A Distinguisher for High Rate McEliece Cryptosystems

Jean-Charles Faugère, Valérie Gauthier-Umaña, Ayoub Otmani, Ludovic Perret, Jean-Pierre Tillich

***Abstract*—The Goppa Code Distinguishing (GCD) problem consists in distinguishing the matrix of a Goppa code from a random matrix. Up to now, it is widely believed that the GCD problem is a hard decisional problem. We present the first technique allowing to distinguish alternant and Goppa codes over any field. Our technique can solve the GCD problem in polynomial-time provided that the codes have rates sufficiently large. The key ingredient is an algebraic characterization of the key-recovery problem. The idea is to consider the dimension of the solution space of a linearized system deduced from a particular polynomial system describing a key-recovery. It turns out that experimentally this dimension depends on the type of code. Explicit formulas derived from extensive experimentations for the value of the dimension are provided for "generic" random, alternant, and Goppa code over any alphabet. Finally, we give explanations of these formulas in the case of random codes, alternant codes over any field and binary Goppa codes.**

***Index Terms*—McEliece's cryptosystem, Algebraic cryptanalysis, Goppa code distinguishing.**

I. Introduction

THIS paper investigates the difficulty of the Goppa Code Distinguishing (GCD) problem which first appeared in [1]. This is a decision problem that aims at recognizing a generator matrix of a binary Goppa code from a randomly drawn binary matrix. Up to now, it is assumed that no polynomial time algorithm exists that distinguishes a generator matrix of a Goppa code from a randomly picked generator matrix. The main motivation for introducing the GCD problem is to reduce the security of the McEliece public-key cryptosytem [2] to the difficulty of decoding a random linear code. Since its apparition, this cryptosystem has withstood many attacks and after more than thirty years now, it still belongs to the very few unbroken public key cryptosystems. This situation substantiates the claim that inverting the encryption function, and in particular recovering the private key from public data, is intractable. The classical methods that are dedicated to inverting the McEliece encryption function without finding a trapdoor all resort to the use of the best general decoding algorithms [3]–[7]. All these algorithms, whose time complexity is exponential, attempt to solve the long-standing problem of decoding random linear code [8]. They also assume (implicitly or explicitly) that there does not exist an algorithm that is able to decode more efficiently McEliece public keys. Let us note that if ever such an algorithm exists, it would permit to solve the GCD problem.

On the other hand, no significant breakthrough has been observed with respect to the problem of recovering the private key [9], [10]. This has led to claim that the generator matrix of a binary Goppa code does not disclose any visible structure that an attacker could exploit. This is strengthened by the fact that Goppa codes share many characteristics with random code. For instance they asymptotically meet the Gilbert-Varshamov bound, they have a trivial permutation group, *etc.* Hence, the hardness of the GCD problem has become a classical belief, and as a consequence, a *de facto* assumption to prove the semantic security in the standard model (IND-CPA in [11] and IND-CCA2 in [12]), and the security in the random oracle model against existential forgery [1], [13] of the signature scheme [1].

We present a deterministic polynomial-time distinguisher for high rate codes. This kind of codes are mainly encountered with the public keys of the signature scheme [1]. It is based on the algebraic attack developped against compact variants of McEliece [14]. In this approach, the key-recovery problem is transformed into the one of solving an algebraic system. By using a linearization technique, we are able to derive a linear system whose rank is different from what one would expect. More precisely, we observe experimentally that this *defect* in the rank is directly related to the type of codes. We provide explicit formulas for "generic" random, alternant, and Goppa code over any alphabet. We performed extensive experiments to confirm that the formulas are accurate. Eventually, we prove the formula in the random case and give explanations in the case of alternant codes over any field and binary Goppa codes. We refer to the full version of the paper [15] for the details of the proofs. We insist on the fact that the existence of our distinguisher does not undermine the security of primitives based on Goppa codes, but basically, it proves that the GCD assumption is false for some parameters.

The paper is organized as follows. In Section II, we introduce the algebraic system that any McEliece cryptosystem must satisfy. In Section III, we construct a linear system deduced from the algebraic system. This defines an algebraic distinguisher. We then provide explicit formulas that predicts the behavior of the distinguisher coming from heavy experimentations. In Section IV, we give a proof of its typical behavior in the random case. In Section V and Section VI, we give explanations of the formulas for alternant and binary

J.C. Faugère and L. Perret are with INRIA, Paris-Rocquencourt Center, SALSA Project - UPMC Univ Paris 06, UMR 7606, LIP6, F-75005, Paris, France CNRS, UMR 7606, LIP6, F-75005, Paris, France e-mail: jean-charles.faugere@inria.fr, ludovic.perret@lip6.fr

V. Gauthier-Umaña is with the Department of Mathematics, Technical University of Denmark, Matematiktorvet, Building 303 S, 2800 Kgs. Lyngby, Denmark, e-mail: v.g.umana@mat.dtu.dk

A. Otmani and J.P. Tillich are with SECRET Project - INRIA Rocquencourt, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex - France, e-mail: ayoub.otmani@inria.fr, jean-pierre.tillich@inria.fr

A. Otmani is also with GREYC - Université de Caen - Ensicaen.

Goppa codes. Lastly, we conclude over the cryptographic implications the distinguisher induces.

