Math. Struct. in Comp. Science (2009), *vol.* 19, *pp.* 9–23. © 2009 Cambridge University Press
doi:10.1017/S0960129508007251 Printed in the United Kingdom

Phase transition of multivariate polynomial systems

GIORDANO FUSCO[†] and ERIC BACH[‡]

[†]*Computer Sciences Department, Stony Brook University, Stony Brook, NY 11794*
Email: fusco@cs.sunysb.edu

[‡]*Computer Sciences Department, University of Wisconsin-Madison, Madison, WI 53706*
Email: bach@cs.wisc.edu

Received 18 October 2007

A random multivariate polynomial system with more equations than variables is likely to be unsolvable. On the other hand, if there are more variables than equations, the system has at least one solution with high probability. In this paper we study in detail the phase transition between these two regimes, which occurs when the number of equations equals the number of variables. In particular, the limiting probability for no solution is $1/e$ at the phase transition, over a prime field.

We also study the probability of having exactly s solutions, with $s \geqslant 1$. In particular, the probability of a unique solution is asymptotically $1/e$ if the number of equations equals the number of variables. The probability decreases very rapidly if the number of equations increases or decreases.

Our motivation is that many cryptographic systems can be expressed as large multivariate polynomial systems (usually quadratic) over a finite field. Since decoding is unique, the solution of the system must also be unique. Knowing the probability of having exactly one solution may help us to understand more about these cryptographic systems. For example, whether attacks should be evaluated by trying them against random systems depends very much on the likelihood of a unique solution.

1. Introduction

A random multivariate quadratic system in n variables is composed of m equations of the form

$$a_{11}x_1^2 + a_{12}x_1x_2 + \cdots + b_1x_1 + \cdots + b_nx_n = c\,,$$

where the coefficients are independently and uniformly distributed on $GF(p)$ (in the case of $p = 2$ the square terms are not present). More generally, a multivariate polynomial system can have terms up to degree d.

In this paper we study the probability that a multivariate polynomial system has no solutions. If the number of equations is much greater than the number of variables, it is very likely that the system has no solution. On the other hand, if there are more variables than equations, we expect to have at least one solution. For $n + \alpha$ random equations in n variables over $GF(p)$ with p prime, we show that the asymptotic probability that they have no common solution is $e^{-p^{-\alpha}}$. The phase transition occurs when the number of equations equals the number of variables. The asymptotic probability in that case is $1/e$.

We also study the probability that a multivariate polynomial system has exactly s solutions, with $s \geqslant 1$. Asymptotically, this probability follows the Poisson distribution

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press

$\lambda^s e^{-\lambda}/s!$, where $\lambda = e^{-\alpha \log p}$. When $s = 1$, its highest value is e^{-1}, which is attained when the number of equations equals the number of variables. For a fixed set of variables, this probability decays very rapidly as the number of equations increases or decreases.

The motivation for studying the probability of exactly s solutions comes from recent developments in cryptography. Many attacks on cryptosystems have been based on solving a large multivariate polynomial system over a finite field (these include Biryukov and De Cannière (2003), Courtois *et al.* (2000) and Courtois and Pierpzyk (2002)). The idea is to express the cryptosystem as a quadratic or cubic system, and then to use an *ad hoc* method to solve it. The solution of this system is unique because it represents the decoded text. One of the methods used to solve these systems is called XL, which was first proposed in Courtois *et al.* (2000). It has been argued in Courtois *et al.* (2000) and subsequent papers that XL takes advantage of the uniqueness of the solution. Knowing the probability of having exactly one solution, we can understand how often XL has the claimed advantage if applied to random quadratic systems.

The quadratic systems from cryptography are not perfectly random, but in the absence of a better theory, we would like to get some insight by assuming that they are. In particular, the asymptotic probability that a random quadratic system has exactly one solution is $1/e$ if the number of equations equals the number of variables, and decays very rapidly if the number of equations increases or decreases.

We ran a large set of experiments to confirm the validity of our results, including some cases that are not covered by our proofs. We found that the variance of the distance between our formulas and the experimental data is small in most cases.

In order to apply our formulas to polynomial systems from cryptanalysis, we also consider particular configurations that occur in that case. Polynomial systems from cryptanalysis have two important properties: their equations are linearly independent and the systems are sparse. Experimental results confirm that our formulas also remain valid in this case of linearly independent equations. We generated different types of sparse systems and our formulas matched the experimental results in most cases.

Finally, we show the results of the application of our formulas to the quadratic systems of some real cryptographic systems. Using the dimensions of those systems, we determine the probability of having exactly one solution. This probability is extremely small, but, on the other hand, there is a huge number of possible quadratic systems of that size.

This paper is organised as follows. Section 2 gives a brief overview of related work. The probability formulas are derived in Section 3. Section 4 contains the results of some experiments that confirm the general validity of our formulas. In Section 5, we apply our formulas to some cryptographic systems.

2. Related work

Given a quadratic system, there is a well-known procedure for determining the number of solutions. The outline of the method is the following. A single quadratic equation can be transformed into canonical form, as described by Jordan (Jordan 1872) for p odd, and Dickson (Dickson 1999) for $p = 2$. From this form it is easy to count the solutions. Then, a system of quadratic equations can be handled by counting the solutions to a

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press

number of single equations. A detailed description of this procedure for $GF(2)$ is given in Woods (1998). This method requires exponential time.

This is not surprising, as Valiant proved in Valiant (1979) that it is #P-complete to count the number of solutions of a multivariate polynomial system of degree 2 or higher.

The problem we study in this paper is different. We are not computing the number of solutions of given quadratic systems, but we are determining the probability that a random system has no solutions or exactly s solutions.

