IB Mathematics SL

Internal Assessment

Elliptical Curve cryptography

Contents

Aim and Rational

Have you ever sent an email, browsed the World Wide Web or sent a WhatsApp message? If you have, you have used Elliptical curve cryptography. Elliptical curve cryptography (ECC) is fundamental to these technologies which use ECC to encryption data that is being transmitted across the Internet.

This is an example of End-to-end encryption, meaning only the sender and reader are able to view the original message even when the data is transmitted across a public medium such as the airwaves Wi-Fi or LTE networks work on, which would otherwise be very insecure as anyone can intercept and read the data being transmitted. There are many different elliptical curves used in encryption and the majority of them have no proof they are secure, other than they have not yet been broken. With increased discussion regarding privacy and data security, and the world becoming more digital and connected than ever, it is important to understand the fundamental technologies that enable us to have private, secure interactions across the Internet and in cases such as online banking, secure access to our money and other digital information. As a computer science student, interested in pursuing a career in network security, but also as an everyday end user of these services and technologies, worried about the trust that is given to these technologies, I wish to explore ECC and the underlying math used in creating these encryption curves.

This paper aims to deliver an intuitive explanation of ECC and explore what makes different curves used in ECC more, or less secure.

The math behind Elliptical Curves

An Elliptic Curve is defined mathematically as;

$$E = \{(x, y) \mid y^2 = x^3 + ax + b\}$$

Where $a, b \in K$ and there is a point at infinity O

The examples of fields K can be are:
- Real numbers ($\mathbb{R}$)
- Rational numbers ($\mathbb{Q}$)
- Complex numbers ($\mathbb{C}$)
- Integers modulo p ($\frac{\mathbb{Z}}{p\mathbb{Z}}$)

In Figures 1.1-1.4 $a, b \in \mathbb{R}$.

The point at infinity O is used to limit the functions domain such that $x \leq O$.

Furthermore, there is a requirement that;

$$\Delta = -16(4a^3 + 27b^2) \neq 0$$

This is required to avoid singular points or in other words, isolated points. An example of a curve where the discriminant (Δ) $= 0$ is visible on Figure 1.3 where there is a singular point at x = 1.

Now that we know how they are defined, lets plot them in the $\mathbb{R}$ field.

