A Nonlinear Multivariate Cryptosystem Based on a Random Linear Code

Daniel Smith-Tone[1,2] and Cristina Tone[1]

[1]Department of Mathematics, University of Louisville,
Louisville, Kentucky, USA
[2]National Institute of Standards and Technology,
Gaithersburg, Maryland, USA

daniel.smith@nist.gov, cristina.tone@louisville.edu

Abstract. We introduce a new technique for building multivariate cryptosystems based on random linear codes. The security of these schemes depends on some of the standard security assumptions in multivariate cryptography as well as some new problems.

Key words: Multivariate Cryptography, k-ary C^*, differential attack

1 Introduction

In the mid 1990s Peter Shor broke the schemes that we currently use for information security in the public key setting, see [1]. If we accept that the construction of the technology to implement his attacks is an engineering challenge as opposed to a physical impossibility, then we admit that out current public key infrastructure is a paper tiger waiting to be crushed.

Since that time, several communities have emerged devoted to various promising avenues to security in a post-quantum world, that is, a world with the large scale quantum computing devices required to undermine current public key cryptography by Shor's techniques. We can largely place these communities in four classes: code-based, isogeny-based, lattice-based and multivariate.

These families are all disparate, though there are sometimes some similarities between code-based and lattice-based techniques. Isogeny-based and multivariate cryptosystems, however, are usually viewed as being far removed from the code-based and lattice-based camps.

An interesting, though impractical, scheme was presented at PKC 2012, see [2], which hacked a lattice technique for use as a multivariate cryptosystem. The main idea is to separate a multivariate quadratic system of formulae into a linear part L and a quadratic part Q playing the roles of the matrix $\mathbf{A}$ and the error distribution χ, respectively, in standard LWE, see [3]. The coefficients of L are very large, whereas the coefficients of Q are very small. When a small input $\mathbf{x}$ is introduced, a small vector $Q(\mathbf{x})$ is sampled and the "lattice point" $L(\mathbf{x})$ is perturbed. As long as the parameters are quite large, and under some additional assumptions, the distribution of $(L, Q(\mathbf{x}) + L(\mathbf{x}))$ is close to that of $(L, L(\mathbf{s}) + \mathbf{e})$ where $\mathbf{e}$ is drawn from an appropriate Gaussian distribution, so that the security of the scheme is based on the LWE assumption and the MQ problem, that is, the problem of solving quadratic systems of equations over a field.

A natural question to ask is whether it is possible to breed a hybrid code-based multivariate scheme and what properties is might possess. In this work, we present a new multivariate encryption scheme inspired and derived from linear codes. While the connection to code-based schemes is not so direct and apparent as the connection to LWE in [2], the construction appears versatile and amenable to adjustment for various security and performance properties as have multivariate schemes in general come to be known. As an example of this, we propose, in addition to the fundamental scheme, a variant with a decryption algorithm approximately 6000 times faster than the original, and, in fact, much faster than any multivariate encryption scheme targeting CCA security at the 128-bit security level, see Table 1 for a comparison with currently credibly secure multivariate encryption schemes including Simple Matrix, Extension Field Cancellation, HFERP and EFLASH, see [4–7].

Table 1. Performance characteristics of multivariate encryption schemes at the best available comparison to the 128-bit security level.

Scheme	Sec.	PK size	Enc.(ms)	Dec.(ms)	Fail Rate
EFLASH(2,134,159,9)	128	165.7KB	1.3	12758	2^{-32}
EFC_{pt^2}(2,83,8)	80*	523KB	4	2481	negl.
EFC_{pt^2}(2,148,8)§	128	3059KB	23‡	∼14000‡	negl.
ABC(2^8,10,13,14,14,364,180)	100†	5537KB	53	59	2^{-32}
ABC(2^8,12,15,16,16,484,240)§	128	13556KB	126‡	147‡	2^{-32}
HFERP(85,70,89,61,3^7+1)	128	1344KB	6	49182	negl.
CBM-CPA(148,132,476)	128	818.3KB	9.1	414168	2^{-16}
PCBM-CCA(148,149,133,475)	128	818.4KB	9.1	423222	negl.
ECBM-CPA(148,131,23,5,298)	128	569.7KB	4.5	66.3	2^{-16}
ECBM-CCA(148,83,71,5,160)	128	327.1KB	4.5	160	2^{-64}
EPCBM-CCA(148,149,132,23,5,297)	128	578.8KB	4.5	67.1	negl.

This manuscript is organized as follows. In Section 2 we present the framework for the new scheme. Then, in Section 3 we examine the decryption failure rate and set constraints on parameters to satisfy reasonable bounds. We then conduct a security analysis against the known attack vectors in Section 4. In Section 5 we introduce a modification dramatically improving performance, both in decryption time and in key size. We then present some concrete parameters for

* The published parameters achieve around 73 bits of security, see [8].
† The published parameters achieve around 98 bits of security, see [9].
§ Extrapolated performance parameters to achieve 128-bit security.
‡ Performance is estimated via scaling.

future scrutiny in Section 6. Finally, we conclude, discussing future directions for this line of reasoning.

2 Nonlinear Multivariate System from a Linear Code

Let $\mathbb{F}_q$ be a finite field with q elements and let C be a rank k random linear code of length n over $\mathbb{F}_q$. Let $\mathbf{G}$ be the generator matrix for C in standard form and let $\mathbf{H}$ be the corresponding parity check matrix.