Recently, much attention has been given to unsatisfiable systems, as there is a direct connection between tautologies and unsatisfiable systems – see, for example, Beame *et al.* (1996), Clegg *et al.* (1996), Buss *et al.* (1997) and Pitassi (1997). The focus of these papers is the study of proof complexity, in particular, the determination of the conditions under which a system is unsatisfiable. Here instead we determine probabilities, such as the chance that a random system is unsatisfiable.

Woods showed (Woods 1998) that there exists a phase transition on multivariate polynomial systems by showing that a system is unsatisfiable when the difference between the number of equations and the number of variables goes to infinity, and that the system has at least one solution when the difference between the number of variables and the number of equations goes to infinity. In this paper we improve Woods's results substantially. In particular, we determine the point at which the phase transition occurs, and compute the limiting value of the probability near the transition point.

To our knowledge this is the first detailed study of phase transitions in polynomial systems. However, there is a well-known phase transition between satisfiability and unsatisfiability for boolean formulas, which has been studied both theoretically and experimentally. We believe that Friegut (Friegut 1999) was the first to prove the existence of a phase transition in boolean formulas. More details have been found experimentally, and surveys of this work have been given by Franco in Franco (2001) and Franco (2005). Our results do say something about boolean formulas, since a boolean formula in conjunctive normal form can be easily transformed into a quadratic system over $GF(2)$ (see, for example, Håstad *et al.* (1993)). However, they are more general, in that we consider polynomial systems of any degree and for any prime field. We also have rigorous theorems to support our experimental observations.

3. Probability of no solutions and of exactly s solutions

The following theorems comprise our main result. Theorem 3.1 is a special case of Theorem 3.2, but we prefer to state it separately to emphasise the phase transition.

Theorem 3.1. Let $d \geqslant 2$ and p be a prime number. Given a multivariate polynomial system of $n + \alpha$ random equations of degree-d in n variables over $GF(p)$, the probability that the system has no solution is $e^{-p^{-\alpha}}$, asymptotically in n.

Corollary 3.1. For a system as in Theorem 3.1, the probability of no solution is e^{-1} if the number of equations equals the number of variables (that is, $\alpha = 0$).

Theorem 3.2. Let $d \geqslant 2$ and p be a prime number and $\lambda = e^{-\alpha \log p}$. Given a multivariate polynomial system of $n + \alpha$ random equations of degree-d in n variables in $GF(p)$, the

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press

probability that the system has exactly $s \geqslant 1$ solutions follows the Poisson distribution $\lambda^s e^{-\lambda}/s!$ asymptotically in n.

Corollary 3.2. For a system as in Theorem 3.2, the probability that the system has exactly $s \geqslant 1$ solutions is $e^{-1}/s!$ if the number of equations equals the number of variables (that is, $\alpha = 0$).

The rest of this section contains the proofs of these results.

Proof of Theorem 3.1. Let p be a prime, and for an n-tuple $x = (x_1, \ldots, x_n)$ of elements from $GF(p)$, let $R_x = (1, \ldots, x_r, \ldots, x_r x_s, \ldots)$. For a system of degree d, R_x contains the monomials up to degree d. Let G be the $p^n \times v$ matrix whose rows are the R_x for distinct x, where $v \approx n^d$ is the number of coefficients in each equation.

Consider the indicator variable

$$Z_x = \begin{cases} 1 & \text{if } x \text{ is a solution to all equations} \\ 0 & \text{otherwise.} \end{cases}$$

Its expectation is $E[Z_x] = p^{-(n+\alpha)}$, and the probability that there is no common solution is

$$E\Big[\prod_x (1 - Z_x)\Big].$$

By the inclusion–exclusion principle, we have

$$\begin{aligned} \prod_x (1 - Z_x) &\geqslant 1 - \sum_x Z_x \\ \prod_x (1 - Z_x) &\leqslant 1 - \sum_x Z_x + \sum_{x,y} Z_x Z_y \\ \prod_x (1 - Z_x) &\geqslant 1 - \sum_x Z_x + \sum_{x,y} Z_x Z_y - \sum_{x,y,z} Z_x Z_y Z_z\,, \end{aligned}$$

and so on. Any partial sum with an even (respectively, odd) number of terms provides a lower (respectively, upper) bound. Also, in these sums, the indices x, y, z, and so on, refer to distinct n-tuples, so each term is effectively a sum over subsets.

Now consider a typical term in the above sum, such as

$$\sum_{x^{(1)},\ldots,x^{(k)}} \prod_i Z_{x^{(i)}}\,.$$

Its expected value is

$$\sum_{x^{(1)},\ldots,x^{(k)}} E\left[\prod_i Z_{x^{(i)}}\right]. \tag{1}$$

A subset for which the corresponding Z's are stochastically independent will contribute $p^{-k(n+\alpha)}$ to the sum. We need to show that most of the subsets are of this type. We say that a subset $\{x^{(1)}, \ldots, x^{(k)}\}$ is in *general position* if the extended vectors $(1, x^{(1)}), \ldots, (1, x^{(k)})$ are linearly independent. Observe that for any general position subset, the random variables

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press

$Z_{x^{(i)}}$, are stochastically independent. The number of general position subsets is

$$\frac{p^n(p^n-1)(p^n-p)\dots(p^n-p^{k-2})}{k!}.$$

Hence, for large n, the general position subsets contribute the value

$$\frac{p^{nk}}{k!}p^{-k(n+\alpha)} = \frac{p^{-\alpha k}}{k!}.$$

If all the rows of G were linearly independent, all subsets would be in general position. Unfortunately this is not true. However, by Lemma 3.1 below, the contribution from subsets not in general position is insignificant compared to this.