II. Algebraic Cryptanalysis of McEliece-like Cryptosystems

The McEliece cryptosystem relies on binary Goppa codes which belong to the class of *alternant codes*. It is convenient to describe this class through a parity-check matrix over an extension field $\mathbb{F}_{q^m}$ of $\mathbb{F}_q$ over which the code is defined. For alternant codes of length $n \leqslant q^m$, there exists a parity-check matrix with a very special form related to rectangular Vandermonde matrices:

$$\boldsymbol{V}_r(\mathbf{x},\mathbf{y}) \stackrel{\text{def}}{=} \begin{pmatrix} y_1 & \cdots & y_n \\ y_1x_1 & \cdots & y_nx_n \\ \vdots & & \vdots \\ y_1x_1^{r-1} & \cdots & y_nx_n^{r-1} \end{pmatrix} \quad (1)$$

where $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ are in $(\mathbb{F}_{q^m})^n$.

Definition 1 (Alternant code): The alternant code of order r over $\mathbb{F}_q$ associated to $\mathbf{x} = (x_1, \ldots, x_n) \in (\mathbb{F}_{q^m})^n$ where all x_i's are distinct and $\mathbf{y} = (y_1, \ldots, y_n) \in \left(\mathbb{F}_{q^m}^*\right)^n$ denoted by $\mathscr{A}_r(\mathbf{x},\mathbf{y})$ is $\left\{\mathbf{c} \in \mathbb{F}_q^n \mid \boldsymbol{V}_r(\mathbf{x},\mathbf{y})\mathbf{c}^T = \mathbf{0}\right\}$. The dimension k satisfies $k \geqslant n - rm$.

A key feature about alternant codes of degree r is the fact that there exists a polynomial time algorithm decoding all errors of weight at most $\frac{r}{2}$ *once a parity-check matrix is given in the form* $\boldsymbol{V}_r(\mathbf{x},\mathbf{y})$.

Definition 2 (Goppa codes): The Goppa code $\mathscr{G}(\mathbf{x},\Gamma)$ over $\mathbb{F}_q$ associated to a polynomial $\Gamma(z)$ of degree r over $\mathbb{F}_{q^m}$ and a certain n-tuple $\mathbf{x} = (x_1, \ldots, x_n)$ of distinct elements of $\mathbb{F}_{q^m}$ satisfying $\Gamma(x_i) \neq 0$ for all $i, 1 \leqslant i \leqslant n$, is the alternant code $\mathscr{A}_r(\mathbf{x},\mathbf{y})$ of order r with y_i being defined by $y_i = \Gamma(x_i)^{-1}$.

Goppa codes, viewed as alternant codes, naturally inherit a decoding algorithm that corrects up to $\frac{r}{2}$ errors. But in the case of *binary* Goppa codes, we can correct twice as many errors (Fact 1). The starting point is the following result given in [16, p. 341].

Theorem 1: A binary Goppa code $\mathscr{G}(\mathbf{x},\Gamma)$ associated to a Goppa polynomial $\Gamma(z)$ of degree r without multiple roots is equal to the alternant code $\mathscr{A}_{2r}(\mathbf{x},\mathbf{y})$, with $y_i = \Gamma(x_i)^{-2}$.

Fact 1 ([17]): There exists a polynomial time algorithm decoding all errors of Hamming weight at most r in a Goppa code $\mathscr{G}(\mathbf{x},\Gamma)$ when $\Gamma(z)$ has degree r and has no multiple roots, if $\mathbf{x}$ and $\Gamma(z)$ are *known*.

We are now able to construct an algebraic system as explained in [14] for the McEliece cryptosystem. This algebraic system is the main ingredient of the distinguisher. We assume that the public matrix is a $k \times n$ generator matrix $\mathbf{G}$. We have seen that the knowledge of $\boldsymbol{V}_r(\boldsymbol{x^*},\boldsymbol{y^*})$ permits to efficiently decode. By definition of $\boldsymbol{G}$, we have:

$$\boldsymbol{V}_r(\boldsymbol{x^*},\boldsymbol{y^*})\boldsymbol{G}^T = \mathbf{0}.$$

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be $2n$ variables corresponding to the x_i^*'s and the y_i^*'s. Observe that such x_i^*'s and y_i^*'s are a particular solution [14] of the following system:

$$g_{i,1}Y_1X_1^j + \ldots + g_{i,n}Y_nX_n^j = 0 \quad (2)$$

with $1 \leqslant i \leqslant k$ and $0 \leqslant j \leqslant r-1$, and where the $g_{i,j}$'s are the entries of the known matrix $\boldsymbol{G}$.

Solving this system boils down to finding an equivalent private key. For compact variants [18], [19] of [2], additional structures permit to drastically reduce the number of variables allowing to solve (2) for a large set of parameters in polynomial-time using dedicated Gröbner bases techniques [14]. The general case is currently an open problem.

