Breaking post-quantum cryptography with arithmetic geometry

Andrew Sutherland

October 19, 2023

Cryptography today

Most public-key cryptosystems in use today are motivated by two number-theoretic questions:

- Factoring: given $n = pq$, determine p (RSA).
- Discrete logarithms (DLP): given a and $b = a^n \bmod p$ determine n.

Both can heuristically be solved in $\exp(1.93(\log n)^{1/3}(\log\log n)^{2/3})$ time via the number field sieve. Much faster than brute force, but too slow for 3072-bit values of n or p.

Better form of the second problem: replace the (multiplicative) group $\mathbb{F}_p^\times$ with the (additive) group of rational points on an elliptic curve $E/\mathbb{F}_p$:

- ECDLP: given $P, Q = nP \in E(\mathbb{F}_p)$ determine n.

The fastest algorithms for solving ECDLP run in time $\exp(\frac{1}{2}\log n)$.
This makes 256-bit ECDLP problems about as hard as 3072-bit DLP-problems.

RSA is primarily used for authentication and ECDLP is primarily used for key exchange.

Elliptic curves over finite fields

Let $p > 3$ be prime, and let $\mathbb{F}_q$ be the field with $q = p^e$ elements (assume $q = p$ for now). For any squarefree cubic $f \in \mathbb{F}_q[x]$ the equation

$$E\colon y^2 = f(x),$$

defines an elliptic curve, whose rational points consist of the set

$$E(\mathbb{F}_q) := \{(x, y) \in \mathbb{F}_p \times \mathbb{F}_p : y^2 = f(x)\} \cup O,$$

where O is the projective point $(0 : 1 : 0)$ on the homogeneous cubic $yz^2 = z^3 f(x/z)$.

The number of rational points has the form

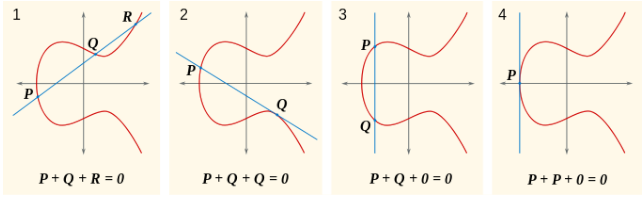
$$\#E(\mathbb{F}_q) = q + 1 - t,$$

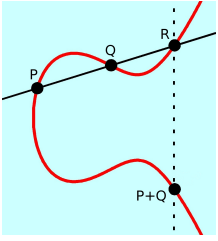
where the Frobenius trace t satisfies $|t| \leq 2\sqrt{q}$ and can be computed in polynomial time.

$E(\mathbb{F}_q)$ can be given the structure of an abelian group with identity O.

The elliptic curve group law

Three points on a line sum to zero.





Ephemeral Diffie–Hellman key exchange

Let $E/\mathbb{F}_q$ be an elliptic curve and fix $P \in E(\mathbb{F}_q)$ (in public). Assume $N = |P| \approx q$ is prime.

Alice and Bob can agree on a random shared secret as follows:

1. Alice chooses random $a \in [1, N]$, sends $aP = \overbrace{P + \cdots + P}^{a}$ to Bob.
2. Bob chooses random $b \in [1, N]$, sends bP to Alice.
3. Alice computes $abP = Q$ and Bob computes $baP = Q$.

The coordinates of Q depend on the random integer ab and can be hashed to yield a shared secret with roughly $\log q$ random bits.

An eavesdropper may know E, P, aP and bP, but not a, b, or Q.[1]
It is believed that computing Q from these values is as hard as ECDLP.

Even though E and P are fixed, we get a new random Q each time.

[1] One should use RSA authentication to protect against a man-in-the-middle attack.

The quantum threat

As shown by Peter Shor, one can easily factor integers and solve the DLP if one has access to a decent-size quantum computer (the complexity is not just polynomial, it is quadratic).

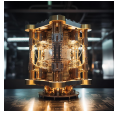
More generally, one can use a quantum computer to solve the hidden subgroup problem (HSP): given a finite abelian group G with unknown cyclic subgroup H and an efficiently computable $f\colon G \to S$ that injects H-cosets of G into S, find a nontrivial element of H.

