ON BURST ERROR CORRECTION AND STORAGE SECURITY OF NOISY DATA*

FELIX FONTEIN , KYLE MARSHALL , JOACHIM ROSENTHAL , DAVIDE SCHIPANI , AND ANNA-LENA TRAUTMANN†

Abstract. Secure storage of noisy data for authentication purposes usually involves the use of error correcting codes. We propose a new model scenario involving burst errors and present for that several constructions.

Key words. Storage security, fuzzy data, biometric authentication, burst errors, two-dimensional codes, Reed-Solomon codes, concatenated codes.

AMS subject classifications. 94B20, 94A62, 68P30.

1. Introduction. Sensible data, as e.g. passwords, are usually stored via the use of a one-way hash function, so that retrieving the data implies comparing the stored hash with the hash of the new input. In many contexts, though, this simple procedure cannot be used due to the noisy or fuzzy nature of the data. An outstanding example is the storage of biometric data, e.g. in the form of fingerprint, iris, voice, DNA etc., for the purpose of authentication: the data derived from different acquisitions of the same biometric feature can slightly change from each other, and the biometric feature can slightly change itself for different reasons [11]. Therefore a certain threshold of tolerance is needed to distinguish legitimate from non legitimate users, but this prevents the standard use of collision resistant hash functions [9].

This problem has led to the proposal of systems for the secure storage of biometric passwords (see [10] for a selected survey of the literature), which essentially act as "tolerant" hash functions. The idea behind most of these methods is a combined use of error correcting codes and hash functions, whose model is the fuzzy commitment scheme [6].

A scheme which is apparently just the dual of this is the syndrome fuzzy hashing construction: in [1] we showed that it offers several advantages with respect to the fuzzy commitment scheme, in particular as far as information leakage is concerned.

The fuzzy commitment scheme has later been generalized to other types of metrics, such as the set difference metric [5] and the edit distance metric [3].

In particular, the fuzzy vault [5] uses polynomial interpolation in order to allow authentication based on the matching of a sufficient number of features, while the fuzzy extractor [3] is a further generalization which combines the previous constructions with particular objects called random extractors. These make the previous schemes stronger with respect to information leakage, although they cannot prevent it [2, 4].

The choice of one scheme rather than another depends not only on the application or model scenario, but also on important issues, like for example privacy concerns, as already mentioned, or suitability for implementation, as we already discussed in [1, 9].

*Research partially supported by Armasuisse and Swiss National Science Foundation Projects no. 132256 and no. 138080.

†The authors are with the Institute of Mathematics, University of Zurich, Switzerland (http://www.math.uzh.ch/aa).

In this paper we will slightly modify the model scenario: we will tolerate up to, say, ℓ error bursts of maximum length b, and possibly assume also the presence of other random errors. This assumption is realistic in many contexts, where errors are likely to appear in bursts, and leads us to focus on and take advantage of burst error correcting codes. We will investigate different choices of codes, depending on the type and the dimension of the bursts, and elaborate on their key features, like error correcting capability and decoding complexity.

The structure of the paper is the following: in Section 2 we review the syndrome fuzzy hashing construction, which can be taken as our model scheme. The rest of the paper is devoted to solutions using burst error-correcting codes: we deal with base field representations of Reed-Solomon codes in Section 3 and with concatenated codes in Section 4. Lastly Section 5 presents some conclusions.

2. Syndrome fuzzy hashing and burst error correction. We review here the syndrome fuzzy hashing construction, as presented in [1].

Suppose we need a tolerance of e errors, then an $[n, k]$-linear block code $C \subset \mathbb{F}_q^n$, able to correct e errors, is selected, and it is described through its $r \times n$ parity-check matrix H, with $r = n - k$. Given a data vector x to be stored, the pair $(H_a(x), Hx)$ is used to represent x, were H_a is a given hash function. When another vector y is acquired and is compared with x, the value $Hx - Hy = H(x - y) = Hv$ is computed, that coincides with the syndrome associated to the difference vector $v = x - y$. Then, syndrome decoding is applied on Hv, according to the chosen code C. If $d(x, y) \leq e$, then v, which has Hamming weight equal to $d(x, y)$, corresponds to a correctable error vector. So, syndrome decoding succeeds and correctly results in v. Then, starting from v and y, x can be computed, as well as $H_a(x)$. The latter coincides with the stored value, so authentication succeeds. Otherwise, syndrome decoding fails or reports $w \neq v$. In such case, $x' = w + y \neq x$ and $H_a(x') \neq H_a(x)$ is obtained, and authentication fails.