III. A Distinguisher of Alternant and Goppa Codes

We present in this part the algebraic distinguisher which is based on the non-linear system (2). We can assume that $\mathbf{G} = (g_{ij})$ with $1 \leqslant i \leqslant k$ and $1 \leqslant j \leqslant n$ is in reduced row echelon form over its k first positions *i.e.* $\mathbf{G} = (\mathbf{I}_k \mid \mathbf{P})$ where $\mathbf{P} = (p_{ij})$ for $1 \leqslant i \leqslant k$, $k+1 \leqslant j \leqslant$ is the submatrix of $\mathbf{G}$ formed by its last $n - k = mr$ columns. We describe now a simple way for solving (2). For any $i \in \{1, \ldots, k\}$ and $e \in \{0, \ldots, r-1\}$, we can rewrite (2) as

$$Y_iX_i^e = \sum_{j=k+1}^{n} p_{i,j}Y_jX_j^e. \quad (3)$$

Thanks to the trivial identity $Y_iY_iX_i^2 = (Y_iX_i)^2$, for all i in $\{1, \ldots, k\}$, we get for all $i \in \{1, \ldots, k\}$:

$$\sum_{j=k+1}^{n} p_{i,j}Y_j \sum_{j=k+1}^{n} p_{i,j}Y_jX_j^2 = \left(\sum_{j=k+1}^{n} p_{i,j}Y_jX_j\right)^2.$$

It is possible to reorder this to obtain:

$$\sum_{j=k+1}^{n-1} \sum_{j'>j}^{n} p_{i,j}p_{i,j'} \left(Y_jY_{j'}X_{j'}^2 + Y_{j'}Y_jX_j^2\right) = 0.$$

We can now linearize this system by letting $Z_{jj'} \stackrel{\text{def}}{=} Y_jY_{j'}X_{j'}^2 + Y_{j'}Y_jX_j^2$. We obtain a system $\mathcal{L}_{\mathbf{P}}$ of k linear equations involving the $Z_{jj'}$'s:

$$\mathcal{L}_{\mathbf{P}} \stackrel{\text{def}}{=} \left\{ \sum_{j=k+1}^{n-1} \sum_{j'>j}^{n} p_{i,j}p_{i,j'}Z_{jj'} = 0 \;\middle|\; i \in \{1, \ldots, k\} \right\}. \quad (4)$$

To solve this system it is necessary that the number of equations is greater than the number of unknowns *i.e.* $k \geqslant \binom{mr}{2}$ with the hope that the rank of $\mathcal{L}_{\mathbf{P}}$ denoted by $\operatorname{rank}(\mathcal{L}_{\mathbf{P}})$ is almost equal to the number of variables. Observe that the linear systems (4) have coefficients in $\mathbb{F}_q$ whereas solutions are sought in the extension field $\mathbb{F}_{q^m}$. But the dimension D of the vector space solution of $\mathcal{L}_{\mathbf{P}}$ does not depend on the underlying field because $\mathcal{L}_{\mathbf{P}}$ can always be seen as a system over $\mathbb{F}_{q^m}$. Remark that we obviously have $D = \binom{mr}{2} - \operatorname{rank}(\mathcal{L}_{\mathbf{P}})$.

We carried out intensive computations with Magma [20] by randomly generating alternant and Goppa codes over the field $\mathbb{F}_q$ with $q \in \{2, 4, 8, 16, 32\}$ for r in the range $\{3, \ldots, 50\}$ and several values of m. Furthermore, in our probabilistic model, a random alternant code is obtained by picking uniformly and independently at random two vectors $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ from $(\mathbb{F}_{q^m})^n$ such that the x_i's are all different and the y_i's are all nonzero. A random Goppa

code is obtained by taking a random vector $(x_1, \ldots, x_n)$ in $(\mathbb{F}_{q^m})^n$ with all the x_i's different and a random *irreducible* polynomial $\Gamma(z) = \sum_i \gamma_i z^i$ of degree r. In our experiments, it appears that D is amazingly *large even in the case where* $k \geqslant \binom{mr}{2}$. It even depends on whether or not the code with generator matrix $\mathbf{G}$ is chosen as a (generic) alternant code or as a Goppa code. Interestingly enough, when $\mathbf{G}$ is chosen at random, $\mathrm{rank}(\mathcal{L}_{\mathbf{P}})$ is equal to $\min\left\{k, \binom{mr}{2}\right\}$ with very high probability. In particular, the dimension of the solution space is typically 0 when k is larger than the number of variables $\binom{mr}{2}$ as one would expect. This will be proved in Section IV.

Although this *defect* in the rank is an obstacle to break the McEliece cryptosystem, it can be used to distinguish the public generator from a random code. But before doing so, let us remark first that although the linear system $\mathcal{L}_{\mathbf{P}}$ is defined over $\mathbb{F}_q$, there exists potentially two vector spaces of solutions depending on whether we focus on $\mathbb{F}_{q^m}$ or $\mathbb{F}_q$. We shall see that this ambiguity can be solved through the following definition.

Definition 3: For any integer $r \geqslant 1$ and $m \geqslant 1$, let us denote by $N \stackrel{\text{def}}{=} \binom{mr}{2}$ the number of variables in the linear system $\mathcal{L}_{\mathbf{P}}$ as defined in (4) and D the dimension of the vector space of solutions of $\mathcal{L}_{\mathbf{P}}$. The *normalized dimension* of $\mathcal{L}_{\mathbf{P}}$ denoted by Δ is defined as $\Delta \stackrel{\text{def}}{=} \frac{D}{m}$.