Let k^* be the largest odd integer not bigger than v. For quadratic systems, k^* is approximately $n^2/2$, and, in general, k^* goes to infinity with n. Then, if

$$\delta = \Pr[\text{no solution}] - \sum_{k=0}^{k^*-2}\frac{p^{-\alpha k}}{k!},$$

we have, given Lemma 3.1,

$$-\frac{p^{-\alpha(k^*-1)}}{(k^*-1)!}(1+o(1)) \leqslant \delta \leqslant \frac{p^{-\alpha k^*}}{k^*!}(1+o(1)).$$

By Stirling's formula, the upper and lower bounds go to 0 as $n \to \infty$, and the sum is the Taylor series for the (entire) exponential function, so the limit of δ is 0, and we conclude

$$\lim_{n\to\infty}\Pr[\text{no solution}] = e^{-p^{-\alpha}}. \qquad \square$$

Proof of Theorem 3.2. The indicator for exactly s solutions is

$$I = \sum_{x^{(1)}} Z_{x^{(1)}} \sum_{x^{(2)}\neq x^{(1)}} Z_{x^{(2)}} \cdots \sum_{\substack{x^{(s)}\neq x^{(1)}\\ \cdots \\ x^{(s)}\neq x^{(s-1)}}} Z_{x^{(s)}} \prod_{\substack{y\neq x^{(1)}\\ \cdots \\ y\neq x^{(s)}}} (1-Z_y).$$

If we expand this and collect terms, we get

$$\sum_{k\geqslant 0}(-1)^k\binom{s+k}{k}\sum_{x^{(1)},\dots,x^{(s+k)}} Z_{x^{(1)}}\cdots Z_{x^{(s+k)}}.$$

As before, the k-th inner sum is over the subsets of $GF(p)^n$ of size $(s+k)$.

For each n, this expansion of I is a finite sum. Furthermore, the Z's are all non-negative, so taking its expectation produces an alternating series. We can therefore evaluate the limit of these expectations by computing limits termwise as we did in the proof of Theorem 3.1.

So let us consider a particular value of k. The number of general position subsets of size $s+k$ is given by

$$\frac{p^n(p^n-1)\cdots(p^n-p^{s+k-2})}{(s+k)!}.$$

Therefore, their contribution to the expectation is asymptotically

$$\frac{p^{n(s+k)}p^{-(s+k)(n+\alpha)}}{(s+k)!} = \frac{p^{-(s+k)\alpha}}{(s+k)!}.$$

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press

Using Lemma 3.1, with k replaced by $s+k$, we see that including subsets not in general position will not change the value of this limit.

Arguing as before, the expectation of I has the limit

$$\sum_{k\geqslant 0}(-1)^k \binom{s+k}{k} \frac{p^{-(s+k)\alpha}}{(s+k)!} = \frac{p^{-s\alpha}}{s!} \sum_{k\geqslant 0}(-1)^k \frac{p^{-k\alpha}}{k!}\,.$$

The value of the last sum is $e^{-p^{-\alpha}} = e^{-\lambda}$, and this gives the desired result. □

The following lemma is used in the proof of Theorems 3.1 and 3.2. It is the key device for our analysis, as it allows us to compute limiting probabilities as if we had full independence.

Lemma 3.1. Let p be a prime, and for an n-tuple $x = (x_1, \ldots, x_n)$ of elements from $GF(p)$, let $R_x = (1, \ldots, x_r, \ldots, x_r x_s, \ldots)$. Let G be the $p^n \times v$ matrix whose rows are the R_x for distinct x. For fixed k and $n \to \infty$, the contribution to (1) from subsets not in general position goes to 0.

Proof of Lemma 3.1. The points $x^{(0)}, \ldots, x^{(\ell)}$ in $GF(p)^n$ are *affinely independent* if the differences $x^{(1)} - x^{(0)}, \ldots, x^{(\ell)} - x^{(0)}$ are linearly independent. For distinct points, this happens if and only if the corresponding subset is in general position.

Let S be a particular k-subset of the rows of G, corresponding to a set of k points. Let $\ell+1$ be the maximum number of points in S that are affinely independent. We may choose coordinates so that the rows for these points are

$$\begin{array}{ccccccccccccc}
 & x_1 & & x_\ell & x_{\ell+1} & & x_n & x_1^2 & & x_n^2 & & x_i x_j & \\
 & & & & & & & & & & & & \\
1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 & \cdots \\
1 & 1 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 & \cdots & 0 & \cdots \\
 & & & & & \vdots & & & & & & & \\
1 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 1 & \cdots & 0 & \cdots
\end{array} \tag{2}$$

for a quadratic system. In general, for a degree d system, the coordinate would follow a similar pattern.

Assume that $\ell+1 < k$, that is, the subset S is degenerate. We claim that the rank of S cannot be $\ell+1$ (it is obviously at least this large). If it were, any other row would be of the form

$$1 \;\; w_1 \;\; \cdots \;\; w_\ell \;\; 0 \;\; \cdots \;\; 0 \;\; w_1^2 \;\; \cdots \;\; w_n^2 \;\; \cdots \;\; w_i w_j \;\; \cdots$$

Since it is a linear combination of rows from (2), we must have all $w_i w_j = 0$. This means that at most one w_i, say w_1, is non-zero. Then we would have (from inspection of the x_1 and x_1^2 columns)

$$w_1 \cdot 1 = w_1^2\,.$$

So $w_1 \in \{0, 1\}$, but this is impossible since the rows came from distinct points. Hence, the rank is at least $\ell+2$.