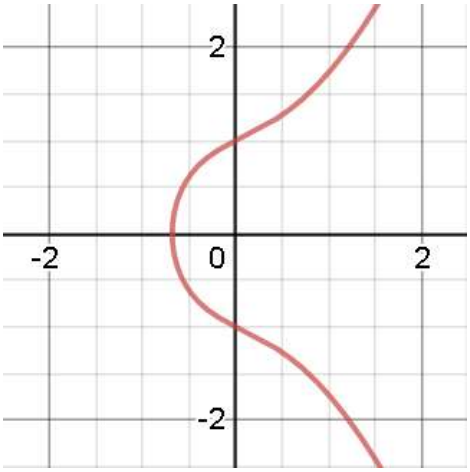

Figure 1.1: $y^2 = x^3 + x + 1$

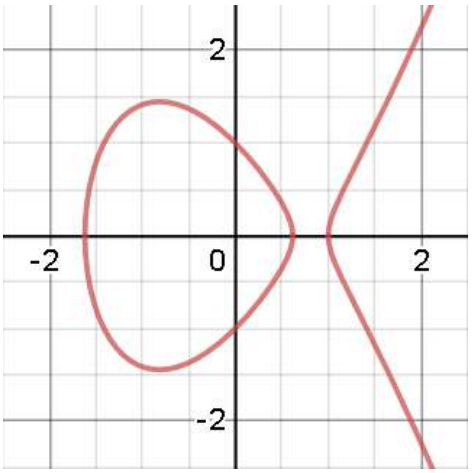

Figure 1.2: $y^2 = x^3 - 2x + 1$

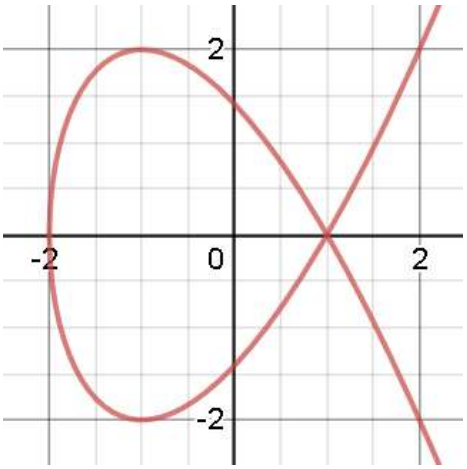

Figure 1.3: $y^2 = x^3 - 3x + 2$

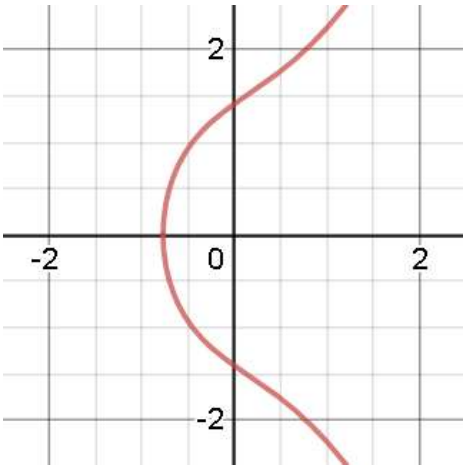

Figure 1.4: $y^2 = x^3 + 2x + 2$

Breaking down the relation

To understand the equation and why it is in the form it is, lets first rearrange:

$$y^2 = x^3 + ax + b \tag{1.1}$$

$$y = \sqrt{x^3 + ax + b} \tag{1.2}$$

Substituting in a = 1 and b = 1, to solve for y where x = 3,

For equation 1.1, we get:

$$y^2 = x^3 + \mathrm{x} + 1$$
$$y^2 = 3^3 + 3 + 1$$
$$y^2 = 31$$
$$\mathrm{y} = \pm\sqrt{31}$$

For equation 1.2, we get:

$$\mathrm{y} = \sqrt{x^3 + \mathrm{x} + 1}$$
$$\mathrm{y} = \sqrt{3^3 + 3 + 1}$$
$$\mathrm{y} = \sqrt{31}$$

Analyzing the two results, we can see equation 1.1 has a vertical symmetry where there are two points (x, y) and $(x, -y)$ while equation 1.2 only has the point (x,y).

This means that the domain of the two functions is the same. However, the range of equation 1.1 will be $\infty < y < \infty$, while the range of equation 1.2 will be $0 \leq y < \infty$.

Relating this back to ECC which uses the horizontal symmetry of the equation in its encryption, it is obvious why the equation is in the form of the equation 1.1 which is also known as Weierstrass equation.

The discriminant

Firstly, a reminder of what is the discriminant. The discriminant is used to find the number of roots an equation has, for example, the quadratic equation's $\Delta = b^2 - 4ac$. If the discriminant is 0, the equation has a double root and if it is positive, the quadratic equation has two real distinct roots.

As mentioned earlier, a requirement for Elliptic curves used in ECC is that the discriminant(Δ) is not equal to 0. (The reason for this requirement are explained later.)

As it is not obvious why the discriminant of the equation 1.1 is $\Delta = -16(4a^3 + 27b^2)$, let's look at how this discriminant is found and what it can tell us about the curve. In order to find the discriminant, we substitute 0 for y as the discriminant of the equation will tell us how many roots the equation has. When we substitute y = 0 in the equation 1.1 we get;

$$0 = x^3 + ax + b$$

Now we have a normal cubic equation. To find the solutions for this equation we use the cubic formula, similar to how we use the quadratic formula to find solutions to quadratic equations. Further, in the same way the quadratic equation has a discriminant, the cubic equation has a discriminant which is $\Delta = -16(4a^3 + 27b^2)$. Notice above the -16 in front of the discriminant of elliptical curves. This causes a sign change of the discriminant and is used only in cubic that use the Weierstrass equation form to simplify calculation when using large numbers. Further explanation is beyond this exploration, but it is important to understand that this only inverts the sign of the discriminant and increases it by a factor of 16, so we must also invert the signs for the number of roots table as well.