Throughout the paper we consider three cases: when the p_{ij}'s are chosen uniformly and independently at random in $\mathbb{F}_q$ then we denote the normalized dimension by Δ_{random}. When $\mathbf{G}$ is chosen as a generator matrix of a random alternant (*resp.* Goppa) code of degree r, we denote it by $\Delta_{\text{alternant}}$ (*resp.* Δ_{Goppa}). Our experiments have revealed that the normalized dimension of the vector space over $\mathbb{F}_q$ of the solutions of (4) is *predictable* and follows formulas.

Experimental Fact 1 (Alternant Case): As long as $N - m\Delta_{\text{alternant}} < k$, with very high probability the normalized dimension $\Delta_{\text{alternant}}$ is equal to $T_{\text{alternant}}$ where by definition:

$$T_{\text{alternant}} \stackrel{\text{def}}{=} \frac{1}{2}(r-1)\left((2e+1)r - 2\frac{q^{e+1}-1}{q-1}\right) \qquad (5)$$

and where $e \stackrel{\text{def}}{=} \left\lfloor \log_q(r-1) \right\rfloor$.

As for the case of random Goppa codes we also obtain formulas different from those of alternant codes. Note however that the Goppa codes are generated by means of a random irreducible $\Gamma(z)$ of degree r and hence $\Gamma(z)$ has no multiple roots. In particular, we can apply Theorem 1 in the binary case.

Experimental Fact 2 (Goppa Case): As long as $N - m\Delta_{\text{Goppa}} < k$, with very high probability the normalized dimension Δ_{Goppa} is equal to T_{Goppa} where by definition:

$$T_{\text{Goppa}} \stackrel{\text{def}}{=} \begin{cases} \frac{1}{2}(r-1)(r-2) = T_{\text{alternant}} \text{ for } r < q-1 \\ \\ \frac{1}{2}r\Big((2e+1)r - 2q^e + 2q^{e-1} - 1\Big) \text{ for } r \geqslant q-1 \end{cases} \qquad (6)$$

and where e is the unique integer such that:

$$q^e - 2q^{e-1} + q^{e-2} < r \leqslant q^{e+1} - 2q^e + q^{e-1}.$$

Based upon these experimental observations, we are now able to define a *distinguisher* between random codes, alternant codes and Goppa codes. This distinguisher will be in particular useful to distinguish McEliece public keys from random matrices.

Definition 4: Let m and r be integers such that $m \geqslant 1$ and $r \geqslant 1$. Let $\mathbf{G}$ be a $k \times n$ matrix whose entries are in $\mathbb{F}_q$ with $n \leqslant q^m$ and $k \stackrel{\text{def}}{=} n - rm$. Without loss of generality, we assume that $\mathbf{G}$ is systematic *i.e.* $\mathbf{G} = (\mathbf{I}_k \mid \mathbf{P})$. Let $\mathcal{L}_{\mathbf{P}}$ be the linear system associated to $\mathbf{G}$ as defined in (4), and Δ the normalized dimension of $\mathcal{L}_{\mathbf{P}}$. We define the *Random Code Distinguisher* $\mathcal{D}$ as the mapping which takes in input $\mathbf{G}$ and outputs b in $\{-1, 0, 1\}$ such that $\mathcal{D}(\mathbf{G}) = -1$ if $\Delta = T_{\text{alternant}}$, $\mathcal{D}(\mathbf{G}) = 0$ if $\Delta = T_{\text{Goppa}}$, and $\mathcal{D}(\mathbf{G}) = 1$ otherwise.

IV. The Random Case

The purpose of this section is to study the behavior of D_{random}, namely the dimension of the solution space of $\mathcal{L}_{\mathbf{P}}$ when the entries of the matrix $\mathbf{P}$ are drawn independently from the uniform distribution over $\mathbb{F}_q$. In this case, we can show that:

Theorem 2: Assume that $N \leqslant k$ and that the entries of $\mathbf{P}$ are drawn independently from the uniform distribution over $\mathbb{F}_q$. Then for any function $\omega(x)$ tending to infinity as x goes to infinity, we have that as mr goes to infinity

$$\mathbf{prob}\Big(D_{\text{random}} \geqslant mr\omega(mr)\Big) = o(1).$$

Notice that if choose $\omega(x) = \log(x)$ for instance, then asymptotically the dimension D_{random} of the solution space is with very large probability smaller than $mr\log(mr)$. When m and r are of the same order – which is generally chosen in practice – this quantity is smaller than $D_{\text{alternant}}$ or D_{Goppa} which are of the form $\Omega(mr^2)$. The main ingredient for proving Theorem 2 consists in analyzing a certain (partial) Gaussian elimination process on the matrix $\boldsymbol{M} \stackrel{\text{def}}{=} (p_{ij}p_{ij'})_{\substack{1 \leqslant i \leqslant k \\ k+1 \leqslant j < j' \leqslant n}}$. Basically it amounts to view the matrix $\boldsymbol{M}$ in block form, each block consisting in the matrix $\boldsymbol{B}_j = (p_{ij}p_{ij'})_{\substack{1 \leqslant i \leqslant k \\ j < j' \leqslant n}}$ with $k+1 \leqslant j < n$. Each $\boldsymbol{B}_j$ is of size $k \times (rm - j)$. Notice that in $\boldsymbol{B}_j$, the rows for which $p_{i,j} = 0$ consist only of zeros.