We construct a quadratic system of formulae as follows. First, randomly select k matrices $\mathbf{A}_i$ in $\mathcal{M}_{n\times(n-k)}(\mathbb{F}_q)$. Next form the products $\mathbf{B}_i = \mathbf{A}_i\mathbf{H}$. Finally, let $F : \mathbb{F}_q^n \to \mathbb{F}_q^k$ be defined by $F(\mathbf{x}) = (F_1(\mathbf{x}), \ldots, F_k(\mathbf{x}))$, where $F_i(\mathbf{x})$ is given by

$$\mathbf{x}\mathbf{B}_i\mathbf{x}^\top.$$

Given knowledge of the code C, preimages under F may be acquired by computing a set of representatives $\mathcal{A}$ of the cosets of C in $\mathbb{F}_q^n$, and linearly solving for a preimage in each coset. Specifically, note that if $\mathbf{y} = F(\mathbf{x})$, then there exists an $\mathbf{x}' \in \mathcal{A}$ and an $\widehat{\mathbf{x}} \in C$ such that $\mathbf{x} = \mathbf{x}' + \widehat{\mathbf{x}}$; moreover, we note that since $\widehat{\mathbf{x}} = \overline{\mathbf{x}}\mathbf{G}$ for some $\overline{\mathbf{x}} \in \mathbb{F}_q^k$, that

$$\begin{aligned}
y_\ell &= (\mathbf{x}' + \widehat{\mathbf{x}})\mathbf{B}_\ell(\mathbf{x}'^\top + \widehat{\mathbf{x}}^\top) \\
&= \mathbf{x}'\mathbf{B}_\ell\mathbf{x}'^\top + \mathbf{x}'\mathbf{B}_\ell\widehat{\mathbf{x}}^\top + \widehat{\mathbf{x}}\mathbf{B}_\ell\mathbf{x}'^\top + \widehat{\mathbf{x}}\mathbf{B}_\ell\widehat{\mathbf{x}}^\top \\
&= \mathbf{x}'\mathbf{B}_\ell\mathbf{x}'^\top + \mathbf{x}'\mathbf{A}_\ell\mathbf{H}\mathbf{G}^\top\overline{\mathbf{x}}^\top + \widehat{\mathbf{x}}\mathbf{B}_\ell\mathbf{x}'^\top + \widehat{\mathbf{x}}\mathbf{A}_\ell\mathbf{H}\mathbf{G}^\top\overline{\mathbf{x}}^\top \\
&= \mathbf{x}'\mathbf{B}_\ell\mathbf{x}'^\top + \widehat{\mathbf{x}}\mathbf{B}_\ell\mathbf{x}'^\top \\
&= \mathbf{x}'\mathbf{B}_\ell\mathbf{x}'^\top + \overline{\mathbf{x}}\mathbf{G}\mathbf{B}_\ell\mathbf{x}'^\top,
\end{aligned}$$

for $1 \le \ell \le k$, form k linear equations in the k unknown coefficients of $\overline{\mathbf{x}}$.

We further note a few simple facts. Efficient derivation of preimages of F requires that $n-k$ be small. Then, necessarily, the matrices $\mathbf{B}_\ell$, which are of rank $n-k$ at most, are of low rank. Given merely the multivariate representation of F_ℓ, however, an adversary does not immediately recover a low rank representation of F_ℓ as a quadratic form. In general, there are around $q^{\binom{n}{2}}$ matrix representations of F_ℓ, many of which have high rank.

Still, the code structure of C can be learned from F in this form by simply searching for roots of the system. Since any code word $\mathbf{x} \in C$ satisfies $F(\mathbf{x}) = \mathbf{0}$, one simply searches with complexity roughly $\mathcal{O}(kq^{n-k})$ for k roots of F which generate a k-dimensional subspace of roots of F and C is recovered. To prevent this attack, we use the plus (+) modifier, adding p additional random formulae to F. These additional formulae are then mixed via an affine transformation T with the k formulae derived from C producing the public key of the code-based multivariate cryptosystem (CBM):

$$P = T \circ (F \| Q),$$

where $Q : \mathbb{F}_q^n \to \mathbb{F}_q^p$ is a random quadratic map and $T : \mathbb{F}_q^{k+p} \to \mathbb{F}_q^{k+p}$ is an invertible linear map. Preimages of P are then calculated by inverting T and following the above procedure for finding a preimage of F.

Including the (+) modifier, the extra equations generated by Q must be satisfied as well, so we may need to check all values of $\mathbf{x}' \in \mathcal{A}$ to find a valid preimage. When p is not much larger than k, we expect in general that the preimage may not be a singleton. Thus, we must make p considerably larger than k so that the equations from Q provide check equations that a single correct input has been found. We can therefore find parameters for which this scheme can be instantiated for public key encryption. For sufficiently large p, the system is statistically injective, in the sense that the probability of selecting an input producing a non-unique output is negligible in n.

3 Decryption Failure Rate

The hidden map F from Section 2 deviates significantly from a random function in that there is a large dimensional subspace on which it is identically zero. This property is not the only manner in which F behaves differently.

One would expect a random function from $\mathbb{F}_q^n$ to $\mathbb{F}_q^k$ to collide in every value approximately q^{n-k} times; moreover, one would expect the distribution of multiplicities for each output to be centered at q^{n-k}. The value, $n-k$ is small by design, however, and the output $\mathbf{0}$ occurs at least q^k times. Thus, the distribution of multiplicities of the outputs must be skewed towards lower values while having a single large value around q^k. We can say somewhat more.

Aside from codewords, there is a higher probability for a collision on the outputs of two elements in the same coset of the code. Recall that given a representative $\mathbf{x}'$ of the coset $\mathbf{x}' + C$, that $F_\ell(\mathbf{x}' + \overline{\mathbf{x}}\mathbf{G}) = \mathbf{x}'\mathbf{B}_\ell\mathbf{x}'^\top + \overline{\mathbf{x}}\mathbf{G}\mathbf{B}_\ell\mathbf{x}'^\top$. Thus, there is a collision $F(\mathbf{x}') = F(\mathbf{x}' + \overline{\mathbf{x}}\mathbf{G})$ if the $k \times k$ matrix

$$\begin{bmatrix}\mathbf{G}\mathbf{B}_1\mathbf{x}'^\top & \cdots & \mathbf{G}\mathbf{B}_k\mathbf{x}'^\top\end{bmatrix}$$

is singular. In particular, there are as many elements in the coset with the same output as $\mathbf{x}'$ as the size of the kernel of this map. Since a $k \times k$ matrix is singular with probability approximately q^{-1} for sufficiently large k and, very roughly, the kernel is of dimension r with probability about q^{-r}, the distribution of multiplicities of outputs is large near zero and decays exponentially in q, with the single exception of a very large multiplicity output of $\mathbf{0}$. We experimentally verified this analysis. The results of a particular example can be found in Table 2.

