ECC 2010, 10/18/2010

My Last 24 Years in Crypto:

A Few Good Judgments and Many Bad Ones

Neal Koblitz, Univ. of Washington, koblitz@math.washington.edu

Much of this talk is based on the article “ECC: The Serpentine Course of a Paradigm Shift” by Ann Hibner Koblitz, N.K., & Alfred Menezes, to appear in the ECC issue of J. Number Theory.

In the meantime it’s available at eprint.iacr.org – do a title word search for “serpentine”.

See especially Section 11 concerning the security implications of isogeny walks.

Conventional wisdom

- In cryptography, for greatest security choose parameters as randomly as possible.
- In elliptic/hyperelliptic curve cryptography it's safest to choose the defining equation to have random coefficients.
- It's okay to use special curves for reasons of efficiency if you insist, but some day that choice might come back to bite you.

In 1991, I proposed the use of the non-supersingular $\mathbf{F}_2$ -curves (also called anomalous binary curves)

$$y^2 + xy = x^3 + 1 \quad \text{or} \quad y^2 + xy = x^3 + x^2 + 1$$

because they seemed to have some efficiency advantages over random curves.

NSA liked these curves, and at Crypto 1997 J. Solinas gave a talk presenting a thorough and definitive treatment of how to optimize ECC operations on these curves.

At present these curves are one of the three sets of NIST-recommended curves (each set containing 5 curves at a range of security levels).

Some people have been mistrustful of this family of curves, in part because of the “conventional wisdom” given above.

However, in the random-vs-special debate about curve selection, Menezes and I found reason to question the conventional wisdom that random is always more secure.

There are various scenarios in which someone who chooses ECC with a special curve might end up better off than someone else who chooses a random curve.

Some such scenarios are suggested by recent work on isogenies. (For more details see Section 11 of the “serpentine course” paper.)

Isogenies

E_1, E_2 defined over $\mathbf{F}_q$

An <u>isogeny</u> ψ: $E_1 \rightarrow E_2$ defined over $\mathbf{F}_q$ is a non-constant rational map defined over F_q that maps ∞ to ∞. Its <u>degree</u> is its degree as a rational map. In our setting the degree is also the order of the kernel of the isogeny.

Any isogeny has a “dual” isogeny going the other way, so we get an equivalence relation of “isogenous” elliptic curves.

Tate’s Theorem: E_1 and E_2 are isogenous over $\mathbf{F_q}$ iff they have the same number of $\mathbf{F_q}$-points.

Low-degree isogenies are easy to construct, but high-degree isogenies are usually not.

Endomorphisms

Let $t = q + 1 - \#E(\mathbf{F}_q)$ denote the trace of an elliptic curve E defined over $\mathbf{F}_q$.

An endomorphism of E is an isogeny from E to itself that is defined over the algebraic closure of $\mathbf{F}_q$.

We shall consider the case of ordinary curves E, meaning that t is prime to the characteristic of $\mathbf{F}_q$. In that case all endomorphisms are defined over $\mathbf{F}_q$.

The endomorphisms form a ring, denoted End(E), that contains the subring $\mathbf{Z}$ of scalar multiplications $P \to nP$.

Let $\Delta = t^2 - 4q < 0$ denote the discriminant of E.

Then $K = Q(\sqrt{\Delta})$ is the CM-field of E.

We have $\Delta = c_0^2\, d$, where $d < 0$ is the discriminant of K.

Then End(E) is an order of the ring of integers $\mathbf{Z}_K$. Its index c in $\mathbf{Z}_K$ is called the conductor of End(E).

The elliptic curves isogenous to a given E can be partitioned according to their endomorphism ring.

These endomorphism classes are determined by the conductor c, and they are in 1-to-1 correspondence with divisors c of c_0.

The number of isomorphism classes of curves in a given endomorphism class is equal to the class number of the order, which is approximately equal to ch_K.

For example, if Δ is squarefree, then all $O(\sqrt{q})$ curves in the isogeny class of E have the same endomorphism ring of conductor 1.

If c_0 is a large prime, then the isogeny class consists of a small number of curves whose endomorphism ring is the full ring of integers Z_K , and the remaining $O(\sqrt{q})$ curves have endomorphism ring of conductor c_0 .

Let ℓ denote a prime. If there is a degree-ℓ isogeny between E_1 and E_2, then either the two curves have the same endomorphism ring, or else the conductors satisfy either

$$c_1 = \ell c_2 \quad \text{or} \quad c_2 = \ell c_1 .$$

By the conductor gap between two endomorphism classes we mean the largest prime that divides one conductor and not the other.

If there is a large conductor gap between two endomorphism classes, then one cannot go from a curve in one class to a curve in the other by a string of low-degree isogenies.

Conversely, by a result of Jao, Miller, and Venkatesan, within an endomorphism class or among classes with small conductor gaps one can efficiently travel randomly and uniformly throughout the set of curves.

Isogenies allow one to transport the discrete log problem from one curve to another. That is, the discrete log problem is “random self-reducible” within a set of endomorphism classes with small conductor gaps.