To start the Gaussian elimination process with $\boldsymbol{B}_1$, we will therefore pick up $rm - 1$ rows for which $p_{i,k+1} \neq 0$. This gives a square matrix $\boldsymbol{M}_1$. We perform Gaussian elimination on $\boldsymbol{M}$ by adding rows involved in $\boldsymbol{M}_1$ to put the first block $\boldsymbol{B}_1$ in standard form. We carry on this process with $\boldsymbol{B}_2$ by picking now $rm - 2$ rows which have not been chosen before and which correspond to $p_{i,k+2} \neq 0$. This yields a square submatrix $\boldsymbol{M}_2$ of size $rm - 2$ and we continue this process until reaching the last block. The key observation is that:

$$\mathrm{rank}(\boldsymbol{M}) \geqslant \mathrm{rank}(\boldsymbol{M}_1) + \mathrm{rank}(\boldsymbol{M}_2) + \cdots + \mathrm{rank}(\boldsymbol{M}_{rm-1}).$$

A rough analysis of this process yields the theorem above. The important point is what happens for different blocks are independent processes, it corresponds to looking at different rows of the matrix $\mathbf{P}$. A more detailed analysis would probably yield a stronger result that $\mathbf{prob}(D_{\text{random}} \geqslant \omega(mr)) = o(1)$, for any function ω going to infinity with mr or allowing to treat the case $N \geqslant k$ where we would like to show that $\mathbf{prob}(D_{\text{random}} \geqslant N - k + \omega(mr)) = o(1)$. This is beyond the scope of this paper (See details in the full version [15]).

V. Interpretation of the Normalized Dimension – The Alternant Case

We consider alternant codes over $\mathbb{F}_q$ of degree r. The goal is to identify a set of vectors of $(\mathbb{F}_{q^m})^n$ which, after decomposing each entry according to a basis of $\mathbb{F}_{q^m}$ over $\mathbb{F}_q$, provides a basis of the solution space of $\mathcal{L}_{\mathbf{P}}$. Let us observe that to set up the linear system $\mathcal{L}_{\mathbf{P}}$ as defined in (4), we have used the trivial identity $Y_iY_iX_i^2 = (Y_iX_i)^2$. Actually, we can use any identity $Y_iX_i^aY_iX_i^b = Y_iX_i^cY_iX_i^d$ with $a,b,c,d \in \{0,1,\ldots,r-1\}$ such that $a+b=c+d$. It is straightforward to check that we obtain the same algebraic system $\mathcal{L}_{\mathbf{P}}$ with:

$$\sum_{j=k+1}^{n}\sum_{j'>j} p_{i,j}p_{i,j'}(Y_jX_j^aY_{j'}X_{j'}^b + Y_{j'}X_{j'}^aY_jX_j^b + Y_jX_j^cY_{j'}X_{j'}^d + Y_{j'}X_{j'}^cY_jX_j^d) = 0. \quad (7)$$

So, the fact that *there are many different ways of combining the equations of the algebraic system together yielding the same linearized system* $\mathcal{L}_{\mathbf{P}}$ explains why the dimension of the vector space solution is large. For larger values of r, the automorphisms of $\mathbb{F}_{q^m}$ of the kind $x \mapsto x^{q^\ell}$ for some $\ell \in \{0,\ldots,m-1\}$ can be used to obtain the identity for any integers a, b, c, d, ℓ, ℓ' such that $aq^{\ell'} + bq^\ell = cq^{\ell'} + dq^\ell$. We get again the linear system $\mathcal{L}_{\mathbf{P}}$ but the decomposition over $\mathbb{F}_q$ of the entries of vectors obtained from such equations give vectors that are dependent of those coming from the identity $Y_iX_i^aY_i^{q^{\ell-\ell'}}X_i^{bq^{\ell-\ell'}} = Y_iX_i^cY_i^{q^{\ell-\ell'}}X_i^{dq^{\ell-\ell'}}$ if we assume $\ell' \leqslant \ell$. Therefore, we are only interested in vectors that satisfy equations obtained with $0 \leqslant a,b,c,d < r$, $0 \leqslant \ell < m$ and $a + q^\ell b = c + q^\ell d$.

Definition 5: Let a, b, c and d be integers in $\{0,\ldots,r-1\}$ and an integer ℓ in $\{0,\ldots,\lfloor \log_q(r-1)\rfloor\}$ such that $a+q^\ell b = c+q^\ell d$. We define $\boldsymbol{Z}_{a,b,c,d,\ell} \stackrel{\text{def}}{=} \Big(\boldsymbol{Z}_{a,b,c,d,\ell}[j,j']\Big)_{k+1\leqslant j<j'\leqslant n}$ where $\boldsymbol{Z}_{a,b,c,d,\ell}[j,j'] \stackrel{\text{def}}{=} Y_jX_j^aY_{j'}^{q^\ell}X_{j'}^{q^\ell b} + Y_{j'}X_{j'}^aY_j^{q^\ell}X_j^{q^\ell b} + Y_jX_j^cY_{j'}^{q^\ell}X_{j'}^{q^\ell d} + Y_{j'}X_{j'}^cY_j^{q^\ell}X_j^{q^\ell d}$, for any j and j' satisfying $k+1 \leqslant j < j' \leqslant n$.