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press

For a fixed value of ℓ, there will be at most

$$p^n(p^n-1)(p^n-p)\cdots(p^n-p^{\ell-1})\times p^{k^2}$$

such degenerate subsets of rows. The first factor is an upper bound for the number of ways to choose $\ell+1$ affinely independent points, and the second factor follows from $\ell \leqslant k$ and affine independence. (Once we have chosen coordinates, only w_i with $i \leqslant \ell$ are eligible to be non-zero.) As $n \to \infty$, we have

$$p^n(p^n-1)\cdots(p^n-p^{\ell-1})\times p^{k^2} \sim p^{(\ell+1)n+k^2}\,.$$

Now, if a collection of rows has rank r, the probability of choosing coefficients so that a degree d function vanishes at the points corresponding to those rows is p^{-r}. Similarly, the probability of choosing m sets of coefficients independently with the same property is p^{-mr}. So, for any fixed ℓ, the contribution of degenerate subsets to (1) is at most

$$\frac{p^{(\ell+1)n+k^2}}{p^{mr}} \leqslant \frac{p^{(\ell+1)n+k^2}}{p^{m(\ell+2)}}\,.$$

This is because of our rank estimate. If we substitute $m = n+\alpha$, this becomes

$$\frac{p^{k^2-(\ell+2)\alpha}}{p^n}\,,$$

which has the limit 0 as n goes to infinity. □

3.1. *Extension of the results to* $\mathbb{Z}/(pq)$

In this section we derive the probability formulas for $\mathbb{Z}/(pq)$ where p and q are distinct primes.

Theorem 3.3. Given a multivariate polynomial system of $n+\alpha$ random equations of degree d in n variables in $\mathbb{Z}/(pq)$ with p and q distinct primes, the probability that the system has no solution is $e^{-p^{-\alpha}}+e^{-q^{-\alpha}}-e^{-(p^{-\alpha}+q^{-\alpha})}$, asymptotically in n.

Corollary 3.3. For a system as in Theorem 3.3, the limiting probability of no solution is $2e^{-1}-e^{-2}$ if the number of equations equals the number of variables (that is, $\alpha = 0$).

Theorem 3.4. Let $\lambda = e^{-\alpha \log p}$ and $\mu = e^{-\alpha \log q}$. Given a multivariate polynomial system of $n+\alpha$ random equations of degree d in n variables in $\mathbb{Z}/(pq)$ with p and q distinct primes, the limiting probability that the system has exactly $s \geqslant 1$ solutions is

$$e^{-\lambda-\mu}\sum_{\substack{uv=s\\u,v\geqslant 1}}\frac{\lambda^u\mu^v}{u!\,v!}$$

asymptotically in n.

Corollary 3.4. For a system as in Theorem 3.4, the limiting probability that the system has exactly $s \geqslant 1$ solutions is $e^{-2}\sum_{\substack{uv=s\\u,v\geqslant 1}}\frac{1}{u!\,v!}$ if the number of equations equals the number of variables (that is, $\alpha = 0$).

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press

Proof of Theorem 3.3. A solution does not satisfy the system modulo pq if it does not satisfy it modulo p or it does not satisfy it modulo q. But calculating in this way we are double counting the probability that it does not satisfy it both modulo p and modulo q.

The probability that there are no solutions modulo p and no solutions modulo q is the product of these two probabilities:

$$e^{-p^{-\alpha}} \cdot e^{-q^{-\alpha}} = e^{-(p^{-\alpha}+q^{-\alpha})}.$$

The probability that there are no solutions modulo pq is the sum of the probability of having no solutions modulo p and no solutions modulo q, minus the probability of no solutions modulo p and modulo q together.

$$e^{-p^{-\alpha}} + e^{-q^{-\alpha}} - e^{-(p^{-\alpha}+q^{-\alpha})}.$$ □

Proof of Theorem 3.4. To have exactly s solutions modulo pq, we must have u solutions mod p and v solutions mod q, where $uv = s$. For different factorisations of s, these events are disjoint. Therefore, for n going to infinity, the probability is

$$e^{-\lambda-\mu} \sum_{\substack{uv=s\\ u,v\geqslant 1}} \frac{\lambda^u \mu^v}{u!\,v!}.$$ □

Note that the previous two results can be further extended to products of several distinct primes. To compute the probability of no solution modulo $n = p_1 p_2 \cdots p_r$, we can use the inclusion–exclusion principle. Alternatively, the asymptotic probability that there is no solution is given by

$$1 - \prod_{i=1}^{r} \left(1 - e^{p_i^{-\alpha}}\right).$$

For the limiting probability of $s \geqslant 1$ solutions, we have the formula

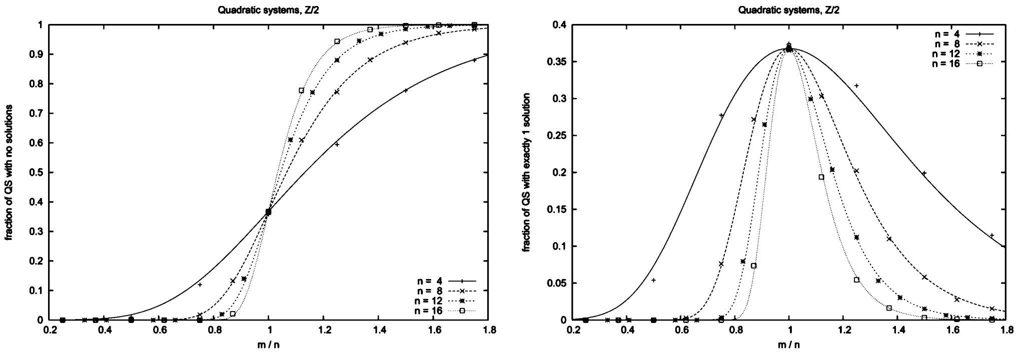
$$e^{-\sum \lambda_i} \sum_{\substack{u_1 u_2 \cdots u_r = s\\ u_i \geqslant 1}} \prod_{i=1}^{r} \frac{\lambda_i^{u_i}}{u_i!}$$

in which $\lambda_i = e^{-\alpha \log p_i}$.