To solve ECDLP on $Q = nP$ with $N = |P|$ prime one uses

$$G := \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z}, \quad H := \langle (n, 1) \rangle, \quad S := \langle P \rangle, \quad \begin{aligned} f\colon G &\to S \\ (x, y) &\mapsto yQ - xP \end{aligned}$$

Any nonzero $(x, y) \in H$ can be used to compute $n = x/y \bmod p$.

A reliable 10^4-qubit quantum computer with 10^{12} gates breaks most current crypto.
As of this writing 3072-bit RSA and 256-bit ECDLP remain well out of reach.



Post-quantum cryptography

In 2016 the National Institute of Standards and Technology (NIST) solicited proposals for "quantum-secure" cryptographic protocols: these run on a classical computer but are intended to be secure against an adversary with a quantum computer.

NIST received 69 submissions for the first round of evaluations in 2017, including proposals based on lattice problems, coding theory, and random walks in isogeny graphs.

At the end of the first round, 26 candidates were chosen for a 2nd round of evaluation in 2019, of which 15 made it to the third round in 2020.

On July 5, 2022 NIST recommended four of these 15 for standardization, and four to enter a fourth round for further evaluation.

Three weeks later, the fourth round candidate *Supersingular Isogeny Key Encapsulation* (SIKE), was spectacularly broken using an attack that had not been considered during the 6 years of post-quantum evaluation. The attack uses arithmetic geometry, not quantum computing.

Isogenies

Let k be a field of characteristic p with algebraic closure $\bar{k}$. Elliptic curves over k form a category whose morphisms are isogenies: group homomorphisms defined by rational maps.

The kernel of an isogeny φ is the group $\ker\varphi := \{P \in E(\bar{k}) : \varphi(k) = 0\}$.
We call an isogeny cyclic when $\ker\varphi$ is a cyclic group.

Isogenies $E \to E$ are endomorphisms; the maps $[n]\colon P \mapsto nP$ are examples.
For $p \nmid n$ we have $\ker[n] \simeq \mathbb{Z}/n\mathbb{Z} \oplus \mathbb{Z}/n\mathbb{Z}$, so $[n]$ is not cyclic.

The degree of an isogeny is the degree $\deg\varphi$ of the rational functions that define it.
When $\deg\varphi$ is coprime p we have $\deg\varphi = \#\ker\varphi$.

Every finite subgroup G of $E(\bar{k})$ is the kernel of an isogeny $\varphi\colon E \to E'$ with $\deg\varphi = \#G$.
The isogeny φ and E' are determined (up to isomorphism) by G; we may write E/G for E'.

Every isogeny $\varphi\colon E \to E'$ has an associated dual isogeny $\hat{\varphi}\colon E' \to E$ with $\hat{\varphi} \circ \varphi = [\deg\varphi]$.

Isogeny graphs

For primes $\ell \neq p$, there are $\ell + 1$ non-trivial cyclic subgroups of $\mathbb{Z}/\ell\mathbb{Z} \oplus \mathbb{Z}/\ell\mathbb{Z}$.
Each is the kernel of an ℓ-isogeny (an isogeny of degree ℓ).