Without loss of generality, we can assume that $d > b$ and set $\delta \stackrel{\text{def}}{=} d-b$. The next proposition shows that some vectors $\boldsymbol{Z}_{c+q^\ell\delta,b,c,b+\delta,\ell}$ can be expressed as a linear combination of vectors defined with $\delta = 1$.

Proposition 1: Let ℓ, δ, b and c be integers such that $\ell \geqslant 0$, $\delta \geqslant 1$, $1 \leqslant b+\delta \leqslant r-1$ and $1 \leqslant c + q^\ell\delta \leqslant r-1$. Let us assume that $\delta \geqslant 2$ and let $b_i \stackrel{\text{def}}{=} b+i-1$ and $c_i \stackrel{\text{def}}{=} c+q^\ell(\delta-i)$. We have

$$\boldsymbol{Z}_{c+q^\ell\delta,b,c,b+\delta,\ell} = \sum_{i=1}^{\delta} \boldsymbol{Z}_{c_i+q^\ell,b_i,c_i,b_i+1,\ell}. \quad (8)$$

Definition 6: Let $\mathcal{B}_r$ be the set of *nonzero* vectors $\boldsymbol{Z}_{c+q^\ell\delta,b,c,b+\delta,\ell}$ obtained with tuples (δ,b,c,ℓ) such that $\delta = 1$ while satisfying $0 \leqslant b < c \leqslant r-2$ if $\ell = 0$, and if $1 \leqslant \ell \leqslant \lfloor \log_q(r-1)\rfloor$:

$$\begin{cases} 0 \leqslant b \leqslant r-2, \\ 0 \leqslant c \leqslant r-1-q^\ell. \end{cases}$$

Proposition 2: For any integer $r \geqslant 3$ the cardinality of $\mathcal{B}_r$ is equal to $T_{\text{alternant}}$.

Proposition 2 gives an explanation of the value of $D_{\text{alternant}}$ and gives the following heuristic.

Heuristic 1: Consider a certain decomposition of the elements of $\mathbb{F}_{q^m}$ in a $\mathbb{F}_q$ basis. Let $\pi_i : \mathbb{F}_{q^m} \to \mathbb{F}_q$ be the function giving the i-th coordinate in this decomposition where $1 \leqslant i \leqslant m$. By extension we denote for $\boldsymbol{z} = (z_j)_{1\leqslant j\leqslant n} \in (\mathbb{F}_{q^m})^n$ by $\pi_i(\boldsymbol{z})$ the vector $(\pi_i(z_j))_{1\leqslant j\leqslant n} \in \mathbb{F}_q^n$. Then, for any j such that $1 \leqslant j \leqslant n$ and for random choices of x_j's and y_j's, the set $\{\pi_i(\boldsymbol{Z}) \mid 1 \leqslant i \leqslant m, \boldsymbol{Z} \in \mathcal{B}_r\}$ forms a basis of the vector space of solutions of $\mathcal{L}_{\mathbf{P}}$.

VI. Interpretation of the Normalized Dimension – The Binary Goppa Case

In this section we will explain Experimental Fact 2 observed for binary Goppa codes. We denote by r the degree of the Goppa polynomial. The theoretical expression T_{Goppa} has a simpler expression in that special case.

Proposition 3: Let $e \stackrel{\text{def}}{=} \lceil \log_2 r\rceil + 1$ and $N \stackrel{\text{def}}{=} \binom{mr}{2}$. When $q = 2$, Formula (6) can be simplified to:

$$T_{\text{Goppa}} = \frac{1}{2}r\Big((2e+1)r - 2^e - 1\Big). \quad (9)$$

Theorem 1 shows that a binary Goppa code of degree r can be regarded as a binary alternant code of degree $2r$. This seems to indicate that we should have $D_{\text{Goppa}}(r) = mT_{\text{alternant}}(2r)$. This is not the case however. It turns out that $D_{\text{Goppa}}(r)$ is significantly smaller than this. In our experiments, we have found out that the vectors of $\mathcal{B}_{2r}$ still form a generating set for the solution space of $\mathcal{L}_{\mathbf{P}}$, but they are not independent anymore. Our goal is therefore to identify the dependencies between $\pi_i(\boldsymbol{Z})$'s with $\boldsymbol{Z}$ in $\mathcal{B}_{2r}$. Although we are firstly interested in linear relations between the $\pi_i(\boldsymbol{Z})$'s, we shall see that many of them come from $\mathbb{F}_{2^m}$-relations that link directly the $\boldsymbol{Z}$'s as shown by the following proposition which exploits the fact that the Y_i's are derived from the Goppa polynomial $\Gamma(z)$ by $Y_i = \Gamma(X_i)^{-1}$.

Proposition 4: Let t, ℓ and c be integers such that $0 \leqslant t \leqslant r-2$, $1 \leqslant \ell \leqslant \lfloor \log_2(2r-1)\rfloor$ and $0 \leqslant c \leqslant 2r - 2^\ell - 1$. We set $c^* \stackrel{\text{def}}{=} c + 2^{\ell-1}$. It holds that:

$$\sum_{b=0}^{r} \gamma_b^{2^\ell} \boldsymbol{Z}_{c+2^\ell,t+b,c,t+b+1,\ell} = \boldsymbol{Z}_{c^*+2^{\ell-1},2t,c^*,2t+1,\ell-1} + \boldsymbol{Z}_{c+2^{\ell-1},2t+1,c,2t+2,\ell-1}. \quad (10)$$

As a consequence, the set $\{\pi_i(\boldsymbol{Z}) \mid 1 \leqslant i \leqslant m, \boldsymbol{Z} \in \mathcal{B}_{2r}\}$ can not be a basis of the linearized system in the Goppa case.