Table 2. The frequency of multiplicities of outputs of the hidden map F— that is, the number of outputs whose preimage under F is of a given size— for an instance of F with parameters $q = 2$, $n = 12$, $k = 10$, and $p = 16$.

Multiplicity	0	2	4	6	8	1030
Frequency	192	384	256	127	64	1

With this observation on the distribution of multiplicities we can estimate a collision probability for P under the standard heuristic that random quadratic functions behave as random oracles and the additional assumption that failures are dominated by multiple preimages in C. For clarity of notation, in the following allow $\#A$ to represent the size of the set A. A collision occurs when the size of the preimage of a valid ciphertext under P is greater than one. Thus, the probability of collision is given by

$$p_{\text{fail}} = Pr\left[\#P^{-1}(\mathbf{y}) > 1 \quad \#P^{-1}(\mathbf{y}) > 0\right] = \frac{Pr\left[\#P^{-1}(\mathbf{y}) > 1\right]}{Pr\left[\#P^{-1}(\mathbf{y}) > 0\right]}.$$

Clearly, since $P = T \circ (F\|Q)$ and T is invertible, we have equivalently,

$$\begin{aligned} p_{\text{fail}} &= \frac{Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2)) > 1\right]}{Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2)) > 0\right]} \\ &= \frac{1 - Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2)) = 1\right] - Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2)) = 0\right]}{1 - Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2)) = 0\right]}, \end{aligned}$$

where $\mathbf{y} = \mathbf{y}_1\|\mathbf{y}_2$.

Considering the above observation about the special status of $\mathbf{y}_1 = \mathbf{0}$, we consider the two probabilities in the above numerator, splitting into the cases $\mathbf{y}_1 = \mathbf{0}$ and $\mathbf{y}_1 \neq \mathbf{0}$. Since we know that $F(C) = \mathbf{0}$ and $|C| = q^k$, we model the random variable $\#F^{-1}(\mathbf{y}_1)$ as a Binomial$(q^n - q^k, q^{-k})$. Similarly, since q^k values in $F^{-1}(\mathbf{0})$ are not random, we model $\#F^{-1}(\mathbf{0})$ as $q^k + X$ where $X \sim$ Binomial$(q^n - q^k, q^{-k})$. We will require the following Lemma related to binomial random variables.

Lemma 1

$$\sum_{k=0}^{n} k\binom{n}{k}(rp)^k(1-p)^{n-k} = npr(1 + (r-1)p)^{n-1}.$$

Proof. Trivial.

First, we consider the probability of no intersection.

$$\begin{aligned} &Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2)) = 0\right] \\ &\qquad = Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2)) = 0 \quad \mathbf{y}_1 = \mathbf{0}\right] Pr\left[\mathbf{y}_1 = \mathbf{0}\right] \\ &\qquad + Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2)) = 0 \quad \mathbf{y}_1 \neq \mathbf{0}\right] Pr\left[\mathbf{y}_1 \neq \mathbf{0}\right] \end{aligned}$$

We expand this expression by splitting the events into disjoint unions based on the value of $\#F^{-1}(\mathbf{y}_1)$.

$$\begin{aligned} &\sum_{s=0}^{q^n-q^k} Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2)) = 0 \wedge \#F^{-1}(\mathbf{y}_1) = s + q^k \quad \mathbf{y}_1 = \mathbf{0}\right] Pr\left[\mathbf{y}_1 = \mathbf{0}\right] \\ &+ \sum_{s=0}^{q^n-q^k} Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2)) = 0 \wedge \#F^{-1}(\mathbf{y}_1) = s \quad \mathbf{y}_1 \neq \mathbf{0}\right] Pr\left[\mathbf{y}_1 \neq \mathbf{0}\right] \end{aligned}$$

Let $Q_{\mathbf{y}_1}$ represent the function Q restricted to $F^{-1}(\mathbf{y}_1)$. Then we may simplify the notation obtaining:

$$\sum_{s=0}^{q^n-q^k} Pr\left[\#Q_{\mathbf{y}_1}^{-1}(\mathbf{y}_2)=0 \wedge \#F^{-1}(\mathbf{y}_1)=s+q^k \quad \mathbf{y}_1=\mathbf{0}\right] Pr\left[\mathbf{y}_1=\mathbf{0}\right]$$
$$+\sum_{s=0}^{q^n-q^k} Pr\left[\#Q_{\mathbf{y}_1}^{-1}(\mathbf{y}_2)=0 \wedge \#F^{-1}(\mathbf{y}_1)=s \quad \mathbf{y}_1\neq\mathbf{0}\right] Pr\left[\mathbf{y}_1\neq\mathbf{0}\right]$$

Under the assumption that Q acts as a random oracle, independence is maintained even with a restricted domain. Therefore, we obtain:

$$\sum_{s=0}^{q^n-q^k} Pr\left[\#Q_{\mathbf{y}_1}^{-1}(\mathbf{y}_2)=0\right] Pr\left[\#F^{-1}(\mathbf{y}_1)=s+q^k \quad \mathbf{y}_1=\mathbf{0}\right] Pr\left[\mathbf{y}_1=\mathbf{0}\right]$$
$$+\sum_{s=0}^{q^n-q^k} Pr\left[\#Q_{\mathbf{y}_1}^{-1}(\mathbf{y}_2)=0\right] Pr\left[\#F^{-1}(\mathbf{y}_1)=s \quad \mathbf{y}_1\neq\mathbf{0}\right] Pr\left[\mathbf{y}_1\neq\mathbf{0}\right]$$