As stated before, this construction is the dual version of the fuzzy commitment scheme, but offers better security in terms of reduced information leakage.

One can transform an $[n, k]$-linear block code into a two-dimensional (array) code by writing the codewords into an $n_1 \times n_2$-array such that $n_1 n_2 = n$. A simple way of doing so is writing the vector entries into the array row by row. For syndrome decoding such a two-dimensional code is transformed back into the vector representation of the original linear block code and can be used as explained before.

Definition 2.1.

1. *A* one-dimensional burst (error) of length *m is a vector of length m over $\mathbb{F}_q$, such that the first and last entries are non-zero.*
2. *A* two-dimensional rectangular burst (error) of size *$m \times m'$ is an $m \times m'$-matrix over $\mathbb{F}_q$, such that the first and last columns and the first and last rows each contain a non-zero element.*

Note that a one-dimensional burst may occur horizontally or vertically in a two-dimensional code.

Burst errors of this type appear in many real-life applications, so that in many contexts the syndrome fuzzy hashing construction can exploit the power of burst error correcting codes.

3. Base field representations of Reed-Solomon codes. Reed-Solomon codes are optimal codes defined over an extension field $\mathbb{F}_{q^m}$. When the entries are expanded over the base field, one gets a code over $\mathbb{F}_q$ that may be used for burst error correction in different ways.

Definition 3.1. *Consider the extension field $\mathbb{F}_{q^m}$ and $n = q^m - 1$ distinct elements $x_1, \ldots, x_n \in \mathbb{F}_{q^m}$, and let $0 < k < n$. Then*

$$RS = \{(f(x_1), \ldots, f(x_n)) \mid f \in \mathbb{F}_{q^m}[x], \deg(f) < k\} \subseteq \mathbb{F}_{q^m}^n$$

is called a Reed-Solomon *(RS) code. It has length n, dimension k and minimum distance $d = n - k + 1$, which means that it is optimal (MDS).*

Expanding as row vectors (Construction I). When expanding the elements of $\mathbb{F}_{q^m}$ in a Reed-Solomon code as elements of $\mathbb{F}_q^m$ via an $\mathbb{F}_q$-isomorphism $\mathbb{F}_{q^k} \cong \mathbb{F}_q^k$, one gets a q-ary linear code of length $n' = nm = m(q^m - 1)$ and dimension $k' = mk$. The minimum distance is $d' \geq d$ but the burst-error correction capability is higher than the random error correction capability. In fact, since at most $\lfloor \frac{n-k}{2} \rfloor$ symbols from $\mathbb{F}_{q^m}$ can be corrupted, it is readily seen that such a code is able to correct any

- single burst of length $\leq m(\lfloor \frac{n-k}{2} \rfloor - 1) + 1$, or
- any ℓ many bursts of length $\leq m(\lfloor \frac{n-k}{2\ell} \rfloor - 1) + 1$.

In the binary case, this expansion of codes over $\mathbb{F}_{2^m}$ into binary codes can be found in [7, Ch. 10 §5]. There one can also find that adding a parity check bit in the end of each m-vector results in a code of length $n' = (m+1)(2^m - 1)$, dimension $k' = mk$ and minimum distance $d' \geq 2d$. The analogue holds for q-ary codes. The higher minimum distance means that one can correct more random errors with this variation of the construction.

Expanding as matrices (Constructions II and III). RS codes in q-ary representation can also be exploited if one wants to compare two data matrices (or higher dimensional data), if the difference pattern is likely to consist of two-dimensional bursts.

To obtain a two-dimensional q-ary representation of the Reed-Solomon codes, it may be convenient to write each symbol from $\mathbb{F}_{q^m}$ into a rectangle or a square array. This can be done in different ways, e.g. row-wise, column-wise, in spiral shapes etc., which does not make a difference in the error-correction capability. Moreover, one can write the RS code itself (row-wise) into a $n_1 \times n_2$-matrix such that $n_1 n_2 = n$. If m is a square and we expand the symbols from $\mathbb{F}_{q^m}$ into squares of length $\sqrt{m}$, then one gets a q-ary array code of length $n_1\sqrt{m} \times n_2\sqrt{m}$.