Proposition 5: The number N_L of equations of the form (10) is $2(r-1)\,(ru + 1 - 2^u)$ where $u \stackrel{\text{def}}{=} \lfloor \log_2(2r-1)\rfloor$.

Notice that each equation of the form (10) involves one vector of $\mathcal{B}_{2r}$ that does not satisfy the other equations. These equations are therefore independent and if we denote by $< \mathcal{B}_{2r} >_{\mathbb{F}_{2^m}}$ the vector space over $\mathbb{F}_{2^m}$ generated by the vectors of $\mathcal{B}_{2r}$ we should have:

$$\dim < \mathcal{B}_{2r} >_{\mathbb{F}_{2^m}} \leqslant |\mathcal{B}_{2r}| - N_L.$$

The experimentations we have made indicate that actually equality holds here. However, this does not mean that the dimension of the vector space over $\mathbb{F}_2$ generated by the set

$\{\pi_i(\boldsymbol{Z}) \mid \boldsymbol{Z} \in \mathcal{B}_{2r}, 1 \leqslant i \leqslant m, \boldsymbol{Z} \in \mathcal{B}_{2r}\}$ is equal to $m \dim < \mathcal{B}_{2r} >_{\mathbb{F}_{2^m}}$. It turns out that there are still other dependencies among the $\pi_i(\boldsymbol{Z})$'s. To see this, let us define the vector $\boldsymbol{Q}_{a,b,c,d,\ell} \stackrel{\text{def}}{=} \left(\boldsymbol{Q}_{a,b,c,d,\ell}[j,j']\right)_{k+1 \leqslant j < j' \leqslant n}$ with:

$$\boldsymbol{Q}_{a,b,c,d,\ell}[j,j'] = \left(\boldsymbol{Z}_{a,b,c,d,\ell}[j,j']\right)^2.$$

Observe also that for any $1 \leqslant i \leqslant m$ we always have $\pi_i(\boldsymbol{Q}_{a,b,c,d,\ell})$ in $\{\pi_i(\boldsymbol{Z}) \mid 1 \leqslant i \leqslant m, \boldsymbol{Z} \in \mathcal{B}_{2r}\}$.

Proposition 6: For any integers $b \geqslant 0$, $t \geqslant 0$, $\delta \geqslant 1$ and ℓ such that $0 \leqslant \ell \leqslant \lfloor \log_2(2r-1) \rfloor - 1$, $b + \delta \leq 2r - 1$ and $t + 2^\ell \delta \leqslant r - 1$, we have

$$\boldsymbol{Z}_{2t+2^{\ell+1}\delta,b,2t,b+\delta,\ell+1} = \sum_{c=0}^{r} \gamma_c^2 \boldsymbol{Q}_{c+2^\ell\delta,b,t+c,b+\delta,\ell}. \qquad (11)$$

Proposition 7: Let N_Q he number of vectors of $\mathcal{B}_{2r}$ satisfying Equation (11) and $u \stackrel{\text{def}}{=} \lfloor \log_2(2r-1) \rfloor$, we have that

$$N_Q = (2r-1)(ru - 2^u + 1).$$

Each of such equation gives rise to m linear equations over $\mathbb{F}_2$ involving the $\pi_i(\boldsymbol{Z})$ for $\boldsymbol{Z}$ in $\mathcal{B}_{2r}$. Therefore, it could be expected that $\Delta_{\text{Goppa}} = |\mathcal{B}_{2r}| - N_L - N_Q$. But, some of the N_Q vectors of $\mathcal{B}_{2r}$ are counted twice because they appear both in linear relations of the form (10) and "quadratic" equations of the form (11). Let $N_{L \cap Q}$ be the number of such vectors. By counting them carefully we can prove that:

Proposition 8: $N_{L\cap Q} = (r-1)\Big((u-\frac{1}{2})r - 2^u + 2\Big)$ where $u \stackrel{\text{def}}{=} \lfloor \log_2(2r-1) \rfloor$.

Proposition 9: For any integer $r \geqslant 2$, we have

$$T_{\text{Goppa}}(r) = |\mathcal{B}_{2r}| - N_L - N_Q + N_{L\cap Q}.$$

VII. Conclusion and Cryptographic Implications

The existence of a distinguisher for the specific case of binary Goppa codes is not valid for any value of r and m but tends to be true for codes that have a rate $\frac{n-mr}{n}$ that is close to one. This kind of codes are mainly encountered with the signature scheme [1]. If we assume that the length n is equal to 2^m and we denote by $r_{\max}$ the smallest integer r such that $N - mT_{\text{Goppa}} \geqslant 2^m - mr$ then any binary Goppa code defined of degree $r < r_{\max}$ can be distinguished (Table I). For example, the binary Goppa code obtained with $m = 13$ and $r = 19$ corresponding to a McEliece public key of 90 bits of security, is distinguishable. More interestingly, all the keys proposed in [21] for the signature scheme can be distinguished. We recall that the existence of such a distinguisher does not undermine the security of [2] and [1]. It only shows that their security should not be reduced to the difficulty of decoding a random linear code by means of the GCD assumption. Therefore this would suggest to directly assume that the McEliece trapdoor function is one-way without any other assumptions.