Each of these probabilities is now readily computed. In the case that $\mathbf{y}_1=\mathbf{0}$, $\#F^{-1}(\mathbf{y}_1)-q^k$ is binomial; otherwise, $\#F^{-1}(\mathbf{y}_1)$ is binomial. Since the probability that a random input to Q produces $\mathbf{y}_2$ is q^{-p}, the probability that none of t outputs is equal to $\mathbf{y}_2$ is $(1-q^{-p})^t$ for either $t=s$ or $t=s+q^k$. Thus we have:

$$\sum_{s=0}^{q^n-q^k}(1-q^{-p})^{s+q^k}\binom{q^n-q^k}{s}q^{-ks}(1-q^{-k})^{q^n-q^k-s}q^{-k}$$
$$+\sum_{s=0}^{q^n-q^k}(1-q^{-p})^{s}\binom{q^n-q^k}{s}q^{-ks}(1-q^{-k})^{q^n-q^k-s}(1-q^{-k})$$
$$=\left(1-q^{-k}+q^{-k}(1-q^{-p})^{q^k}\right)\left(1-q^{-k-p}\right)^{q^n-q^k}.$$

A similar process for $Pr[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2))=1]$ produces expressions in the form of Lemma 1 with q^n-q^k in place of n, $1-q^{-p}$ in place of r and q^{-k} in place of p. Simplifying the massive expression we obtain:

$$Pr\left[\#(F^{-1}(\mathbf{y}_1)\bigcap Q^{-1}(\mathbf{y}_2))=1\right]$$
$$=(q^{-k-p}*(1-q^{-p})^{q^k-1}+q^{-p}*(1-q^{-k}))*(q^{n-k}-1)*(1-q^{-k-p})^{q^n-q^k-1}$$
$$+q^{-p}*(1-q^{-p})^{q^k-1}*(1-q^{-k-p})^{q^n-q^k}.$$

With some tedious but trivial manipulation, we can show that the resulting expression for $lg(p_{fail})$ is approximately equal to $q^{2k-n-p-1}$. Thus we have proven the following theorem.

Theorem 1 *Under the heuristic that* $F_{\overline{C}}$ *and* Q *are random functions, the collision probability for the CBM satisfies the bound*

$$p_{fail} < q^{2k-n-p}.$$

Furthermore, if

$$\lim_{n\to\infty} \frac{p}{n} > 1,$$

then p_{fail} *is negligible in* n.

We performed some small scale experiments which agree with the above probability to within a factor of $q = 2$ as k and n increase, suggesting that the heuristic of Theorem 1 is sufficiently close to reality to be meainingful. A range of values of k, n and p exhibiting a transition between loose and tight approximation by the above estimate are presented in Table 3.

In addition to possible lack of injectivity, there is another issue affecting CMB. If the plaintext happens to be a code-word, then the decryption method fails to provide any linear relations. In this case, inversion can still be achieved with high probability at the cost of searching through the code for the appropriate preimage. Such searches are performed in the experiment presented in Table 3. For practical parameters, there are q^k codewords and k must be quite large; thus, inversion in this case is infeasible. Therefore, when we consider decryption failures, we find that the failure rate is the larger between q^{k-n} and q^{2k-n-p}. Thus, there is a phase transition for CBM around $p = k$. For plain CBM the decryption failure rate is dominated by q^{k-n}, but for the fairly rich space of possible variants, it is possible for this transition to take place.

4 Security Analysis

Attacks on multivariate cryptosystems largely fall in to a few categories: algebraic, rank, differential, statistical and ad hoc. We analyze the scheme presented in Section 2 with respect to these categories.

4.1 Algebraic Attack

The most fundamental attack in multivariate cryptography is the algebraic attack, that is, directly solving the system of equations $\mathbf{y} = P(\mathbf{x})$. The complexity of this method is determined by the size of the Macaulay matrix at the solving degree. In practice, coincidence of the solving degree and the first fall degree in Gröbner basis calculations is sufficiently common that we conservatively assume that they are equal and select parameters for which the first fall degree is sufficiently large to guarantee security from this attack.

Following the analysis of [6], we calculate the semi-regular degree, i.e. the first fall degree assuming that as few relations as possible exist among the polynomials at each degree, as the first non-positive coefficient in the series expansion of

$$S_{n,m}(t) = \frac{(1-t^q)^n(1-t^2)^m}{(1-t)^n(1-t^{2q})^m}.$$

Table 3. The log decryption failure rate for small scale variants of the scheme. Values are computed by encrypting all possible plaintexts and counting the number of plaintexts that cannot be uniquely decrypted. All experiments use the value $q = 2$.

$n = 13$

$k = 8$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−7.093	−7.830	−8.415	−9.000	−12.000	−11.000
$k = 9$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−6.625	−7.415	−8.300	−9.193	−9.415	−11.000
$k = 10$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−5.238	−6.193	−6.715	−8.046	−9.000	−9.000
$k = 11$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−3.294	−4.206	−5.212	−6.057	−7.219	−7.913

$n = 14$

$k = 9$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−6.245	−7.557	−8.415	−9.193	−10.415	−11.000
$k = 10$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−5.950	−6.591	−7.715	−8.678	−10.000	−11.415
$k = 11$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−4.378	−5.081	−5.902	−6.810	−8.000	−9.046
$k = 12$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−2.623	−3.381	−4.142	−5.090	−6.000	−6.830

$n = 15$

$k = 10$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−6.102	−7.006	−7.660	−9.193	−10.830	−11.415
$k = 11$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−5.128	−5.967	−6.923	−7.956	−9.046	−10.300
$k = 12$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−3.621	−4.323	−5.127	−6.145	−7.099	−8.023
$k = 13$	p	12	13	14	15	16	17
	$lg(p_{fail})$	−2.207	−2.653	−3.358	−4.218	−5.155	−5.961

As seen in Section 3, to keep the decryption failure rate low we require that $2k - n - p$ is small. If we set this quantity to at most -64, we obtain $p \geq 2k - n + 64$. If we further assume that the MinRank is $r = 2(n - k)$ and is fixed at some value, then following the analysis in [10], we have an upper bound on the degree of regularity of the minimum of r and the semi-regular degree.

For practical values of n it is easy to find values of k for which the semi-regular degree is bounded by r and for which the above formula holds for p. Thus, under the assumption that the scheme is semi-regular, the complexity of the algebraic attack is $\mathcal{O}\left(\binom{n+r}{r}^{\omega}\right)$, where $2 \leq \omega < 3$.

