H}\mathbf{x} \equiv \mathbf{0} \pmod 2$ can be written as $\mathbf{x} = \mathbf{G}\boldsymbol{\xi} \bmod 2$ for some $\boldsymbol{\xi}$. Using instead the coefficients from the characteristic polynomial of $\mathbf{A}_{11}$, we obtain the generator matrix $\mathbf{G}_1$ of the fine code, in the same manner.

Remark: Note that $a_0 = \det \mathbf{A} \bmod 2$; we should thus ensure that $a_0 \neq 0$, i.e., that $\mathbf{A}$ be invertible modulo-2. Otherwise $\mathbf{G}$ would vanish in its first row, and any code word $\mathbf{x}$ would have zero as its first entry. ◇

To each code word $\mathbf{x} = [x_0, x_1, \ldots, x_{n-1}]^T$ we may associate a polynomial

$$\mathbf{x}(\lambda) = x_0 + x_1\lambda + x_2\lambda^2 + \cdots + x_{n-1}\lambda^{n-1}.$$

In view of the Toeplitz structure of $\mathbf{G}$ (constant along any diagonal), the operation $\mathbf{x} = \mathbf{G}\boldsymbol{\xi}$ translates to

$$x(\lambda) = p(\lambda)\, \xi(\lambda)$$

where $\xi(\lambda) = \sum_k \xi_k\lambda^k$ is built from the information bits $\boldsymbol{\xi}$ fed to the generator matrix, and $p(\lambda)$ is the characteristic polynomial of $\mathbf{A}$, giving thus the generator polynomial for the code. It is straightforward to check that

$$p(\lambda) = q(\lambda)\, r(\lambda)$$

where $q(\lambda)$ [resp., $r(\lambda)$] is the characteristic polynomial of $\mathbf{A}_{11}$ [resp., $\mathbf{A}_{22}$] in (4). Thus nested codes have generator polynomials forming a divisibility chain.

The connection to cyclic codes is straightforward to show:

Property 2 *A Krylov-generated parity check matrix* $\mathbf{H}$ *describes a cyclic code if and only if* $\mathbf{A}^n \equiv \mathbf{I} \pmod 2$.

For sufficiency, let $\mathbf{x} = [x_0, x_1, \ldots, x_{n-1}]^T$ be a code word, so that $\mathbf{0} = \mathbf{H}\mathbf{x} = \mathbf{b}x_0 + \mathbf{A}\mathbf{b}x_1 + \cdots + \mathbf{A}^{n-1}\mathbf{b}x_{n-1}$. This still vanishes if multiplied by $\mathbf{A}$, so that

$$\begin{aligned} \mathbf{0} &= \mathbf{A}\mathbf{b}x_0 + \mathbf{A}^2\mathbf{b}x_1 + \cdots + \mathbf{A}^n\mathbf{b}x_{n-1} \\ &\equiv \mathbf{b}x_{n-1} + \mathbf{A}\mathbf{b}x_0 + \cdots + \mathbf{A}^{n-1}\mathbf{b}x_{n-2} \qquad \pmod 2 \end{aligned}$$

since $\mathbf{A}^n\mathbf{b} \equiv \mathbf{b}$. This confirms that $[x_{n-1}, x_0, \ldots, x_{n-2}]^T$ is also a code word. For necessity, the generator polynomial $p(\lambda)$ of any cyclic code must divide $\lambda^n - 1$ (e.g., [24, §5.2.3], [22, §5.4]), so that $\lambda^n - 1 = p(\lambda)\, Q(\lambda)$ for some polynomial $Q(\lambda)$. Replacing λ by $\mathbf{A}$ gives $\mathbf{A}^n - \mathbf{I} = p(\mathbf{A})\, Q(\mathbf{A}) \equiv \mathbf{0}$ since $p(\mathbf{A}) \equiv \mathbf{0}$ by the Cayley-Hamilton theorem. ◇