TABLE I
A binary Goppa code of length $n = 2^m$ and degree $r < r_{\max}$ is distinguishable from a random code.

m	8	9	10	11	12	13	14	15
$r_{\max}$	5	8	8	11	16	20	26	34
m	16	17	18	19	20	21	22	23
$r_{\max}$	47	62	85	114	157	213	290	400

References

[1] N. T. Courtois, M. Finiasz, and N. Sendrier, "How to achieve a McEliece-based digital signature scheme," *Lecture Notes in Computer Science*, vol. 2248, pp. 157–174, 2001.

[2] R. J. McEliece, *A Public-Key System Based on Algebraic Coding Theory*. Jet Propulsion Lab, 1978, pp. 114–116, dSN Progress Report 44.

[3] P. J. Lee and E. F. Brickell, "An observation on the security of McEliece's public-key cryptosystem," in *Advances in Cryptology - EUROCRYPT'88*, ser. Lecture Notes in Computer Science, vol. 330/1988. Springer, 1988, pp. 275–280.

[4] J. S. Leon, "A probabilistic algorithm for computing minimum weights of large error-correcting codes," *IEEE Transactions on Information Theory*, vol. 34, no. 5, pp. 1354–1359, 1988.

[5] J. Stern, "A method for finding codewords of small weight," in *Coding Theory and Applications*, ser. Lecture Notes in Computer Science, G. D. Cohen and J. Wolfmann, Eds., vol. 388. Springer, 1988, pp. 106–113.

[6] A. Canteaut and F. Chabaud, "A new algorithm for finding minimum-weight words in a linear code: Application to McEliece's cryptosystem and to narrow-sense BCH codes of length 511," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 367–378, 1998.

[7] D. J. Bernstein, T. Lange, and C. Peters, "Attacking and defending the McEliece cryptosystem," in *PQCrypto*, ser. LNCS, vol. 5299, 2008, pp. 31–46.

[8] E. Berlekamp, R. McEliece, and H. van Tilborg, "On the inherent intractability of certain coding problems," *IEEE Transactions on Information Theory*, vol. 24, no. 3, pp. 384–386, May 1978.

[9] J. Gibson, "Equivalent goppa codes and trapdoors to mcelieces public key cryptosystem," in *Advances in Cryptology EUROCRYPT 91*, ser. Lecture Notes in Computer Science, D. Davies, Ed. Springer Berlin / Heidelberg, 1991, vol. 547, pp. 517–521.

[10] P. Loidreau and N. Sendrier, "Weak keys in the mceliece public-key cryptosystem," *IEEE Transactions on Information Theory*, vol. 47, no. 3, pp. 1207–1211, 2001.

[11] R. Nojima, H. Imai, K. Kobara, and K. Morozov, "Semantic security for the McEliece cryptosystem without random oracles," *Des. Codes Cryptography*, vol. 49, no. 1-3, pp. 289–305, 2008.

[12] R. Dowsley, J. Müller-Quade, and A. C. A. Nascimento, "A CCA2 secure public key encryption scheme based on the McEliece assumptions in the standard model," in *CT-RSA*, 2009, pp. 240–251.

[13] L. Dallot, "Towards a concrete security proof of Courtois, Finiasz and Sendrier signature scheme," in *WEWoRC*, 2007, pp. 65–77.

[14] J.-C. Faugère, A. Otmani, L. Perret, and J.-P. Tillich, "Algebraic cryptanalysis of mceliece variants with compact keys," in *EUROCRYPT*, ser. Lecture Notes in Computer Science, H. Gilbert, Ed., vol. 6110. Springer, 2010, pp. 279–298.

[15] J.-C. Faugère, V. Gauthier, A. Otmani, L. Perret, and J.-P. Tillich, "A distinguisher for high rate McEliece cryptosystems," Cryptology ePrint Archive, Report 2010/331, 2010, http://eprint.iacr.org/.

[16] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*, 5th ed. Amsterdam: North–Holland, 1986.

[17] N. Patterson, "The algebraic decoding of Goppa codes," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 203–207, 1975.

[18] T. P. Berger, P. Cayrel, P. Gaborit, and A. Otmani, "Reducing key length of the McEliece cryptosystem," in *Progress in Cryptology - Second International Conference on Cryptology in Africa (AFRICACRYPT 2009)*, ser. Lecture Notes in Computer Science, B. Preneel, Ed., vol. 5580, Gammarth, Tunisia, Jun. 21-25 2009, pp. 77–97.

[19] R. Misoczki and P. S. L. M. Barreto, "Compact McEliece keys from Goppa codes," in *Selected Areas in Cryptography (SAC 2009)*, Calgary, Canada, Aug. 13-14 2009.

[20] W. Bosma, J. J. Cannon, and C. Playoust, "The Magma algebra system I: The user language," *J. Symb. Comput.*, vol. 24, no. 3/4, pp. 235–265, 1997.

[21] M. Finiasz and N. Sendrier, "Security bounds for the design of code-based cryptosystems," in *Asiacrypt 2009*, ser. LNCS, M. Matsui, Ed., vol. 5912. Springer, 2009, pp. 88–105.