The Figures 1.1 – 1.3 are all examples of elliptical curves, however each of them has a different discriminant and therefore completely different shape. Calculating the discriminant for the Figures 1.1 – 1.3;

Δ of Figure 1.1:

$$y^2 = x^3 + x + 1$$
$$\Delta = -16(4(1)^3 + 27(1)^2)$$
$$\Delta = -16(4 + 27)$$
$$\Delta = -496$$

Δ of Figure 1.2:

$$y^2 = x^3 - 2x + 1$$
$$\Delta = -16(4(-2)^3 + 27(1)^2)$$
$$\Delta = -16(-32 + 27)$$
$$\Delta = 80$$

Δ of Figure 1.3:

$$y^2 = x^3 - 3x + 2$$
$$\Delta = -16(4(-3)^3 + 27(2)^2)$$
$$\Delta = -16(-108 + 108) = 0$$

Comparing the discriminants with the curves seen on the Figures above, we can see that when:

$\Delta < 0$, there is one unique root (x − intercepts) – seen on Figure 1.1

$\Delta > 0$, there are three unique roots – seen on Figure 1.2

$\Delta = 0$, there are two real roots – seen on Figure 1.3

Furthermore, we can see that in Figure 1.3 there is a repeated root. This repeated root of the curve removes the ability to use it for ECC which is why there is the requirement that $\Delta \neq 0$.

Translations

If we look at the two parameters a and b, let's explore what each of them does.

$$y^2 = x^3 + ax + b$$

If we look back at Figures 1.1 and 1.2 , the decrease in a caused a stretch in the horizontal axis, changing the x-intercepts and keeping the y intercepts the same.

Parameter b on the other hand, changes the y-intercept. This can be shown algebraically as;

$$y^2 = 0^3 + a0 + b = b$$

$$y = \sqrt{b}$$

This is seen on Figure 1.1 and 1.2 where b = 1 and the y-intercept = 1, and similarly on Figure 1.4 where b = 2 and y-intercept = $\sqrt{2}$.

Now that we understand the basic properties of the relation, we can start to explore Group Operations which are the processes done in ECC to encrypt data. Group Operations include: Addition, Point Doubling and Scalar Multiplication.

Group Operations

Group Operations are in essence, curve arithmetic's. For example, Addition is adding two points on the curve to get a third point. Elliptical Curves are great for curve arithmetic's because any line that passes through two points on the curve, in general, must pass through a third point (with some exceptions, for example in tangent lines, lines with a gradient of 0 and vertical lines).

Addition

As explained earlier, addition is adding two points, P and Q to get a third point P+Q.

This is conveyed in Figure 3.1 below. Figure 3.1 shows an example of addition where:

$y^2 = x^3 + 2x + 2$ and P(0,1.4) and Q(1, 2.2). As seen on Figure 3.1, The Point P+Q is the third intercept of the line connecting P and Q, reflected in the x-axis.

To calculate this algebraically, we first must find the equation of the line PQ, which is defined as:

$$y = sx + d$$

We start by finding the slope of the curve(s) using:

$$s = \frac{y_p - y_q}{x_p - x_q} Equation(3.1)$$

Then substituting in the values for P and Q we get:

$$s = \frac{1.4 - 2.2}{0 - 1} = 0.82$$

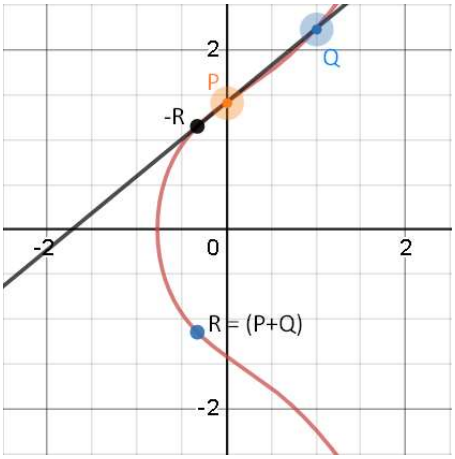

Figure 3.1 – Point Addition

Next, find the y-intercept d by substituting x_1 , y_1 and s into the equation $y = sx + d$ and solve for d. However, because P is a y-intercept we can skip that step and use y_p as d. Now we know the line of the Curve PQ is $y = 0.82x + 1.4$.