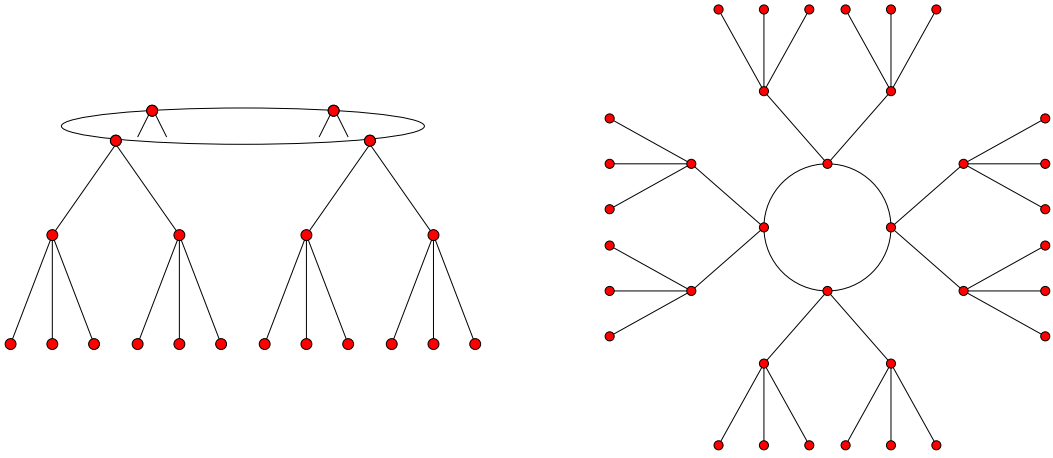
The isogeny graph $\mathcal{G}_\ell(k)$ has elliptic curves E/k as vertices and ℓ-isogenies as edges. Each vertex in $\mathcal{G}_\ell(k)$ has degree at most $\ell + 1$. It will be less than $\ell + 1$ if an ℓ-isogeny leads to an elliptic curve over $\bar{k}$ that is not defined over k (this happens), hence not a vertex in $\mathcal{G}_\ell(k)$.

The graph $\mathcal{G}_\ell(k)$ is not connected. For $k = \mathbb{F}_q$ elliptic curves E_1 and E_2 cannot lie in the same component unless $\#E_1(\mathbb{F}_q) = \#E_2(\mathbb{F}_q)$, and even then they might lie in different components.

The components of $\mathcal{G}_\ell(k)$ have a rich and varied structure, including "volcanoes", "whirlpools", and "spines", however the components most suitable for cryptography are Ramanujan graphs.

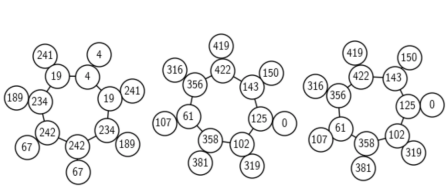
Elliptic curves with Frobenius trace $t \equiv 0 \bmod p$ are supersingular. In characteristic p all such curves are defined over $\mathbb{F}_{p^2}$ and the supersingular subgraph of $\mathcal{G}_\ell(\mathbb{F}_{p^2})$ is a connected $(\ell + 1)$-regular Ramanujan graph: random walks of length $\log_\ell p$ yield a uniform distribution over the $\approx p/12$ supersingular vertices, independent of the starting point.

Ordinary isogeny graphs

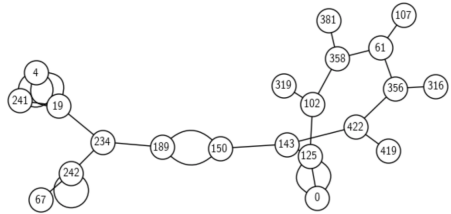


Side and top views of a 3-volcano over a finite field taken from *Isogeny volcanoes*.

Supersingular isogeny graphs



(a) The $\mathcal{G}_2(\mathbb{F}_p)$ for $p = 431$

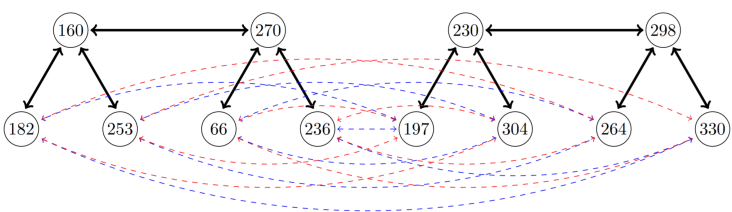


(b) The spine $\mathcal{S} \subset \mathcal{G}_2(\overline{\mathbb{F}_p})$ for $p = 431$.

Figure 3.3: *Stacking, folding and attaching by an edge for $\mathfrak{p} = 431$ and $\ell = 2$. The leftmost component of $\mathcal{G}_2(\mathbb{F}_p)$ folds, the other two components stack, and the vertices 189 and 150 get attached by a double edge.*

Image taken from *Adventures in Supersingularland* by Sarah Arpin, Catalina Camacho-Navarro, Kristin Lauter, Joelle Lim, Kristina Nelson, Travis Scholl, and Jana Sotáková.