5. LINEAR ALGEBRAIC DECODING

Let $\mathbf{x} = \mathbf{G}_1\boldsymbol{\xi}$ be a word from the fine code, in which certain positions have been erased, with the indices of the known positions denoted $i_1, i_2, \ldots, i_m$. We may then write the equations for the known bits using a submatrix built from rows i_1, $\ldots$, i_m of the generator matrix $\mathbf{G}_1$ for the fine code as

$$\begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_m} \end{bmatrix} = \underbrace{\begin{bmatrix} g_{i_1,1} & g_{i_1,2} & \cdots & g_{i_1,n-l} \\ g_{i_2,1} & g_{i_2,2} & \cdots & g_{i_2,n-l} \\ \vdots & \vdots & \ddots & \vdots \\ g_{i_m,1} & g_{i_m,2} & \cdots & g_{i_m,n-l} \end{bmatrix}}_{\overline{\mathbf{G}}_1} \underbrace{\begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{n-l} \end{bmatrix}}_{\boldsymbol{\xi}} . \quad (5)$$

Here the elements of $\boldsymbol{\xi}$ are the unknowns. Provided the matrix $\overline{\mathbf{G}}_1$ has full column rank (here, $n-l$), a unique solution exists for $\boldsymbol{\xi}$ which may be obtained in polynomial time using, e.g., Gaussian elimination in modulo-2 arithmetic. Once $\boldsymbol{\xi}$ is obtained, the entire code word $\mathbf{x}$ may be regenerated as $\mathbf{x} = \mathbf{G}_1\boldsymbol{\xi}$, and the erased positions are then recovered. When $\overline{\mathbf{G}}_1$ has rank less than $n-l$, a unique solution for the code word $\mathbf{x}$ does not exist.

Simulation example. Consider a nested code with length $n = 500$, and a message length of $k = 150$ bits. Any practical code must have rate below capacity, i.e., $(k/n) \le C = 1 - 2\alpha$, giving $\alpha \le (1 - k/n)/2 = 0.35$ as the maximum fraction of stolen bits. Here $l = n\alpha = 175$ is the number of rows in $\mathbf{H}_1$ (the parity-check matrix for the fine code). $\mathbf{A}$ and $\mathbf{b}$ were randomly generated, and tested for whether $\mathbf{A}$ was invertible and $(\mathbf{A}, \mathbf{b})$ controllable [and likewise for $(\mathbf{A}_{11}, \mathbf{b}_1)$], and retained once affirmative responses were returned.

From the message $\mathbf{m}$, a code word $\mathbf{x}$ is generated as

$$\begin{matrix} l \text{ rows} \{ \\ k \text{ rows} \{ \end{matrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{m} \end{bmatrix} = \begin{bmatrix} \mathbf{H}_1 \\ \mathbf{H}_\Delta \end{bmatrix} \mathbf{x}_p, \quad \mathbf{x} = \mathbf{x}_p + \mathbf{G}\boldsymbol{\xi} \in \mathcal{B}(\mathbf{m})$$

in which $\mathbf{x}_p$ can be obtained using linear algebra and $\boldsymbol{\xi}$ is chosen randomly. Both Bob and Eve know the nested code $\mathbf{H} = \left[\begin{smallmatrix} \mathbf{H}_1 \\ \mathbf{H}_\Delta \end{smallmatrix}\right]$, but not the message $\mathbf{m}$, much less which code word $\mathbf{x}$ is selected from $\mathcal{B}(\mathbf{m})$. Knowing only that $\mathbf{H}_1\mathbf{x} = \mathbf{0}$, Bob and Eve both attempt to construct generator bits $\boldsymbol{\xi}$ from their available bits using (5), giving in either case a code word estimate $\hat{\mathbf{x}}$, and thus a message estimate $\widehat{\mathbf{m}} = \mathbf{H}_\Delta\hat{\mathbf{x}}$.

Figure 2 plots the bit error rate in estimating the message $\mathbf{m}$ versus the actual fraction β of bits stolen by Eve (with $\beta \le \alpha$), averaged over 500 independent realizations of the erasure positions in $\mathbf{x}$ for each theft fraction β. Bob's bit error rate is identically zero up to fractions exceeding 0.33, indicating that his erasure code is close to capacity. The rise in bit error rate as β approaches α (the capacity limit) is due to an increased number of $\overline{\mathbf{G}}_1$ submatrices in (5) having insufficient rank.