We ran a series of experiments on small-scale variants to compare their behaviour to that of semi-regular schemes. For these experiments we chose to keep $n - k$ as close as reasonably possible to 10 to ensure that the MinRank is sufficiently high to model the behaviour of larger schemes. To study larger degrees of regularity in comparison to the semi-regular degree, we chose a small p, satisfying the formula $p = 2n - k - 8$. We also chose to examine parameter sets at the boundary of different semi-regular degrees to verify that these systems of equations really behave as generic systems. We found that in all trials the observed degree of regularity always exactly matched the semi-regular degree. The data are presented in Table 4.

Table 4. Degree of Regularity d_{reg} for small schemes at the transitions points of semi-regular degree d_{sr} with k as close as possible to $n - 10$ such that the scheme is not degenerate. Ten experiments are conducted for each parameter set, all having the same results.

(n, k, p)	$(10, 2, 10)$	$(11, 3, 11)$	$(23, 13, 25)$	$(24, 14, 26)$	$(36, 26, 38)$	$(37, 27, 39)$
d_{reg}	3	4	4	5	5	6
d_{sr}	3	4	4	5	5	6

4.2 Rank Attacks

Since for each coordinate of F there exists a matrix representation of rank $n-k$, we may suspect that there is a relevant rank attack on the scheme separating F from Q. There are a couple of systematic forms we must consider to analyze rank attacks. One such representation, the upper triangular representation, seems to offer no weakness. Even though the non-standard matrix has rank $n - k$ which may be low, the upper triangular form in general has much larger rank, see Table 5. On the other hand, if we ignore diagonal entries, symmetric representations have a rank bound of $2(n - k)$ since we can construct such a symmetric representation by adding the non-standard representation of rank $n - k$ with its transpose.

Given a low rank linear combination α of the symmetric forms ignoring the diagonal elements, one can take the corresponding linear combination $\mathbf{T}_\alpha$ of the upper triangular representations representing a function in the span of the F_i. Once recovered, there is an efficient way to expose the code C, undermining the scheme. One can randomly select $\mathcal{O}(kq^{n-k})$ vectors and likely find k generators $\mathbf{c}_i$ of C by testing whether linear combinations of roots $\mathbf{c}_i\mathbf{T}_\alpha\mathbf{c}_i^\top = 0$ are also roots. We note explicitly that the secret non-standard matrix representations of F_i share the same kernel, but there is no need for the systematic representations to share this property.

Under the assumption that the rank of the systematic matrix representations of F_i is no more than $2(n-k)$ and given $p+1$ systematic matrix representations of public polynomials, we are guaranteed that there is a linear combination eliminating the p plus polynomials. Therefore, we may simply select $p+1$ of the public symmetric forms and run a MinRank attack obtaining a solution. Using the standard "linear algebra search" technique of solving MinRank, one obtains

Table 5. MinRank for some small example instances of the code-based multivariate scheme. In each instance, there exists a non-standard matrix representation of a linear combination of the public quadratic forms of rank $n-k=2$; however, in each case the MinRank achieved is larger. In each case, $q=2$ and the systematic form used for the MinRank is the upper-triangular representation.

n ($k=n-2$, $p=n+4$)	10	11	12	13	14	15
MinRank	4	5	5	7	7	8

a complexity of

$$\mathcal{O}\left((p+1)^{\omega} q^{2\lceil\frac{p+1}{n}\rceil(n-k)}\right).$$

4.3 Differential Attacks

Consider the differential $DP(\mathbf{a},\mathbf{x}) = P(\mathbf{a}+\mathbf{x}) - P(\mathbf{a}) - P(\mathbf{x}) + P(\mathbf{0})$ of the public key P. We may expand this quantity as follows:

$$\begin{aligned} DP(\mathbf{a},\mathbf{x}) &= D(T\circ(F\|Q))(\mathbf{a},\mathbf{x}) \\ &= T(DF\|DQ)(\mathbf{a},\mathbf{x}). \end{aligned}$$

The special structure of DF implies that P has a subspace differential invariant, see [9, Definition 2]. Specifically, suppose that $\mathbf{M}$ is a linear projection onto C. Then we obtain

$$DF(\mathbf{Ma},\mathbf{Mx}) = F(\mathbf{c_a}+\mathbf{c_x}) - F(\mathbf{c_a}) - F(\mathbf{c_x}) + F(\mathbf{0}) = \mathbf{0}.$$

As noted in the previous subsection, since the systematic matrix forms of F_i are not necessarily of low rank, it is inefficient to recover the differential invariant from rank techniques. The alternative, however, of modeling the differential invariant as a cubic system of equations in the unknown coefficients of M and T^{-1} is no better, even though there are several dependencies in the system. Finding such an M in this way requires solving kn^2 cubic equations in $kn+km$ variables, which is much more complex than the brute force attack.

4.4 Statistical and ad hoc Attacks

As shown in, [11, 12], statistical cryptanalysis techniques in multivariate cryptography can be quite varied and can possibly allow for hybridized statistical/algebraic attacks. Security against all such attacks will be an ongoing research topic in this area in general.

To address the question of whether there are any straightforward and effective statistical attacks, we analyze the code-based multivariate scheme in two ways. First, we analyze the difference in distribution between $P(C)$ and $P(\overline{C})$, that is, ciphertexts derived from codeword and non-codeword plaintexts, respectively. Second, we examine the difference in Hamming weight between ciphertexts and random vectors.

To compare the distributions $P(C)$ and $P(\overline{C})$ we chose to select a statistic sensitive to any change in distribution between two empirical distributions, the Kolmogorov-Smirnov statistic. We select subsets X_1 and X_2 of C and $\overline{C}$, respectively, of the same size and compute

$$KS_N = \sup_{x \in \mathbb{F}_q^m} F_{1,N}(x) - F_{2,N}(x) ,$$

where $F_{i,N}$ is the empirical distribution of $P(X_i)$ with respect to a fixed total order $\prec$ on $\mathbb{F}_q^m$ and $|X_i| = N$. At significance level α, the test detects a distinction in the distributions if

$$KS_N > \sqrt{\frac{-ln(\alpha)}{N}}.$$

For small parameters we chose $X_1 = C$ and observed that the rejection rate approaches α as p increases for fixed n and k. Expecting more power for larger values of N we increase parameter sizes and allow X_1 to be a random size N subset of C. In this case, we see that for sufficiently large data sets that the distributions seem to converge.