Supersingluar isogeny graphs



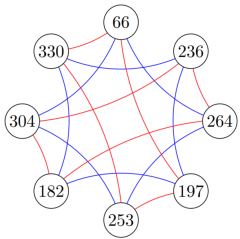


FIGURE 5. A whirlpool with two components.

Image taken from *Orienting supersingular isogeny graphs* by Leonardo Colò and David Kohel.

Supersingular Isogeny Diffie–Hellman (SIDH)

SIDH is a variant of the ephemeral Diffie–Hellman protocol that replaces scalar multiplication by a random integer with a random walk in a supersingular isogeny graph:

Alice and Bob use a supersingular elliptic curve $E_0/\mathbb{F}_{p^2}$ to establish a shared secret as follows:

1. Alice chooses a random integer a encoded in base-2 and computes E_a by taking an a-walk in the 2-isogeny graph, starting from E_0; she sends E_a to Bob.
2. Bob chooses a random integer b encoded in base-3 and computes E_b by taking a b-walk in the 3-isogeny graph, starting from E_0; he sends E_b to Alice.
3. Alice computes $(E_b)_a$ (an a-walk from E_b), and Bob computes $(E_a)_b$ (a b-walk from E_a).
4. Alice and Bob use $j((E_b)_a) = j((E_a)_b)$ is the shared secret,
   where $j(E) := 1728\frac{4A^3}{4A^3+27B^2}$ is the j-invariant of $y^2 = x^3 + Ax + B$.

No known classical/quantum algorithm efficiently computes $j((E_b)_a)) = j((E_a)_b)$ from E_a, E_b.

But this description glosses over a crucial detail.
Alice and Bob need to to compute their two a-walks and b-walks compatibly!

Supersingular Isogeny Key Encapsulation (SIKE)

SIKE uses $E_0 \colon y^2 = x^3 + 6x^2 + x$ and $p = 2^e 3^f - 1$ and $\#E_0(\mathbb{F}_{p^2}) = (p+1)^2 = 2^{2e}3^{2f}$.
Public parameters E_0, P_a, Q_a, P_b, Q_b with $E_0[2](\mathbb{F}_{p^2}) = \langle P_a, Q_a \rangle$ and $E_0[3](\mathbb{F}_{p^2}) = \langle P_b, Q_b \rangle$.

1. Alice computes $\varphi_a \colon E_0 \to E_a = E_0/\langle P_a + aQ_a \rangle$ as a chain of 2-isogenies $E_0 \to \cdots \to E_a$ (this is her a-walk) and sends Bob $\varphi_a(P_b)$, $\varphi_a(Q_b)$, and E_a.
2. Bob computes $\varphi_b \colon E_0 \to E_a = E_0/\langle P_b + bQ_b \rangle$ as a chain of 3-isogenies $E_0 \to \cdots \to E_b$ (this is his b-walk) and sends Alice $\varphi_b(P_a)$, $\varphi_b(Q_a)$, and E_b.
3. Alice uses $\varphi_b(P_a), \varphi_b(Q_a)$ to compute $\varphi_b(P_a + aQ_a) = \varphi_b(P_a) + a\varphi_b(Q_a)$, which allows her to compute $(E_b)_a$ as $E_b/\langle \varphi_b(P_a + aQ_a) \rangle$.
4. Bob uses $\varphi_a(P_b), \varphi_a(Q_b)$ to compute $\varphi_a(P_b + bQ_a) = \varphi_a(P_b) + b\varphi_a(Q_b)$, which allows him to compute $(E_a)_b$ as $E_a/\langle \varphi_a(P_b + bQ_b) \rangle$.

In addition to E_a and E_b, this protocol sends $\varphi_a(P_b)$, $\varphi_a(Q_b)$, $\varphi_b(P_a)$, $\varphi_b(Q_a)$ in the clear.
This additional information is what makes the new attack possible, but the details are subtle.

The breaking of SIKE

On July 30, 2022, Castryck and Decru posted *An efficient key recovery attack on SIDH*, which gives a practical algorithm to compute b given E_0, P_a, Q_a, and E_b, $\varphi_b(P_a)$, $\varphi_b(Q_b)$.