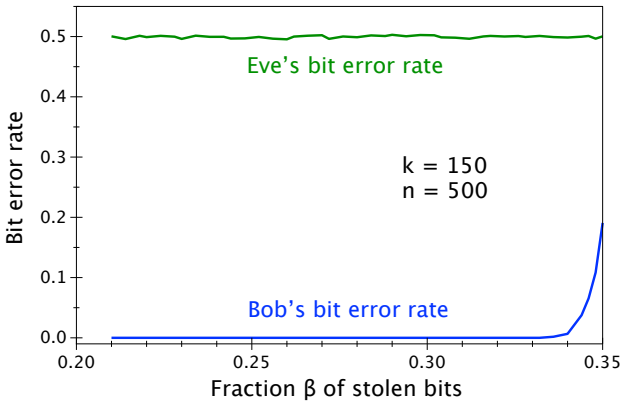

Fig. 2. Bit error rates for Bob and Eve versus the actual fraction $\beta \le \alpha$ of stolen bits, with $\alpha = 0.35$ the capacity limit.

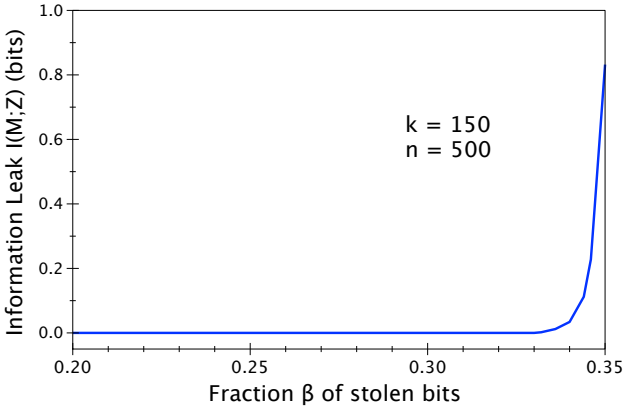

Fig. 3. Number of bits leaked versus fraction β of stolen bits.

Eve's system of equations persistently presents a rank-deficient $\overline{\mathbf{G}}_1$, offering at least $2^{n(1-\alpha-\beta)}$ solutions for $\boldsymbol{\xi}$ and thus for her estimate $\hat{\mathbf{x}}$. A BER of 0.5 implies that her message estimate is statistically no better than a coin flip. ◇

Let $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_\ell$ be a basis for the right null-space of $\overline{\mathbf{G}}_1$ in Eve's system (5), and set $\mathbf{F} \triangleq \mathbf{H}_\Delta\mathbf{G}_1\,[\boldsymbol{\eta}_1\ \boldsymbol{\eta}_2\ \cdots\ \boldsymbol{\eta}_\ell]$.

Property 3 *Eve's uncertainty is* $H(M|Z = \mathbf{z}) = \mathrm{rank}(\mathbf{F})$.

The proof is simple and so omitted. We can estimate the equivocation $I(M;Z) = H(M) - \sum_{\mathbf{z}} H(M|Z = \mathbf{z})\Pr(\mathbf{z})$ by averaging over realizations of $\mathbf{z}$, as plotted in Figure 3, which shows identically zero leakage for $\beta \le 0.33$, and less than one bit leaked as $\beta \to \alpha$. For longer block lengths n, concatenated codes [25], polar codes [26] or LDPC designs [5], [12] would logically prove of interest, if a coherent nested code design procedure can be harnessed; see, e.g., [27].

6. REFERENCES

[1] S. M. White, "Social engineering," in *IEEE Information Assurance Worskhop*, pp. 388–389, 2003.

[2] L. Laribee, D. S. Barnes, N. C. Rowe, and C. H. Martell, "Analysis and defensive tools for social engineering attacks on computer systems," in *Int. Conf. Engineering of Computer-Based Systems*, pp. 261–267, 2006.

[3] E. Rabinovitch, "Staying protected from 'social engineering'," *IEEE Communications Mag.*, vol. 45, pp. 20–21, Sept. 2007.

[4] R. Liu, Y. Liang, H. V. Poor, and P. Spasojević, "Secure nested codes for type II wiretap channels," in *Information Theory Workshop*, (Tahoe City, CA), pp. 337–342, Sept. 2007.