Definition. The L-conductor-gap class of E is the set of all endomorphism classes in the isogeny class of E that have conductor gap < L with End(E).

Suppose that an algorithm were found that solves the discrete log problem in time T_1 in a proportion ε of all elliptic curves over $\mathbf{F}_q$, where the property of being a "weak" curve is independent of isogeny and endomorphism class.

Then the discrete log can be found on any curve in an L-conductor gap class in time roughly $T_1 + T_2/\varepsilon$, where T_2 denotes the time for constructing a low-degree isogeny – assuming, of course, that the L-conductor-gap class contains more than $1/\varepsilon$ curves.

It is the possibility of random isogeny walks through a conductor-gap class that under certain circumstances might make a random curve less secure than a special curve.

For a random curve all isogenous curves are in the same conductor-gap class, because Δ has negligible probability of being divisible by the square of a large prime.

Let's look at a hypothetical scenario.

In this example $\mathbf{F}_q$ is a prime-degree extension of $\mathbf{F}_2$.

We'll suppose that some version of Weil descent or another approach some day leads to a faster-than-sqrt attack on a small but non-negligible proportion of curves defined over $\mathbf{F}_q$.

NIST's 2000 Digital Signature Standard recommends 5 elliptic curves over prime fields and 10 over binary fields. For each of 5 binary fields they suggest one random curve and one anomalous binary curve.

The largest binary field is the degree-571 extension of $\mathbf{F}_2$, which should provide more than the 256 bits of security needed to protect a high-security AES private key.

The conventional wisdom is that, if anything, the random curve B-571 is a safer choice than the anomalous binary curve K-571.

However, let's suppose that a proportion ε of all curves over this field could be attacked by a new faster-than-sqrt algorithm, and that the "weak" property is independent of isogeny and endomorphism class.

The curve B-571 has squarefree discriminant and so isogeny walks can fan out from B-571 throughout its isogeny class, which consists of roughly 2^{285} curves. After $O(1/\varepsilon)$ isogenies, the DLP on B-571 can be transported to a weak curve.

In contrast, K-571 has discriminant $\Delta = -7c_0^2$ with c_0 the product of a 22-bit prime and a 263-bit prime.

The endomorphism ring of K-571 has conductor 1, i.e., it is the full ring of integers of $Q(\sqrt{-7})$. Thus, the 2^{262}-conductor-gap class of K-571 has only about 2^{22} curves, and so if ε is much less than 2^{-22}, the DLP on K-571 probably cannot be transported to a weak curve by isogenies.

Under our hypothetical assumptions, K-571 is likely to be safer than B-571.

What conclusions do we want to draw?

Not that we should prefer special curves over random ones.

All we can say is that we don’t really know.

It’s a judgment call.

To give a similar example over a prime field, suppose we choose a random prime B and a random even number A such that

(i) $p = A^2 + B^2$ is prime;

(ii) either $n = (p+1)/2 - A$
or else $n = (p+1)/2 + A$
is prime.

Then the elliptic curve E over $\mathbf{F}_p$ defined (for suitable a in $\mathbf{F}_p$) by $y^2 = x^3 - ax$ is the only curve (up to isomorphism) in its conductor-gap class.

Remarks. 1. The only NIST-recommended curves over a prime field are random ones.

2. In his Ph.D. thesis Wenhan Wang has found that a very similar situation exists for genus-2 curves. That is, curves over a prime field whose Jacobians have a large endomorphism ring are often isolated, in the sense that you can't travel widely from them using isogeny walks.

An abbreviated history of embarrassing misjudgments I've made in the last 24 years

First major one:

In the late 1980's it seemed (to me at least) that any elliptic curve group would be secure as long as its order is prime or almost prime.

So for pedagogical reasons why not use the simplest possible curves? And this is what I often did (in my introductory book and in articles and talks).

It's an elementary exercise to show that the curve

$y^2 = x^3 - x$ over $\mathbf{F}_p$ with $4|(p+1)$

or

$y^2 + y = x^3$ over $\mathbf{F}_p$ with $3|(p+1)$

has group order $p+1$.

Just choose p so that $(p+1)/4$ or $(p+1)/6$ is prime, and ECC is secure, or so I thought.

These curves also have some nice efficiency advantages for computing point multiples, especially over extension fields of $\mathbf{F}_2$ and $\mathbf{F}_3$.

Then in 1991 Menezes-Okamoto-Vanstone showed that the Weil pairing gives a reduction of the ECDLP to the DLP on the multiplicative group of an extension of the field of definition.

And for supersingular curves, such as the two written above, the extension degree is very small. Usually it's 2, as in the above cases.

I felt chagrined and embarrassed.

This killed supersingular curves for ECC and made me feel foolish for having used them so often as illustrative examples.

Next embarrassing episode:

In the early 1990's, Mike Fellows and I became captivated by the notion that, despite the fiasco with knapsacks, good cryptosystems could in fact be constructed from NP-hard combinatorial problems.

We even wrote a paper with the exuberant title "Combinatorial Cryptosystems Galore!"

There was only one actual example that we spent some time developing, and it had a sorry history.