- They provide Magma code that solves the \$50,000 SIKE challenge (217-bit prime) from Microsoft in about 5 minutes (they claimed the cash prize on July 22).
- Their code breaks standard security level in the SIKE proposal (434-bit prime) in about an hour, and the highest security level in the SIKE proposal (751-bit prime) in about 20 hours. Later improvements reduced these times to under a minute and under an hour.
- On August 8 Maino and Martindale post *An attack on SIDH with arbitrary starting curve*, which generalizes the attach to any SIDH scheme that exposes images of torsion points.
- Castryck–Decru and Maino–Martindale present their results in a pair of talks given at the Algorithmic Number Theory Symposium (ANTS XV) on August 10.
- On August 17 Robert posts *Evaluating isogenies in polylogarithmic time*, giving a less practical but more general algorithm that is provably polynomial–time.

The 1997 paper *The number of curves of genus two with elliptic differentials*, by Ernst Kani is the key ingredient used by all of these results. It has nothing to do with cryptography.

Abelian surfaces

An abelian variety is a projective algebraic variety (zero locus of a set of homogeneous polynomials) that is also an algebraic group (the group is defined by regular rational maps).

Elliptic curves are abelian varieties of dimension one.
A product of elliptic curves is an abelian surface, an abelian varieties of dimension two.
A product of supersingular elliptic curves yields a superspecial abelian surface.

But most abelian surfaces (even superspecial ones) are Jacobians $\mathrm{Jac}(X)$ of a genus 2 curve

$$X\colon y^2 = f(x),$$

where $f(x)$ is a square free sextic. Points on $\mathrm{Jac}(X)$ are represented by pairs of points on X.

As with elliptic curves, one can define (ℓ,ℓ)-isogeny graphs of (principally polarized) abelian surfaces in which edges are (ℓ,ℓ)-isogenies with kernels isomorphic to $\mathbb{Z}/\ell\mathbb{Z} \oplus \mathbb{Z}/\ell\mathbb{Z}$. One way to construct such an isogeny is to take the product of two ℓ-isogenies of elliptic curves.

Note that for an abelian surface A and $\ell \neq p$ we have $A[\ell] \simeq \mathbb{Z}/\ell\mathbb{Z} \oplus \mathbb{Z}/\ell\mathbb{Z} \oplus \mathbb{Z}/\ell\mathbb{Z} \oplus \mathbb{Z}/\ell\mathbb{Z}$.

The group law on the Jacobian of a genus 2 curve

We define $\{A, B\} + \{C, D\} + \{E, F\} = 0$.

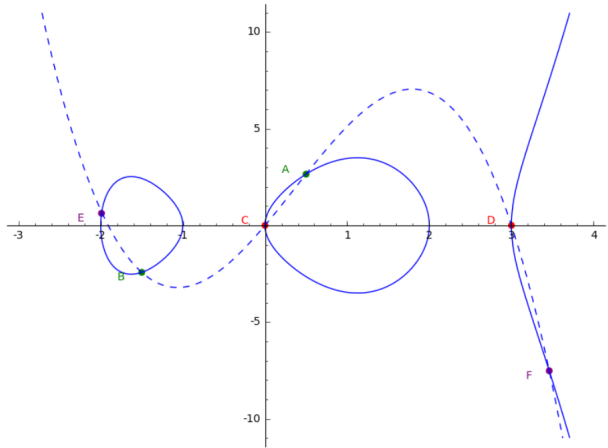


Image credit: *Cryptographic applications of isogeny graphs of genus 2 and 3 curves*, Chloe Martindale, 2019.

The key idea

There are about $p^2/288$ products of supersingular elliptic curves defined over $\mathbb{F}_{p^2}$.
But the total number of superspecial abelian surfaces defined over $\mathbb{F}_{p^2}$ is roughly $p^3/2880$.

For cryptographic-size p it is **astronomically unlikely** we will ever encounter a product of elliptic curves in a random walk in the (ℓ, ℓ)-isogeny graph of superspecial abelian surfaces.

But in the SIDH setup, we can use the curves E_0, E_b and the points $\varphi_b(P_a)$ and $\varphi_b(Q_a)$ to explicitly construct p a scenario where this is guaranteed to happen in the $(2, 2)$-isogeny graph, and we can do this for any value of b, including all but the first or last 3-adic digit of b.