We further perform a goodness-of-fit test comparing the empirical distribution of $P(X)$ for $X \subseteq \mathbb{F}_q^n$ with $|X| = N$ with the uniform distribution Unif($\mathbb{F}_q^m$). The test asserts that the distributions differ when

$$KS'_N = \sup x \in \mathbb{F}_q^m \; F_N(x) - F(x) \; > \frac{\delta_\alpha}{\sqrt{N}},$$

where $F(x) = \frac{1}{q^m}|\{x' \in \mathbb{F}_q^m : x' \preceq x\}|$ and $\vartheta(\frac{1}{2}, \frac{2i}{\pi}\delta_\alpha^2) = 1 - \alpha$ where ϑ is the Jacobi theta function. We use the trick from [?] of replacing δ_α with $\delta_\alpha + \frac{1}{6\sqrt{N}} + \frac{\delta_\alpha - 1}{N}$ to maximize the accuracy of the tests for small sample sizes. In our case, we used $N = 2048$ for all of the tests. Again, for fixed n and k as p increases we observe that the rejection rate approaches α. Some data from our experiments are presented in Table 6.

Table 6. Rejection rates (R_r) for 100 trials of the Kolmogorov-Smirnov test for sample data values at the $\alpha = 0.05$ level. Test A compares the distributions of $P(X_1)$ and $P(X_2)$, while Test B compares $P(X)$ with Unif($\mathbb{F}_q^m$). In all cases $N = 2048$.

					Test A	$n = 14$	$k = 12$			
p	7	8	9	10	11	12	13	14	15	16
R_r	10	8	2	7	4	7	4	5	4	6

					Test B	$n = 24$	$k = 18$			
p	17	18	19	20	21	22	23	24	25	26
R_r	10	9	7	6	9	7	5	4	5	6

We also conduct experiments comparing the Hamming weight distribution $H(P(\mathbb{F}_q^n))$ to the distribution of the Hamming weight of random vectors in $\mathbb{F}_q^m$, Binomial$(m, 0.5)$. The results of the experiments are presented in Table 7. Again, the data indicate that as p increases the statistical differences become small.

We note that while these data suggest that for larger parameters and in particular for larger values of p that the distribution of ciphertexts is "smoothed" towards uniform in distribution and towards binomial in Hamming weight, it is not easy to judge the rate of observations required to attain significance levels of cryptographic relevance. While the tests seem to have sufficient power to provide results with $N = 2^{11}$ at the $\alpha = 0.05$ level, it is difficult to justify how many observations are required to achieve significance at $\alpha = 2^{-f(n)}$. Verifying that the number of samples must be very large to measure a distinction in the distributions is an open question.

Table 7. Rejection rate (R_r) for 100 trials of the Kolmogorov-Smirnov goodness-of-fit test comparing the Hamming weight distribution on $N = 2048$ ciphertexts with parameters $n = 14$ and $k = 12$ with Binomial$(m, 0.5)$ at the $\alpha = 0.05$ level.

					Test A	$n = 14$	$k = 12$			
p	1	2	3	4	5	6	7	8	9	10
R_r	100	100	42	7	6	2	7	4	3	5

5 Modifications

One clear problem with CBM is the poor decryption failure rate. Since the legitimate user needs to perform q^{n-k} linear algebra steps to invert F, this quantity must remain small; however, inversion is infeasible even with an unique preimage when $\mathbf{x} \in C$. Also, as seen in Section 4, the linear algebra search MinRank attack has a complexity that is only a factor of $2\frac{p}{n}$ greater in the exponent than decryption by a legitimate user. Thus, to achieve a high level of security, extremely large parameters must be used. We propose a couple of modifications that provide the degrees of freedom required to make CBM more versatile.

5.1 PCBM

The first modification exploits polynomial morphisms to avoid infeasible inversion. Specifically, we can consider an embedding of the plaintext space insisting that output of the affine transformation U is never a codeword. We repeat the construction of CBM from Section 2 using n' in place of n and adding to every equation a random linear form. Thus, we have

$$F_i(\mathbf{x}) = \mathbf{x}\mathbf{B}_i\mathbf{x}^\top + \mathbf{x} \cdot \mathbf{b}_i,$$

where $\mathbf{b}_i$ is a random vector of dimension n'. We then choose an invertible affine transformation $T : \mathbb{F}_q^{k+p} \to \mathbb{F}_q^{k+p}$, set $n = n' - 1$ and select an injective affine transformation $U : \mathbb{F}_q^n \to \mathbb{F}_q^{n'}$ such that $Im(U) \cap C = \emptyset$. The public key is then given by $P = T \circ (F\|Q) \circ U$. Clearly, the inversion process for F is the same because the extra summand from the $\mathbf{b}_i$ merely changes the linear form associated with the ith equation.

With this construction, infeasible inversion due to the plaintext begin a codeword is impossible, and the decryption failure analysis simplifies to essentially the same as the injectivity probability in Section 3. Since, due to rank concerns, practical parameters make the decryption failure probability extremely low, PCBM is statistically injective.

5.2 ECBM

In this subsection we propose a modification of the scheme that decouples decryption for the legitimate user from a search through all cosets of the code. Specifically, we propose to use an embedded small instance of EFLASH, see [7], to encode the syndrome corresponding to the plaintext, thus identifying uniquely the correct coset in which to solve for the valid preimage.