As I recount in my book *Random Curves*:

“Mike Fellows and I… constructed a system based on… *ideal membership…* that involved polynomials, and we challenged people to try to crack it.

“The most attractive feature of our cryptosystem was the name that Mike thought up for it: *Polly Cracker.*

“It was very inefficient, and before long some papers were published that indeed cracked the code.”

Back to ECC:

During the first 15 years of ECC my feeling was that it didn't matter what field you worked over. You had to avoid generic algorithms by working in groups of large prime order, and after MOV you had to avoid supersingular curves.

But otherwise you could use whatever field you most enjoy working with, and security is unaffected by that choice.

Late 1990's: Frey proposes Weil descent to attack the DLP on curves over composite degree extension fields;
then Gaudry, Hess, Smart, Galbraith, Menezes, Teske, and others find weak curves over certain binary fields.

Fortunately, other people (such as Scott Vanstone) had had better instincts than I had, and all commercial implementations and all ECC standards used prime fields or prime-degree extensions of $\mathbf{F}_2$.

I was very bad at anticipating future developments.

In early 1998 I published *Algebraic Aspects of Cryptography.* In a section titled “Cultural Background” I discussed the Birch and Swinnerton-Dyer Conjecture, after which I essentially apologized to my readers for taking their time with something that, while mathematically important, has no relevance for cryptography.

A mere 8 months later I was eating those words, after I received an email from J. Silverman outlining a striking new approach to the ECDLP.

It was a variant – somewhat backwards – version of index calculus, and for that reason Silverman called it “xedni calculus.”

What was most alarming for ECC people was that Silverman used the heuristics of the BSD Conjecture (and an analytic rank formula of Mestre) to boost the likelihood of a successful attack on the ECDLP.

After a lot of initial worry about xedni (fueled by our concern that RSA would use xedni as a weapon in their public relations battle with ECC, which was still going strong in 1998), I found that we could use the height function to show that xedni wouldn't work.

I was so thrilled about this success in defending ECC that I gave a talk at ECC 2000 titled

"Miracles of the Height Function:
A Golden Shield Protecting ECC"

At around the same time a paper by Silverman and Suzuki made a detailed examination of index calculus and explained why it wouldn't work.

Essentially, the Silverman-Suzuki paper elaborated on the argument that Vic Miller made in his original ECC paper in 1985.

At ECC 2007, Silverman made a similar analysis for all 4 ways one could try index or xedni with liftings to global fields.

But alas! Index calculus has reared its evil head during the last few years.

For example, Gaudry and Diem found subexponential index calculus algorithms for the ECDLP on elliptic curves defined over the degree-m extension of $\mathbf{F}_q$ as m and q grow suitably.

Regrettably, much cryptographic writing exudes a brash certainty about the work.

Abstracts and introductions to papers often read as if they were written by marketing people or as part of a patent application, full of hype with little connection to reality.

For example (from iacr.org/2007/438):

> "…permits savings on bandwidth and storage… substantially improves computational efficiency and scalability over any existing scheme with suitable functionality…

> “In contrast to the only prior scheme to provide this functionality, ours offers improved security… We provide formal security definitions and support the proposed scheme with security proofs…”

This paper by Boldyreva-Gentry-O’Neill-Yum constructed pairing-based “sequential aggregate signatures” (meaning that several parties in sequence compose a single compact signature).

The amusing thing about this example is that about a year later Hwang, Lee, and Yung showed that a crucial security proof in this paper was fallacious.

They also broke the corresponding protocol.

The claim has often been made that reduction arguments constitute “proofs of security” that can be offered to the public as a guarantee.

From the preface to the book by Katz and Lindell:

> "…cryptographic constructions can be proven secure with respect to a clearly-stated definition of security and relative to a well-defined cryptographic assumption.
>
> "This is the essence of modern cryptography, and what has transformed cryptography from an art to a science. The importance of this idea cannot be over-emphasized."

And anyone who's dismayed by the large number of fallacious proofs in the provable security literature is supposed to be consoled by the prospect that advances in "theorem-proving" software will soon make it possible to prove the security of our protocols automatically, with no longer any possibility of flaws in the proofs; human mistakes and failings will supposedly disappear from the process of establishing guarantees of security.

For more discussion of this dubious claim, see "Another look at automated theorem-proving", http://www.iacr.org/2007/401.pdf

Anyone who's bewildered by the exotic nature of some of the cryptographic assumptions that underlie security proofs for many of the pairing-based protocols is supposed to be reassured by Boyen's exuberant explanation:

> "The newcomer to this particular branch of cryptography will… be astonished by the sheer number, and sometimes creativity, of these assumptions…
>
> "…in comparison to the admittedly quite simpler algebraic structures of twentieth-century public-key cryptography… the new 'bilinear' groups offer a much richer palette of cryptographically useful trapdoors than their 'unidimensional' counterparts…"

On the one hand, we see the trend of bold and boastful writing by crypto researchers.

On the other hand, we see a long history of misjudgments and uncertainty that continues to the present.

How can we reconcile the disciplinary culture of our field with reality?