[5] A. Thangaraj, S. Dihidar, A. R. Calderbank, S. W. McLaughlin, and J.-M. Merolla, "Applications of LDPC codes to the wiretap channel," *IEEE Trans. Information Theory*, vol. 53, pp. 2933–2945, Aug. 2007.

[6] A. Mills, B. Smith, T. C. Clancy, E. Soljanin, and S. Vishwanath, "On secure communication over wireless erasure networks," in *Int. Symp. Information Theory*, (Toronto), pp. 161–165, July 2008.

[7] A. G. Dimakis, V. Prabhakaran, and K. Ramchandran, "Decentralized erasure codes for distributed network storage," *IEEE Trans. Information Theory*, vol. 52, pp. 2809–2816, June 2006.

[8] H.-Y. Lin and W.-G. Tzeng, "A secure decentralized erasure code for distributed network storage," *IEEE Trans. Parallel and Distributed Systems*, vol. 21, pp. 1586–1594, Nov. 2010.

[9] S. G. Harihara, B. Janakiram, M. G. Chandra, K. G. Aravind, S. Kandhe, P. Balamuralidhar, and B. S. Adiga, "Spreadstore: An LDPC erasure code scheme for distributed storage systems," in *Int. Conf. Data Storage and Data Engineering*, pp. 154–158, 2010.

[10] A. D. Wyner, "The wire-tap channel," *Bell System Technical Journal*, vol. 54, pp. 1355–1387, Oct. 1975.

[11] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Trans. Information Theory*, vol. 24, pp. 339–348, May 1978.

[12] A. Subramanian, A. Thangaraj, M. Bloch, and S. W. McLaughlin, "Strong secrecy on the binary erasure wiretap channel using large-girth LDPC codes," *IEEE Trans. Information Forensics and Security*, vol. 6, pp. 585–594, Sept. 2011.

[13] A. Wyner, "The common information of two dependent random variables," *IEEE Trans. Information Theory*, vol. 21, no. 2, pp. 163–179, 1975.

[14] T. Han and S. Verdú, "Approximation theory of output statistics," *IEEE Trans. Information Theory*, vol. 39, pp. 752–772, May 1993.

[15] M. R. Bloch and N. J. Laneman, "Secrecy from resolvability." (online) arXiv:1105.5419, 2011.

[16] A. Subramanian, A. T. Suresh, S. Raj, A. Thangaraj, M. Bloch, and S. McLaughlin, "Strong and weak secrecy in wiretap channels," in *Proc. Int. Symp. Turbo Codes*, (Brest, France), pp. 30–34, Sept. 2010.

[17] T. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ: Wiley, 2nd ed., 2006.

[18] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. Information Theory*, vol. 48, pp. 1250–1276, June 2002.

[19] G. H. Golub and C. F. van Loan, *Matrix Computations*. Baltimore, MD: Johns Hopkins Univ. Press, 2nd ed., 1989.

[20] T. Kailath, *Linear Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[21] M. Tomlinson, C. Tijhai, and M. Ambroze, "Analysis of the distribution of the number of erasures correctable by a binary linear code," *IET Communications*, vol. 1, no. 3, pp. 539–548, 2007.

[22] S. G. Wilson, *Digital Modulation and Coding*. Upper Saddle River, NJ: Prentice-Hall, 1996.

[23] R. A. Roberts and C. T. Mullis, *Digital Signal Processing*. Reading, MA: Addison-Wesley, 1987.

[24] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 2nd ed., 1983.

[25] A. Graell i Amat and E. Rosnes, "Good concatenated code ensembles for the binary erasure channel," *IEEE Trans. Selected Areas in Communications*, vol. 27, pp. 928–943, Aug. 2009.

[26] H. Mahdavifar and A. Vardy, "Achieving the secrecy capacity of wiretap channels using polar codes," *IEEE Trans. Information Theory*, vol. 57, pp. 6428–6443, Oct. 2011.

[27] C. Kelley and J. Kliewer, "Algebraic construction of graph-based nested codes from protographs," in *Proc. Int. Symp. Information Theory*, (Austin, TX), pp. 829–833, June 2010.