Theorem 3.2. *A q-ary RS code with the parameters from above is able to correct*

- *any single square burst of area* $\leq \left(\sqrt{m}\left\lfloor\sqrt{\frac{n-k}{2}} - 1\right\rfloor + 1\right)^2$, *or*
- *any ℓ many square bursts of area* $\leq \left(\sqrt{m}\left\lfloor\sqrt{\frac{n-k}{2\ell}} - 1\right\rfloor + 1\right)^2$, *or*
- *any single one-dimensional burst of length* $\leq \sqrt{m}(\lfloor \frac{n-k}{2} \rfloor - 1) + 1$, *or*
- *any ℓ many one-dimensional bursts of length* $\leq \sqrt{m}(\lfloor \frac{n-k}{2\ell} \rfloor - 1) + 1$.

Proof. The one-dimensional burst error correction capability is the same as in the row vector expansion case, thus it remains to show the square burst case.

Let x denote the side length of the burst. The burst corrupts the most extension field symbols of the code when e.g. the entry of its upper left corner lies on the entry of the lower right corner of some expanded symbol. Then the burst corrupts at most

$$\left(\left\lceil \frac{x-1}{\sqrt{m}} \right\rceil + 1\right)^2$$

extension field symbols. Since the number of corrupted extension field elements has to be less than or equal to the error-correction capability of the RS code, it has to hold that

$$\left(\left\lceil\frac{x-1}{\sqrt{m}}\right\rceil+1\right)^2\leq\frac{n-k}{2}\iff x\leq\sqrt{m}\left\lfloor\sqrt{\frac{n-k}{2}}-1\right\rfloor+1$$

which implies the formula. In the case that there are ℓ square bursts it has to hold that

$$\ell\left(\left\lceil\frac{x-1}{\sqrt{m}}\right\rceil+1\right)^2\leq\frac{n-k}{2}\iff x\leq\sqrt{m}\left\lfloor\sqrt{\frac{n-k}{2\ell}}-1\right\rfloor+1.$$

□

Note, that in analogy to Construction I, one can again add a parity check element to all vector expansions of the extension field elements to achieve a larger minimum distance.

Another way of expanding the elements of $\mathbb{F}_{q^m}=\mathbb{F}_q[\alpha]$ is using companion matrices. For this let $p(x)\in\mathbb{F}_q[x]$ be the minimal polynomial of α and $P\in\mathbb{F}_q^{m\times m}$ its companion matrix. Then $\mathbb{F}_{q^m}\cong\mathbb{F}_q[P]$ and we can represent the extension field elements by q-ary matrices. If the RS code is written in an $n_1\times n_2$-matrix like above, then this results in a q-ary array code of length $n_1m\times n_2m$.

Theorem 3.3. *A q-ary RS code expanded with companion matrices with the parameters from above is able to correct*

- *any single square burst of area* $\leq\left(m\left\lfloor\sqrt{\frac{n-k}{2}}-1\right\rfloor+1\right)^2$*, or*
- *any ℓ many square bursts of area* $\leq\left(m\left\lfloor\sqrt{\frac{n-k}{2\ell}}-1\right\rfloor+1\right)^2$*, or*
- *any single one-dimensional burst of length* $\leq m(\lfloor\frac{n-k}{2}\rfloor-1)+1$*, or*
- *any ℓ many bursts of length* $\leq m(\lfloor\frac{n-k}{2\ell}\rfloor-1)+1$*.*

The proof is analogous to the one of Theorem 3.2 and hence omitted.

It can be noticed that this second expansion can offer a comparable burst error correction capability with respect to the other expansion (considering q-ary expansions of comparable length), but allowing to start with a RS code of smaller length and having lower decoding complexity. The rate though would be relatively smaller, which may reduce the security of the scheme.

4. Concatenated codes. When the error pattern is a mix of burst and random errors, concatenated codes provide the desired flexibility.