We don't know what b is, but there are only 3 possibilities for its first or last digit, and we can try them all. If we guess wrong, our construction won't produce a walk that hits a product of elliptic curves, and this gives us a way to test whether a candidate digit value is correct or not.

The key tool that enables this construction are isogeny diamonds, which were introduced by Ernst Kani in 1997 (long before anyone was thinking about isogeny-based cryptography).

Isogeny diamonds

Let $\phi\colon A \to B$ and $\psi\colon A \to C$ be isogenies of elliptic curves with $\gcd(\deg\phi, \deg\psi) = 1$.
We then have an (orthogonal) isogeny diamond

$$\begin{array}{ccc} A & \xrightarrow{\phi} & B \\ \downarrow{\scriptstyle \psi} & & \downarrow{\scriptstyle \psi'} \\ C & \xrightarrow{\phi'} & D \end{array}$$

where ϕ' and ψ' are determined by $\ker\phi' = \psi(\ker\phi)$ and $\ker\psi' = \phi(\ker\psi)$.

Lemma (Kani 1997)

Let $d = \deg\phi + \deg\psi$ and let $G := \{(\phi(P), \psi(P)) : P \in A[d]\} \subseteq B \times C$. Then

$$\begin{pmatrix} \hat{\phi} & \hat{\psi} \\ -\psi' & \phi' \end{pmatrix}$$

is a (d,d)-isogeny $B \times C \to A \times D$ with kernel G.
In particular, the abelian surface $(B \times C)/G$ is a product of elliptic curves.

Bob's ultimate isogeny diamond

Recall that in the SIDH setup we have $\#E_0(\mathbb{F}_{p^2}) = 2^{2e}3^{2f}$. Suppose we have an efficiently computable isogeny $\psi\colon E_0 \to C$ of degree $2^e - 3^f$, and that $\varphi_b = E_0 \to E_b$ decomposes into a composition of 3-isogenies as $\varphi_b = \phi_f \circ \phi_{f-1} \circ \cdots \circ \phi_2 \circ \phi_1$.

$$\begin{array}{ccccccccc} E_0 & \xrightarrow{\phi_1} & E_1 & \xrightarrow{\phi_2} & \cdots & \xrightarrow{\phi_{f-1}} & E_{f-1} & \xrightarrow{\phi_f} & E_f = E_b \\ \downarrow{\scriptstyle \psi} & & & & & & & & \\ C & & & & & & & & \end{array}$$

Consider the isogeny diamond determined by E_0, φ_b, ψ. We don't know φ_b, but we know $\varphi_b(P_a)$ and $\varphi_b(Q_a)$, and we can compute $\psi(P_a)$ and $\psi(Q_a)$, so we can compute generators for $G = \{(\varphi_b(P), \psi(P)) : P \in E_0[2^e] = \langle P_a, Q_a \rangle\}$, which is the kernel of a $(2^e, 2^e)$-isogeny from $E_b \times C \to E_0 \times D$ to a product of elliptic curves (one of which is E_0).

We can verify this by taking a walk in the $(2, 2)$-isogeny graph of abelian surfaces with kernel G and checking that we indeed hit a product of elliptic curves on the eth step.

Walking the $(2,2)$-isogeny graph using Richelot isogenies

We first apply gluing formulas of Howe–Leprévost–Poonen to compute quadratic polynomials $g_1, g_2, g_3 \in \mathbb{F}_{p^2}[x]$ such that the Jacobian of the genus 2 curve

$$X_1 \colon y^2 = f(x) = g_1(x)g_2(x)g_3(x)$$

is isomorphic to $(E_b \times C)/G_1$ where $G_1 := \langle (2^{e-1}\varphi_b(P), 2^{e-1}\psi(P)) : P = P_a, Q_a \rangle$. This is the first $(2,2)$-isogeny in our walk. Subsequent steps are computed using Richelot isogenies.