From these considerations, it is clear that the ‘obstruction’ to finding formulas valid for all n is in getting formulas that work mod p^k, when $k \geqslant 2$. We see no difficulty in extending our results to non-prime finite fields, but such extensions are left to the reader.

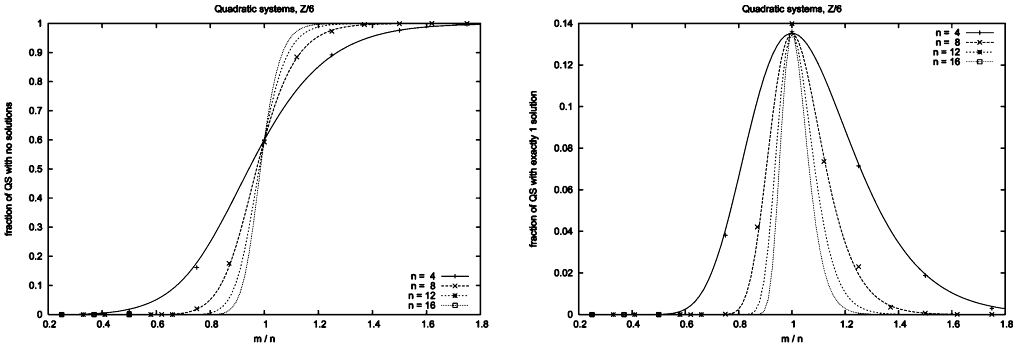
4. Experimental results

We have run a large set of experiments, and these confirm the validity of our results, including for cases that are not covered by our proofs. We generated 10,000 random polynomial systems for each configuration and counted the number of solutions in each case. Figure 1*a* shows the fraction of quadratic systems with no solutions in $\mathbb{Z}_2$. Figure 1*b* shows the fraction of quadratic systems with exactly one solution in $\mathbb{Z}_2$. The continuous line represents the value of the functions described in the previous section, while the discrete symbols give results from the experiments. We can see that the experimental



(*a*) (*b*)

Fig. 1. Fraction of quadratic systems with (*a*) no solutions and (*b*) exactly one solution in $\mathbb{Z}_2$.



(*a*) (*b*)

Fig. 2. Fraction of quadratic systems with (*a*) no solutions and (*b*) exactly one solution in $\mathbb{Z}_6$.

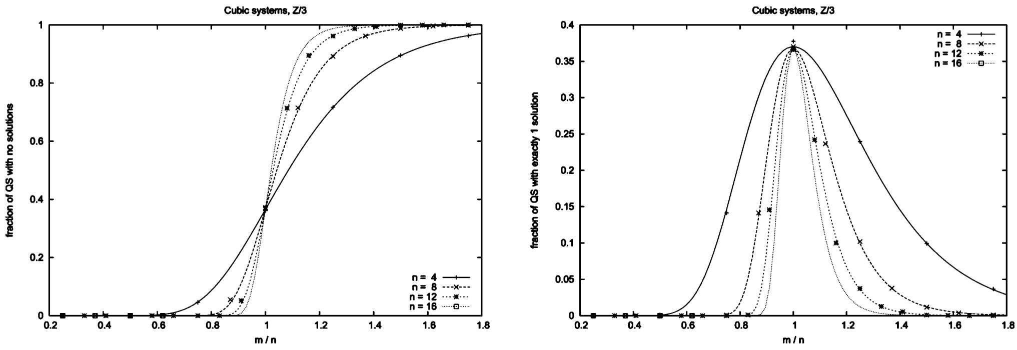
results are consistent with the formulas even in the case of a small number of variables. This is better than we were expecting since the formulas were derived for n going to infinity. Figures 2*a* and 2*b* show similar results for $\mathbb{Z}_6$. Figures 3*a* and 3*b* show that similar results hold for cubic systems. Table 1*a* shows the variance of the experimental values with respect to the formulas for the quadratic systems. The data in this table was obtained by varying n from 4 to 16 and m from 1 to 28.

These experiments were designed to investigate a wider range of applicability than we considered in our theorems. The fact that the variance is small makes us believe that our theorems are valid more generally than our proofs would indicate.

4.1. *Linearly independent equations*

We considered the case of non-linear systems with linearly independent equations. This is motivated by the fact that the quadratic systems used in cryptanalysis are restricted to linearly independent equations.

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press



(*a*) (*b*)

Fig. 3. Fraction of cubic systems with (*a*) no solutions and (*b*) exactly one solution in $\mathbb{Z}_3$.

	no solutions	1 solution
$\mathbb{Z}_2$	$1.66 \cdot 10^{-5}$	$1.95 \cdot 10^{-5}$
$\mathbb{Z}_3$	$7.30 \cdot 10^{-6}$	$7.56 \cdot 10^{-6}$
$\mathbb{Z}_5$	$2.48 \cdot 10^{-6}$	$1.74 \cdot 10^{-6}$
$\mathbb{Z}_6$	$1.54 \cdot 10^{-5}$	$1.60 \cdot 10^{-6}$
$\mathbb{Z}_7$	$2.00 \cdot 10^{-6}$	$2.78 \cdot 10^{-6}$

(*a*)

	no solutions	1 solution
$\mathbb{Z}_2$	$1.79 \cdot 10^{-5}$	$2.09 \cdot 10^{-5}$
$\mathbb{Z}_3$	$7.30 \cdot 10^{-6}$	$7.56 \cdot 10^{-6}$
$\mathbb{Z}_5$	$2.58 \cdot 10^{-6}$	$1.82 \cdot 10^{-6}$
$\mathbb{Z}_6$	$1.65 \cdot 10^{-5}$	$1.72 \cdot 10^{-6}$
$\mathbb{Z}_7$	$2.89 \cdot 10^{-6}$	$4.02 \cdot 10^{-6}$

(*b*)

Table 1. Variance of experimental values with respect to the formulas for (*a*) uniform at random equation and (*b*) linearly independent equations.