Definition 4.1. *Let* $\mathcal{C}_{\text{out}}\colon(\mathbb{F}_{q^k})^K\to(\mathbb{F}_{q^k})^N$ *and* $\mathcal{C}_{\text{in}}^{(i)}\colon\mathbb{F}_q^k\to\mathbb{F}_q^n$ *for* $i=1,\dots,N$ *be encoding functions. The* concatenated code $\mathcal{C}=\mathcal{C}_{\text{in}}^{(1)}\|\mathcal{C}_{\text{in}}^{(2)}\|\dots\|\mathcal{C}_{\text{in}}^{(N)}$ *is the q-ary code of length Nn and dimension Kk given by*

$$\mathcal{C}=\left\{(\mathcal{C}_{\text{in}}^{(1)}(y_1),\mathcal{C}_{\text{in}}^{(2)}(y_2),\dots,\mathcal{C}_{\text{in}}^{(N)}(y_N))\;\middle|\;(y_1,y_2,\dots,y_N)=\mathcal{C}_{\text{out}}(x),x\in(\mathbb{F}_{q^k})^K\right\}.$$

Here we assume the isomorphism $\mathbb{F}_{q^k}\cong\mathbb{F}_q^k$, so that the maps $\mathcal{C}_{\text{in}}$ agree in the domain. The codes $\mathcal{C}_{\text{in}}^{(1)},\dots\mathcal{C}_{\text{in}}^{(N)}$ are called the *inner codes* and $\mathcal{C}_{\text{out}}$ is called the *outer code.* The decoding process is a two-step process in which the inner codes are decoded first, followed by the outer code. Suppose that the inner codes are identical, that is

$$\mathcal{C}=\underbrace{C\|C\|\dots\|C}_{n\text{ times}},$$

and we can represent an element $c \in \mathcal{C}$ as

$$c = (c_{1,1}, \dots, c_{1,n} \| \dots \| c_{N,1}, \dots, c_{N,n}). \tag{4.1}$$

Assume that C and $\mathcal{C}_{\text{out}}$ have efficient decoding procedures. Suppose that the decoding procedure for C can correct t errors and the decoding procedure for $\mathcal{C}_{\text{out}}$ can correct s errors. If fewer than $t+1$ errors occur in the coordinates of the ith copy of C then the inner decoding step can correct these errors and the block would no longer count as an error in the outer code. If the number of errors is greater than t then the n symbols corresponding to the ith copy of C count as only one error to the outer decoder. One consequence of this multi-layer decoding process is that $\mathcal{C}$ can correct any burst of length smaller than $n(s-1)+2t+1$, beside the possibility of correcting other random errors. Alternatively, it is not difficult to see that $\mathcal{C}$ also has the capability to correct multiple burst errors, provided their lengths are not too long.

Construction IV. In the two-dimensional case, concatenated codes can be interleaved in ways that have advantages against different types of error patterns. Let $b \mid N$ and $a \mid n$. Arrange the elements of c in (4.1) in the $\frac{Na}{b} \times \frac{nb}{a}$ array given by

$$\left(\begin{array}{ccc|c|ccc}
c_{1,1} & \cdots & c_{b,1} & \cdots & c_{1,n/a} & \cdots & c_{b,n/a} \\
c_{b+1,1} & \cdots & c_{2b,1} & \cdots & c_{b+1,n/a} & \cdots & c_{2b,n/a} \\
 & \vdots & & \vdots & & \vdots & \\
c_{N-b+1,1} & \cdots & c_{N,1} & \cdots & c_{N-b+1,n/a} & \cdots & c_{N,n/a} \\
\hline
 & \vdots & & \ddots & & \vdots & \\
\hline
c_{1,\frac{(a-1)n}{a}+1} & \cdots & c_{b,\frac{(a-1)n}{a}+1} & \cdots & c_{1,n} & \cdots & c_{b,n} \\
c_{b+1,\frac{(a-1)n}{a}+1} & \cdots & c_{2b,\frac{(a-1)n}{a}+1} & \cdots & c_{b+1,n} & \cdots & c_{2b,n} \\
 & \vdots & & \vdots & \vdots & & \\
c_{N-b+1,\frac{(a-1)n}{a}+1} & \cdots & c_{N,\frac{(a-1)n}{a}+1} & \cdots & c_{N-b+1,n} & \cdots & c_{N,n}
\end{array}\right).$$

Using this interleaving pattern, any submatrix of size $\frac{N}{b} \times b$ contains only one symbol from each inner code. It is straightforward to see that this interleaving scheme can correct at least t rectangular bursts of size $\frac{N}{b} \times b$, since it can correct t such bursts using the inner codes alone. Additionally, this interleaving pattern can correct at least s random errors.