Index g_1, g_2, g_3 so that the roots of $g_1(x) = x^2 + g_{11}x + g_{10}$ and $g_2(x) = x^2 + g_{21}x + g_{20}$ define elements of $\mathrm{Jac}(X)$ that generate $G_2 := \langle (2^{e-2}\varphi_b(P), 2^{e-2}\psi(P)) : P = P_a, Q_a \rangle$. Let

$$\delta := \det \begin{pmatrix} g_{10} & g_{21} & 1 \\ g_{20} & g_{30} & 1 \\ g_{30} & g_{31} & g_{32} \end{pmatrix}$$

Provided $\delta \neq 0$, the next step in our walk is the Jacobian of $X_2 \colon y^2 = g_1'(x)g_2'(x)g_3'(x)$, where $g_i'(x) := \delta^{-1} \left(\frac{dg_j}{dx} g_k - g_j \frac{dg_k}{dx} \right)$ for $(i,j,k) = (1,2,3), (2,3,1), (3,1,2)$.
We will have $\delta = 0$ if and only if the next step in our walk is a product of elliptic curves.

We will necessarily hit $\delta = 0$ when we try to compute the eth step in our walk!
If this does not happen, the starting point of our walk must have been wrong.

Guessing Bob's penultimate isogeny diamond

Recall the diagram of isogenies

$$\begin{array}{ccccccccc} E_0 & \xrightarrow{\phi_1} & E_1 & \xrightarrow{\phi_2} & \cdots & \xrightarrow{\phi_{f-1}} & E_{f-1} & \xrightarrow{\phi_f} & E_f = E_b \\ \downarrow{\scriptstyle \psi} & & & & & & & & \\ C & & & & & & & & \end{array}$$

Let $\varphi_n := \phi_n \circ \phi_{n-1} \circ \cdots \circ \phi_1$ for $1 \leq n \leq f$, so that $\varphi_f = \varphi_b$. We don't know what ϕ_f is or what $\varphi_{f-1}(P_a)$ and $\varphi_{f-1}(Q_a)$ are, but there are only 4 possibilities, because there are only 4 cyclic 3-isogenies ρ from E_b, one of which is the dual isogeny $\hat{\phi}_f$, and $3\varphi_{f-1} = \hat{\phi}_f \circ \varphi_f$.

We can determine which ρ is correct by trying all 4: let $\varphi := \rho \circ \varphi_f$ and take a walk in the $(2,2)$-isogeny graph of abelian surfaces as above using the kernel

$$G = \langle (c\varphi(P), \psi_1(P)) : P = P_a, Q_a \rangle,$$

where c is the multiplicative inverse of 3 modulo 2^e and $\psi_1 \colon E_0 \to C_1$ has degree $2^e - 3^{f-1}$. With very high probability, we will find $\delta = 0$ on the eth step only when $\rho = \hat{\phi}_f$.

Lather, rinse, repeat

Having determined $\hat{\phi}_f$ we can compute its codomain E_{f-1} and $\varphi_{f-1}(P_a)$ and $\varphi_{f-1}(Q_a)$. We also know the most significant 3-adic digit of b: the kernel of ϕ_f is $\varphi_{f-1}(P_b + bQ_b)$.

If we replace $E_f, \varphi_f(Q_a), \varphi_f(P_a)$ with $E_{f-1}, \varphi_{f-1}(P_a), \varphi_{f-1}(P_a)$, equivalently, decrement f, we are right back where we started, but with a smaller value of f. Repeat until $f = 0$. We can thus compute Bob's secret b one 3-adic digit at a time.

We have intentionally glossed over an important point: how do we construct an efficiently computable $\psi\colon E_0 \to C$ of degree $d = 2^e - 3^f$?

- If $d = u^2 + v^2$ is a sum of two squares we can write it as $d = (u + iv)(u - iv)$. $\mathrm{End}(E_0)$ contains an endomorphism corresponding to i, so use $\psi = u + iv \in \mathrm{End}(E_0)$.
- If d is not a sum of two squares, try $d = 2^{e-\alpha} - 3^{f-\beta}$ with α, β small (works for SIKE).
- More generally, check if $d = 2^{e-\alpha} - 3^{f-\beta}$ is a product of small primes.
- More generally still, take Robert's approach: write d as a sum of four squares (always possible) and take walks in isogeny graphs of 8-dimensional abelian varieties.