Let $\mathbb{K}$ be a degree $d > n - k$ extension of $\mathbb{F}_q$, and let $f : \mathbb{K} \to \mathbb{K}$ be a C^* monomial, $f(x) = x^{q^\theta+1}$ where $(q^\theta + 1, q^d - 1) = 1$. Let $\phi : \mathbb{F}_q^d \to \mathbb{K}$ be an $\mathbb{F}_q$-vector space isomorphism. Then $E = \phi^{-1} \circ f \circ \phi$ is the vector-valued representation of the monomial function f over $\mathbb{F}_q$. Further define

$$E'(\mathbf{x}) = \Pi_a \circ E(V(\mathbf{x}\mathbf{H}^\top)),$$

where Π_a is a codimension a projection and $V : \mathbb{F}_q^{n-k} \to \mathbb{F}_q^d$ is linear of full rank. Finally, let U and T be invertible linear maps of dimensions n and m, respectively, and we compute the public key

$$P = T(F\|E'\|Q) \circ U.$$

Inversion of P is accomplished as follows. Given a ciphertext $\mathbf{y}$, the user first computes $\mathbf{v} = T^{-1}(\mathbf{y})$. Next, $\mathbf{v}$ is parsed into $\mathbf{v}_1$, $\mathbf{v}_2$ and $\mathbf{v}_3$, corresponding to the outputs of F, E' and Q, respectively. The user then randomly searches through the q possible preimages of $\mathbf{v}_2$ under Π_a, inverts E via exponentiation by h satisfying $h(q^\theta + 1) = 1 \pmod{q^d - 1}$, and searches through all preimages $\mathbf{s}$ of these values under V. In this way, the user obtains $q^{d+k+1-n}$ values $\mathbf{s}$ one of which is the valid syndrome corresponding to the preimage $\mathbf{u} = U(\mathbf{x})$ of $\mathbf{v}_1$ under F. The valid syndrome reveals the coset of C containing $\mathbf{u}$, and inversion of F to recover $\mathbf{u}$ proceeds as in Section 2. Finally the plaintext is recovered as $\mathbf{x} = U^{-1}(\mathbf{u})$.

We note a few consequences of using this modification of the code-based scheme. First, at the cost of the inversion of an embedded small EFLASH instance there is no longer an enumeration of cosets step in the inversion of F. Thus the inversion of F is sped up by as much as a factor of q^{n-k-a}. Second, since the complexity of inversion is decoupled from the quantity $n - k$, this value can be made much larger, making the MinRank attack much less efficient. Finally, since we are introducing a C^* monomial map in the scheme, we must reconsider differential and Q-rank attacks.

Luckily, it is straightforward to see that the analysis proving resistance to differential and Q-rank attacks for EFLASH, see [7], are applicable in this context as well. Note that even though n can be chosen much larger than d, the input to the public key is compressed to a dimension of $n-k<d$ before the application of the EFLASH instance $\Pi_a \circ E \circ V$; thus, there is a valid projection and an entire EFLASH instance in the central map. Therefore, P has no differential symmetries or invariants, and can be built to have Q-rank $2a$.

6 Parameter Selection

In selecting parameters, we consider the analyses of the previous section as well as efficiency. The most inefficient operation is inversion of the hidden map $F\|Q$; therefore, we begin by describing an efficient approach.

In key generation we fix the values of our coset representatives, $\mathcal{A}$ and pre-compute the constants $\mathbf{x}'\mathbf{B}_\ell\mathbf{x}'^\top$ and the linear forms $\mathbf{G}\mathbf{B}_\ell\mathbf{x}'^\top$ for each $1\leq\ell\leq k$. Collectively, these values form an affine map $\overline{\mathbf{B}}:\mathbb{F}_q^k\to\mathbb{F}_q^k$. Inversion of $F\|Q$ is then accomplished by finding all preimages $\overline{\mathbf{x}}$ of $\overline{\mathbf{B}}$ and checking that $Q(\mathbf{x}'+\overline{x})$ is the appropriate output. Thus the complexity of inversion is approximately $q^{n-k}k^\omega$.

We find that the limiting attack from Section 4 is the linear algebra search variant of the MinRank attack. With a complexity on the order of $q^{2\frac{p(n-k)}{n}}$, we find that the legitimate user only has an advantage of a factor of $2\frac{p}{n}$ in the exponent over an adversary. Thus $\frac{p}{n}$ must be made large to allow efficient inversion while maintaining security. We examine the case that $\frac{p}{n}$ is sufficiently larger than 3 that the adversary must choose 4 vectors in the MinRank calculation. Then parameters achieving the 128-bit security level are given by $q=2$, $n=148$, $k=132$ and $p=476$. For these parameters the semi-regular degree is 8 which achieves slightly over 128-bit security for the algebraic attack. For a static key version (PCBM), we propose the parameters $q=2$, $n=148$, $n'=149$, $k=133$ and $p=475$. The performance is fairly abysmal for these parameters, with decryption for our non-optimized implementation taking approximately 400 seconds.

The EFLASH variant of the code-based cryptosystem is much more efficient. Parameters achieving the 128-bit security level are $q=2$, $n=148$, $k=131$, d=23, $a=5$ and $p=298$ for ephemeral use with a decryption failure rate of 2^{-16}. With these parameters our non-optimized magma implementation decrypts in 66ms, about 6000 times faster than the code-based scheme without modification.

The EFLASH variant also allows us the freedom to choose parameters for static use. Since the decryption complexity is no longer related to $n-k$, we can set this quantity to a large value. This change has two effects. First, with a sufficiently large value of $n-k$, we no longer need an extremely large value for p to prevent the MinRank attack. In fact, if we chose $n-k$ around 65, then even with $p\approx n$ the MinRank attack does not affect our 128-bit security claim.

Secondly, the allowance of smaller values of p reduces key sizes. Therefore, for static keys we propose the parameters $q = 2$, $n = 148$, $k = 83$, $d = 71$, $a = 5$ and $p = 160$ achieving a decryption failure rate of 2^{-64}.

In addition we propose a parameter set incorporating both of the mentioned modifiers, $EPCBM$. This scheme sacrifices a miniscule amount of speed and key size to allow for static keys. The proposed parameters for 128-bit security are $q = 2$, $n = 148$, $n' = 149$, $k = 132$, $d = 23$, $a = 5$ and $p = 297$.

7 Conclusion

The code-based multivariate encryption scheme (CBM) presented here is an interesting and novel avenue to explore in the attempt to find an efficient and secure multivariate public key encryption scheme. While the literature contains a few multivariate encryption schemes with a claim to solid theoretical foundations, none of these schemes have achieved noteworthy performance at the security levels necessary for future public key applications.