Construction V. Without interleaving, one could consider the following pattern. Let $a \mid N$ and $b \mid n$. Arrange the elements of c in (4.1) in the $\frac{Nn}{ab} \times ab$ array given by

$$\left(\begin{array}{ccc|c|ccc}
c_{1,1} & \cdots & c_{1,b} & \cdots & c_{a,1} & \cdots & c_{a,b} \\
c_{1,b+1} & \cdots & c_{1,2b} & \cdots & c_{a,b+1} & \cdots & c_{a,2b} \\
 & \vdots & & \vdots & & \vdots & \\
c_{1,n-b+1} & \cdots & c_{1,n} & \cdots & c_{a,n-b+1} & \cdots & c_{a,n} \\
\hline
 & \vdots & & \ddots & & \vdots & \\
\hline
c_{N-a+1,1} & \cdots & c_{N-a+1,b} & \cdots & c_{N,1} & \cdots & c_{N,b} \\
c_{N-a+1,b+1} & \cdots & c_{N-a+1,2b} & \cdots & c_{N,b+1} & \cdots & c_{N,2b} \\
 & \vdots & & \vdots & \vdots & & \\
c_{N-a+1,n-b+1} & \cdots & c_{N-a+1,n} & \cdots & c_{N,n-b+1} & \cdots & c_{N,n}
\end{array}\right).$$

Using this pattern, any single burst of maximal size $((s_1-1)\frac{n}{b}+1)\times((s_2-1)b+1)$ can be corrected by the outer code, where $s_1, s_2 \in \mathbb{N}$ with $s_1 s_2 \leq s$, and as long as each submatrix of size $\frac{n}{b} \times b$ which does not collide with a burst contains at most t errors, all errors can be corrected. Therefore, this scheme works well if there is at most one larger burst, or few smaller ones, and a lot of random errors spread more or less uniformly in the matrix.

Note that Constructions II and III can be seen as special cases of Construction V, where the inner code is the trivial code with $b = \sqrt{m}$ for Construction II, and with the encoding $\mathbb{F}_{q^m} \cong \mathbb{F}_q[P] \subseteq \mathbb{F}_q^{m\times m}$ and $b = m$ for Construction III.

Construction VI. Consider another interleaving pattern given by the $n \times N$ array (say $N \geq n$):

$$\begin{pmatrix} c_{1,1} & c_{2,1} & c_{3,1} & c_{4,1} & \cdots & c_{N,1} \\ c_{N,2} & c_{1,2} & c_{2,2} & c_{3,2} & \cdots & c_{N-1,2} \\ & & \vdots & & & \\ c_{N-n+2,n} & c_{N-n+3,n} & c_{N-n+4,n} & c_{N-n+5,n} & \cdots & c_{N-n+1,n} \end{pmatrix}$$

Each inner code is interleaved diagonally, and it is clear from the construction that any burst of size $1 \times n$ or $n \times 1$ corrupts only one symbol from each inner code. Additionally, an error pattern consisting of a diagonal burst would corrupt an entire inner codeword but be treated as a single error to the outer decoder.

Using a fuzzy scheme in which the witness is prone to burst errors as well as random errors, an interleaving scheme like that in Construction IV may be advantageous. In case less burst errors happen, but more random errors, a scheme like the ones in Construction V or VI might be better suited.

5. Conclusions.

One of the big advantages of using burst error correcting codes relies in the decoding complexity, since we are correcting errors in words of a certain length but using decoding procedures for codes of smaller lengths. For example in Constructions I, II and III the decoding complexity is dominated by the Reed-Solomon decoder in the extension field. Using the standard Gorenstein-Peterson-Zierler decoding procedure this is given by

$$O\Big(n \cdot \frac{n-k}{2}\Big) = O\Big((q^m-1)\frac{q^m-1-k}{2}\Big) = O(q^{2m})$$

operations in $\mathbb{F}_{q^m}$. More recent algorithms for decoding cyclic codes [8] can even achieve a complexity of

$$O\Big(\sqrt{n}\log n \cdot \frac{n-k}{2}\Big) = O(mq^{\frac{3}{2}m}).$$

Note, that in Construction III the conversion of the elements of $\mathbb{F}_q[P]$ to elements of $\mathbb{F}_q[\alpha]$ is more complex then the conversion of elements of $\mathbb{F}_q^m$ to elements of $\mathbb{F}_{q^m}$ needed in Constructions I and II, but the complexity is still polynomial in m and thus does not change the overall complexity in the Big-O notation.