The formulas derived in Section 3 hold in the case of linearly independent equations also. This is because the equations of a random polynomial system are linearly independent with very high probability. In fact, a system of degree q with n variables has more than n^q coefficients, which implies that the matrix of the coefficients is rectangular even when we consider $m > n$. As shown in Gerth (1986), it is very likely that a random rectangular matrix has maximal rank.

This is confirmed by the experimental data. We ran the same experiment as the one described at the beginning of Section 4, but enforcing the requirement that the equations must be linearly independent by eliminating the linearly dependent equations. Table 1*b* shows that the variance is very small in this case also.

4.2. *Sparse systems*

In this section we check our formulas on sparse systems. Again the motivation comes from cryptanalysis, where the quadratic systems are usually sparse. In order to simulate the sparseness, we consider three kind of sparse systems:

1 Each coefficient can be 0 with probability z and non-zero with probability $1 - z$. Note that the known term can still assume any value with equal probability.

2 Each equation contains exactly a fraction f of the variables, that is, the coefficients of the remaining variables are 0.

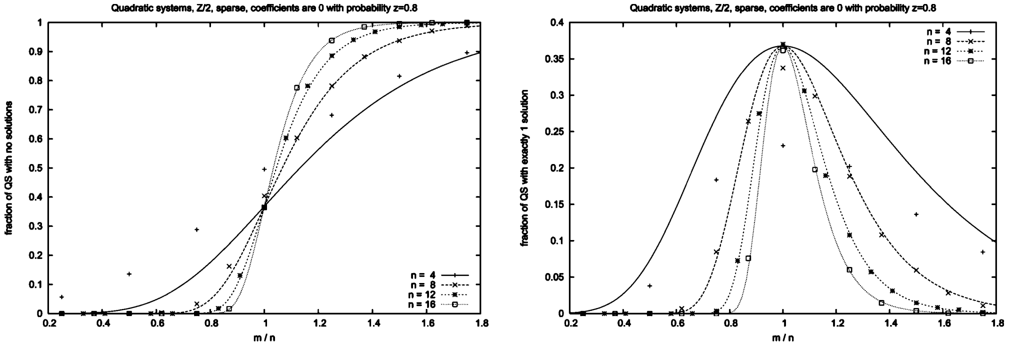
z	no solutions	1 solution
0.5	$3.66 \cdot 10^{-6}$	$6.30 \cdot 10^{-6}$
0.6	$1.62 \cdot 10^{-5}$	$9.04 \cdot 10^{-6}$
0.7	$7.14 \cdot 10^{-5}$	$3.21 \cdot 10^{-5}$
0.8	$1.82 \cdot 10^{-3}$	$6.81 \cdot 10^{-4}$
0.9	$1.32 \cdot 10^{-2}$	$3.31 \cdot 10^{-3}$

(*a*)

	no solutions	1 solution
$\mathbb{Z}_2$	$3.74 \cdot 10^{-5}$	$3.27 \cdot 10^{-5}$
$\mathbb{Z}_3$	$8.05 \cdot 10^{-5}$	$3.62 \cdot 10^{-5}$
$\mathbb{Z}_5$	$1.54 \cdot 10^{-4}$	$7.81 \cdot 10^{-5}$
$\mathbb{Z}_6$	$1.19 \cdot 10^{-3}$	$7.17 \cdot 10^{-5}$
$\mathbb{Z}_7$	$1.27 \cdot 10^{-4}$	$4.00 \cdot 10^{-5}$

(*b*)

Table 2. Variance of experimental values with respect to the formulas for (*a*) different values of z with random system in $\mathbb{Z}_3$ and (*b*) different fields when the coefficients are zero with probability $z = 2/3$.



(*a*) (*b*)

Fig. 4. Fraction of quadratic systems with (*a*) no solutions and (*b*) exactly one solution in $\mathbb{Z}_2$ with coefficients set to 0 with probability $z = 2/3$.

3 Bi-affine equations. These are the type of equations used in Rijndael's cryptanalysis (see, for example, Courtois and Pierpzyk (2002)).

Case 1: The coefficients have higher probability of being 0. This is the most generic type of sparseness we consider. The variance between the formulas from Section 3 and the experimental results is very small for values of z up to 0.7. Table 2*a* shows how the variance varies using different values of z with a random system in $\mathbb{Z}_3$. A similar situation is obtained in other prime fields. Table 2*b* shows the value of the variance of random systems in different fields where the coefficients are zero with probability $z = 2/3$.

If z is smaller than 0.7, the results are very similar to Figures 1*a* and 1*b*. Figures 4*a* and 4*b* show the result obtained with random systems in $\mathbb{Z}_2$ where the coefficients are zero with probability $z = 0.8$. The plot shows that the formula does not provide a good approximation for a system with $n = 4$ variables, but it still works for bigger values of n.

Case 2: Each equation contains exactly a fraction f of the variables. In this case the variance from the experiments is much higher. Table 2*a* shows the values of the variance of random system in $\mathbb{Z}_3$ generated by varying f from 0.1 to 0.5. Similar results are obtained in other fields where f is fixed to 0.5, as shown in Table 2*b*.