Without modification, the code-based scheme of Section 2 seems to lie solidly in the region of poor performance inhabited by the past multivariate encryption schemes. To avoid truly colossal keys one must endure decryption with precomputed keys that still takes minutes at the 128-bit security level. The reason for this slowness is that decryption is analogous to a form of syndrome decoding without an error-prone message provided. To use this analogy, the plaintext is like a noisy codeword and the ciphertext is like a very noisy hint about the noisy codeword and its syndrome in the form of several inner products of the noisy codeword with its syndrome. Given the private key, the inner products can be extracted, but the syndrome must be guessed before the message and then noisy message are recovered; however, the analogy stretches rather thin here since there is no distance bound for the error and consequently a need merely for uniqueness in the "noisy codeword" and not the codeword itself.

In contrast, modifying the scheme by including a miniature version of EFLASH embedded in the system enhances the performance dramatically. It is no longer the case that a search for the correct syndrome must be undertaken. The correct syndrome is enciphered by the EFLASH component. Thus, the very efficient decoding process given the syndrome and hint allows for rapid decryption. In addition, due to the decoupling of the rank from decryption efficiency, the rank can be increased a great deal resulting in much smaller keys.

We also propose a technique that bypasses the main culprit in decryption failure; specifically, we can ensure that the input to the central map F is never a codeword, an occurance which precludes efficient inversion. With this method we allow decryption failure rates so low that the scheme is often both injective and practically invertible on its range eliminating decryption failures altogether.

There are several directions to explore that this work inspires. First, we may consider whether there is any mechanism by which we can connect the security of this scheme with the syndrome decoding problem. Currently there is no bound on the weight of the coset representative used in decryption, which is why the decoding analogy is not extremely tight, and it is not clear how to force the coset representative to be of small norm without revealing the code structure, which is not allowable for this scheme. In another direction, with the necessity of so many equations, there are numerous modifications that can be added to try to optimize performance. Perhaps one could embed another sufficiently high rank encoding of the syndrome with a different technique that at the appropriate scale is more efficient than the EFLASH modification while maintaining security. For now the possibilities are wide open.

References

1. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM J. Sci. Stat. Comp. **26, 1484** (1997)
2. Huang, Y., Liu, F., Yang, B.: Public-key cryptography from new multivariate quadratic assumptions. In Fischlin, M., Buchmann, J.A., Manulis, M., eds.: Public Key Cryptography - PKC 2012 - 15th International Conference on Practice and Theory in Public Key Cryptography, Darmstadt, Germany, May 21-23, 2012. Proceedings. Volume 7293 of Lecture Notes in Computer Science., Springer (2012) 190–205
3. Regev, O.: On lattices, learning with errors, random linear codes, and cryptography. In Gabow, H.N., Fagin, R., eds.: Proceedings of the 37th Annual ACM Symposium on Theory of Computing, Baltimore, MD, USA, May 22-24, 2005, ACM (2005) 84–93
4. Tao, C., Xiang, H., Petzoldt, A., Ding, J.: Simple matrix - A multivariate public key cryptosystem (MPKC) for encryption. Finite Fields and Their Applications **35** (2015) 352–368
5. Szepieniec, A., Ding, J., Preneel, B.: Extension field cancellation: A new central trapdoor for multivariate quadratic systems. In Takagi, T., ed.: Post-Quantum Cryptography - 7th International Workshop, PQCrypto 2016, Fukuoka, Japan, February 24-26, 2016, Proceedings. Volume 9606 of Lecture Notes in Computer Science., Springer (2016) 182–196
6. Ikematsu, Y., Perlner, R.A., Smith-Tone, D., Takagi, T., Vates, J.: HFERP - A new multivariate encryption scheme. [13] 396–416
7. Cartor, R., Smith-Tone, D.: EFLASH: A new multivariate encryption scheme. In Cid, C., Jr., M.J.J., eds.: Selected Areas in Cryptography - SAC 2018 - 25th International Conference, Calgary, AB, Canada, August 15-17, 2018, Revised Selected Papers. Volume 11349 of Lecture Notes in Computer Science., Springer (2018) 281–299
8. Joux, A., Vitse, V.: A crossbred algorithm for solving boolean polynomial systems. In Kaczorowski, J., Pieprzyk, J., Pomykala, J., eds.: Number-Theoretic Methods in Cryptology - First International Conference, NuTMiC 2017, Warsaw, Poland, September 11-13, 2017, Revised Selected Papers. Volume 10737 of Lecture Notes in Computer Science., Springer (2017) 3–21
9. Moody, D., Perlner, R.A., Smith-Tone, D.: An asymptotically optimal structural attack on the ABC multivariate encryption scheme. In Mosca, M., ed.: Post-Quantum Cryptography - 6th International Workshop, PQCrypto 2014, Waterloo, ON, Canada, October 1-3, 2014. Proceedings. Volume 8772 of Lecture Notes in Computer Science., Springer (2014) 180–196

10. Ding, J., Hodges, T.J.: Inverting HFE systems is quasi-polynomial for all fields. In Rogaway, P., ed.: Advances in Cryptology - CRYPTO 2011 - 31st Annual Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2011. Proceedings. Volume 6841 of Lecture Notes in Computer Science., Springer (2011) 724–742
11. Fouque, P., Granboulan, L., Stern, J.: Differential cryptanalysis for multivariate schemes. In Cramer, R., ed.: Advances in Cryptology - EUROCRYPT 2005, 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Aarhus, Denmark, May 22-26, 2005, Proceedings. Volume 3494 of Lecture Notes in Computer Science., Springer (2005) 341–353
12. Ding, J., Perlner, R.A., Petzoldt, A., Smith-Tone, D.: Improved cryptanalysis of hfev- via projection. [13] 375–395
13. Lange, T., Steinwandt, R., eds.: Post-Quantum Cryptography - 9th International Conference, PQCrypto 2018, Fort Lauderdale, FL, USA, April 9-11, 2018, Proceedings. Volume 10786 of Lecture Notes in Computer Science., Springer (2018)