The decoding complexity for Constructions IV, V and VI depends on the choices of the inner and outer codes. If for example you choose a RS-code over $\mathbb{F}_{q^k}$ for the outer code and a q-ary narrow-sense primitive BCH-code of length $n = q^m - 1$ for the inner code, then the decoding complexity is $O(Nnt) = O((q^k-1)(q^m-1)t) = O(q^{k+m}t)$ operations in $\mathbb{F}_{q^m}$, where t is the error-correction capability of the inner code, and $O(q^{2k})$ operations in $\mathbb{F}_{q^k}$.

Depending on the application, any of the six presented constructions can be advantageous for burst error correction in the syndrome fuzzy hashing scenario. The differences appear in the type and number of correctable errors as well as in the decoding complexity for given code word size in the q-ary expansion. Naturally, this list of constructions is not complete and there exist other codes that can be useful for burst-error correction in the storage of noisy data.

We want to conclude with a list of recommendations for which situations the above constructions can be used:

- Constructions I and II are relatively fast to decode (as the inner code is trivial), while having a good burst error correction capability. They can also correct a few random errors, but not too many.
- Construction III can correct bursts similarly well as Construction II, and compensates a relatively lower rate with a faster decoding procedure, which can be relevant when deploying the syndrome fuzzy hashing scheme on embedded devices.
- Construction IV works well with several and larger bursts, but not too many random errors.
- Construction V works well with a few large bursts and many uniformly enough distributed random errors.
- Construction VI is well suited for rectangular bursts of size $k \times \ell$ with k much smaller than ℓ (or vice versa), and for random errors.

REFERENCES

[1] M. Baldi, M. Bianchi, F. Chiaraluce, J. Rosenthal, D. Schipani. On fuzzy syndrome hashing with LDPC coding. In *Proc. 4th ACM International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL)*, pp. 1-5, Barcelona, 2011.

[2] I. Buhan, J. Doumen, P. Hartel. Controlling leakage of biometric information using dithering. In *Proc. EUSIPCO 2008*, Lausanne, 2008.

[3] Y. Dodis, R. Ostrovsky, L. Reyzin, A. Smith. Fuzzy extractors: how to generate strong keys from biometrics and other noisy data. *SIAM J. Comp.*, vol. 38, no. 1, pp. 97–139, 2008.

[4] Y. Dodis, A. Smith. Correcting errors without leaking partial information. In *Proc. ACM STOC 2005*, pp. 654–663, Baltimore MD, 2005.

[5] A. Juels, M. Sudan. A fuzzy vault scheme. *Designs, Codes and Cryptography*, vol. 38, no. 2, pp. 237–257, 2006.

[6] A. Juels, M. Wattenberg. A fuzzy commitment scheme. In *Proc. 6th ACM Conference on Computer and Communications Security*, pp. 28–36, Singapore, 1999.

[7] F.J. MacWilliams, N.J.A. Sloane. The Theory of Error-Correcting Codes. North Holland, 1977.

[8] D. Schipani, M. Elia, J. Rosenthal. On the Decoding Complexity of Cyclic Codes up to the BCH Bound. In *Proc. IEEE Int. Symp. on Information Theory*, pp.835–839, St. Petersburg, 2011.

[9] D. Schipani, J. Rosenthal. Coding solutions for the secure biometric storage problem. In *Proc. IEEE Information Theory Workshop 2010*, Dublin, 2010.

[10] P. Tuyls, B. Skoric, T. Kevenaar (Eds.). *Security with noisy data.* Springer, 2007.

[11] U. Uludag, S. Pankanti, S. Prabhakar, A.K. Jain. Biometric cryptosystems: Issues and challenges. In *Proc. of the IEEE*, vol. 92, no. 6, pp. 948–960, 2004.