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press

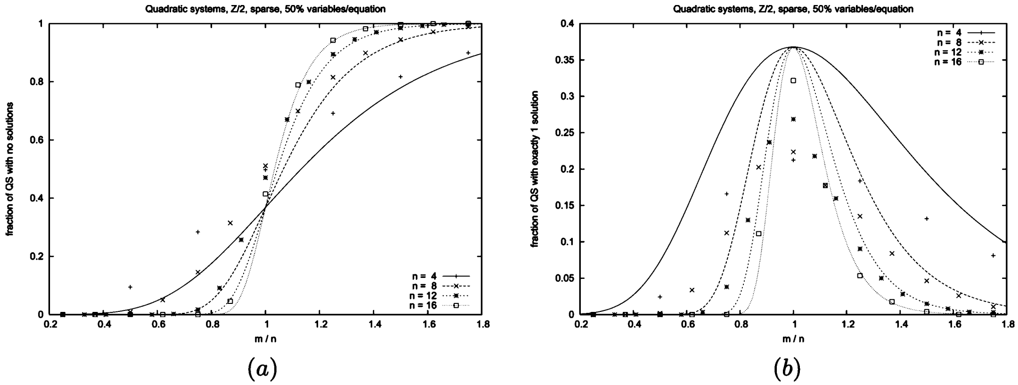


Fig. 5. Fraction of quadratic systems with (*a*) no solutions and (*b*) exactly one solution in $\mathbb{Z}_2$ with exactly 50% variables per equation.

f	no solutions	1 solution
0.1	$4.16 \cdot 10^{-1}$	$1.74 \cdot 10^{-2}$
0.2	$2.58 \cdot 10^{-1}$	$1.65 \cdot 10^{-2}$
0.3	$8.78 \cdot 10^{-2}$	$1.45 \cdot 10^{-2}$
0.4	$3.65 \cdot 10^{-3}$	$9.54 \cdot 10^{-3}$
0.5	$5.21 \cdot 10^{-3}$	$2.87 \cdot 10^{-3}$

(*a*)

	no solutions	1 solution
$\mathbb{Z}_2$	$3.34 \cdot 10^{-3}$	$2.56 \cdot 10^{-3}$
$\mathbb{Z}_3$	$5.33 \cdot 10^{-3}$	$2.94 \cdot 10^{-3}$
$\mathbb{Z}_5$	$9.17 \cdot 10^{-2}$	$4.11 \cdot 10^{-3}$
$\mathbb{Z}_6$	$1.87 \cdot 10^{-2}$	$1.02 \cdot 10^{-3}$
$\mathbb{Z}_7$	$1.91 \cdot 10^{-2}$	$7.90 \cdot 10^{-3}$

(*b*)

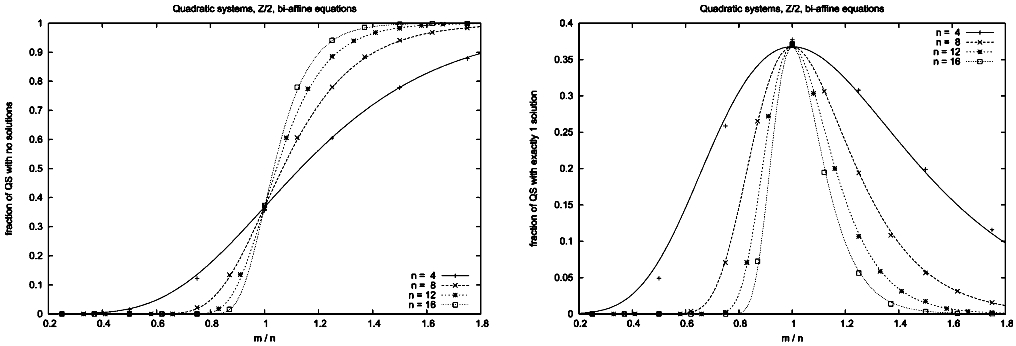
	no solutions	1 solution
$\mathbb{Z}_2$	$9.17 \cdot 10^{-6}$	$2.00 \cdot 10^{-5}$
$\mathbb{Z}_3$	$2.63 \cdot 10^{-5}$	$3.99 \cdot 10^{-5}$
$\mathbb{Z}_5$	$1.24 \cdot 10^{-6}$	$9.04 \cdot 10^{-6}$
$\mathbb{Z}_6$	$4.14 \cdot 10^{-5}$	$1.23 \cdot 10^{-5}$
$\mathbb{Z}_7$	$4.10 \cdot 10^{-6}$	$5.10 \cdot 10^{-6}$

(*c*)

Table 3. Variance of experimental values with respect to the formulas for (*a*) random system in $\mathbb{Z}_3$ generated varying f from 0.1 to 0.5 and (*b*) different fields when f is fixed to 0.5. (*c*) Variance of experimental values with respect to the formulas for different fields with bi-affine equations.

An explanation of these results is that this model reduces the freedom of the random equations, which, in fact, are no longer perfectly uniform at random. For this reason, the formulas no longer describe the phenomenon exactly, and the variance from the experiment is much higher. This is also evident from figures 5*a* and 5*b*.

Case 3: Bi-affine equations. Bi-affine equations are only used for quadratic systems. The variables are partitioned into two sets of equal size. Each quadratic term is composed of one variable from the first set and one from the second (that is, two variables from the same set never appear multiplied together). The variance in this case is small – see Table 3*c*. The results for $\mathbb{Z}_2$ are plotted in Figures 6*a* and 6*b*.



(*a*) (*b*)

Fig. 6. Fraction of quadratic systems with (*a*) no solutions and (*b*) exactly one solution in $\mathbb{Z}_2$ with bi-affine equations.

Cryptosystem	n	m	α
Khazad	6464	7664	1200
Misty1	3856	3856	0
Kasumi	4264	4264	0
Camellia-128	3584	6224	2640
Rijndael-128	3296	6296	3000
Serpent-128	16640	17680	1040

(*a*)

Cryptosystem	Total # of systems	Pr[1 solution]
Khazad	$6.86 \cdot 10^{6249185}$	$5.81 \cdot 10^{-362}$
Misty1	$1.68 \cdot 10^{2239709}$	$1/e$
Kasumi	$4.20 \cdot 10^{2738543}$	$1/e$
Camellia-128	$1.64 \cdot 10^{1934992}$	$1.91 \cdot 10^{-795}$
Rijndael-128	$5.40 \cdot 10^{1636625}$	$8.13 \cdot 10^{-904}$
Serpent-128	$3.58 \cdot 10^{41683551}$	$8.49 \cdot 10^{-314}$

(*b*)

Table 4. (*a*) Sizes of quadratic systems from cryptography. (*b*) Total number of quadratic systems and the probability of exactly 1 solution.

5. Equations from cryptographic systems

In this section we apply the formula for exactly one solution to the sizes of quadratic systems for some well-known cryptographic systems. The results obtained with the experimental data (see Section 4) give us confidence in using the formula in this case, even if this is not a case covered by our proofs. The data in Table 4*a* are from Biryukov and De Cannière (2003). All the equations are in $\mathbb{Z}_2$. For the quadratic systems of Misty1 and Kasumi, the parameters m and n are in the range of applicability of our formulas.

Table 4*b* shows that for many systems the probability of having exactly one solution is extremely small. However, the number of systems with exactly one solution is not that small if we consider that the total number of possible systems is huge.

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press

One inference that can be drawn from this study is that quadratic systems with unique solutions are relatively rare, so rare that in most cases studying the performance of solution algorithms for random systems might not tell us much about their efficacy in attacking specific cryptosystems.

6. Conclusions and open problems

In this paper, we have shown that the probability that a random polynomial system has no solution has a phase transition when the number of equations equals the number of variables. The value of the probability at the phase transition is $1/e$ if the computation is over a prime field.

We have also shown that probability of having exactly s solutions, $s \geqslant 1$, follows a Poisson distribution with parameter $\lambda = e^{-\alpha \log p}$ for prime fields.

We then extended the result to $\mathbb{Z}/(pq)$ with p and q distinct primes. Extending the result to the case of $\mathbb{Z}/(p^r)$ with p prime is an open problem.

Adapting the formulas in the case of sparse systems where each equation contains exactly a fixed number of variables is also an open problem.

Acknowledgments

We would like to thank Dieter van Melkebeek and Janos Simon for their helpful comments. We would also like to express our thanks for the support provided by NSF grant CCF-0523680.

References

Biryukov, A. and De Cannière, C. (2003) Block ciphers and systems of quadratic equations. Proc. FSE 2003. *Springer-Verlag Lecture Notes in Computer Science* **2887** 274–289.

Beame, P., Impagliazzo, R., Krajíček, J., Pitassi, T. and Pudlák, P. (1996) Lower bounds on Hilbert's Nullstellensatz and propositional proofs. *Proc. London Math. Soc.* **73** 1–26.

Buss, S., Impagliazzo, R., Krajíček, J., Razborov, A. A. and Sgall, J. (1997) Proof complexity in algebraic systems and bounded depth Frege systems with modular counting. *Comput. Complex.* **6** 256–298.

Clegg, M., Edmonds, J. and Impagliazzo, R. (1996) Using the Groebner basis algorithm to find proofs of unsatisfiability. *Proc. 28th Ann. ACM Symp. Theory Comput.* 174–183.

Courtois, N., Klimov, A., Patarin, J. and Shamir, A. (2000) Efficient algorithms for solving overdefined systems of multivariate polynomial equations. In: Proc. Eurocrypt 2000. *Springer-Verlag Lecture Notes in Computer Science* **1807** 392–407.

Courtois, N. and Pierpzyk, J. (2002) Cryptanalysis of block ciphers with overdefined systems of equations. In: Proc. Asiacrypt 2002. *Springer-Verlag Lecture Notes in Computer Science* **2501** 267–287.

Dickson, L. E. (1899) Determination of the structure of all linear homogeneous groups in a Galois field which are defined by a quadratic invariant. *Amer. J. Math.* **21** 193–256.

Friegut, E. (1999) Necessary and sufficient conditions for sharp thresholds of graph properties and the k-SAT problem. *Amer. J. Math.* **12** 1017–1054.

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press

Franco, J. (2001) Results related to threshold phenomena research in satisfiability: lower bounds. *Theoret. Comput. Sci.* **265** (1-2) 147–157.

Franco, J. (2005) Typical case complexity of satisfiability algorithms and the threshold phenomenon. *Disc. Appl. Math.* **153** (1-3) 89–123.

Gerth, F. (1986) Limit probabilities for coranks of matrices over GF(q). *Lin. Multilin. Alg.* **19** 79–93.

Håstad, J., Phillips, S. and Safra, S. (1993) A well-characterized approximation problem. *Inf. Proc. Lett.* **47** (6) 301–305.

Jordan, C. (1872) Sur la forme canonique des congruences du second degré et le nombre de leurs solutions. *J. Math. Pures. Appls.* **17** (2) 368–402. (Abstract of results in *C. R. Acad. Sci.* (1872) **74** 1093–1095.)

Pitassi, T. (1997) Algebraic propositional proof systems. In: Immerman, N. and Kolaitis, P. G. (eds.) *Descriptive Complexity and Finite Models*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science **31** 215–244.

Valiant, L. G. (1979) The complexity of enumeration and reliability problems. *SIAM J. Comput.* **8** 4120–421.

Woods, A. R. (1998) Unsatisfiable systems of equations, over a finite field. *Proc. 39th Ann. Symp. Found. Comput. Sci.* 202–211.

https://doi.org/10.1017/S0960129508007251 Published online by Cambridge University Press