To find the third point we must find the third intercept of the two curves

$y = 0.82x + 1.4$ and $y^2 = x^3 + 2x + 2$.

In general terms, these two equations can be writen as $y = sx + d$ and y= $\sqrt{x^3 + ax + b}$.

Notice, $sx + d = \sqrt{x^3 + ax + b}$ can be re-written as $(sx + d)^2 = x^3 + ax + b$.

Making the equation equal to zero and expanding the brackets we get:

$$0 = x^3 + ax + b - s^2x^2 - 2sdx - d^2$$

Let x_P, x_Q, x_R be the set of solutions to this equation, where x_P and x_Q are the known roots at points P and Q.

From the factor theorem, we know that $(x - x_P)(x - x_Q)(x - x_R) = 0$. Explanation of the factor theorem is beyond this exploration, however given we know that there are three solutions to the equation above, we can factorize it as noted. This is similar to how quadratic equations with two roots can be factorized in the form $(x - x_1)(x - x_2) = 0$. By expanding this we get that;

$$(x - x_P)(x - x_Q)(x - x_R) = x^3 + x^2(-x_P - x_Q - x_R) + x(x_Px_Q + x_Px_R + x_Qx_R) - x_Px_Qx_R$$

Now if we equate the two expanded equations we get:

$$x^3 + ax + b - s^2x^2 - 2sdx - d^2$$

$$= x^3 + x^2(-x_P - x_Q - x_R) + x(x_Px_Q + x_Px_R + x_Qx_R) - x_Px_Qx_R$$

equating the factors of x^2; $\quad -s^2 = -x_P - x_Q - x_R$

solving for x_R: $\quad x_R = s^2 - x_P - x_Q \quad Equation(3.2)$

Then, we can use Equation 3.1, and substitute (x_P, y_P) for (x_R, y_R);

$$s = \frac{y_R - y_Q}{x_R - x_Q}$$

solve for y_R: $$y_R = s(x_R - x_Q) + y_Q \qquad Equation(3.3)$$

Now we use equations 3.2 and 3.3 to find point P+Q from our example above.

Where we had the elliptical curve $y^2 = x^3 + 2x + 2$ and the two points P(0,1.4) and Q(1,2.2) connected by the line $y = 0.82x + 1.4$.

$$x_R = 0.82^2 - 0 - 1 = -0.32$$

$$y_R = 0.82(-0.32 - 1) + 2.2 = 1.15$$

Knowing that $P + Q = -$R, we take the negative value of y_R, which is equal to -1.15. Now we have the point R which we were looking for at the point R$(-0.32, -1.15)$.

This is an example of curve arithmetic which is used in many encryption technologies. However, looking at Equation3.1, we can see a limitation of Group Addition given that $x_P \neq x_Q$ as that would make the denominator zero. This means that in order to find P+P = 2P, we must use another group operation -Multiplication.

Multiplication

As mentioned above, Multiplication is the process of finding the nP, for example P+P=2P.

at Point P. This is because where we used to have points x_P and x_Q, we now just have x_P. This means to find 2P, we must draw a tangent curve at point P and find the intercept with the elliptical curve which will yield -2P Then similar to addition, we must take the reflection of that point in the x-axis to find point 2P.

Algebraically this is done by first finding the gradient of the curve at point P by finding the first derivative of the elliptical curve

$$y^2 = x^3 + ax + b$$

To do this, we use implicit differentiation:

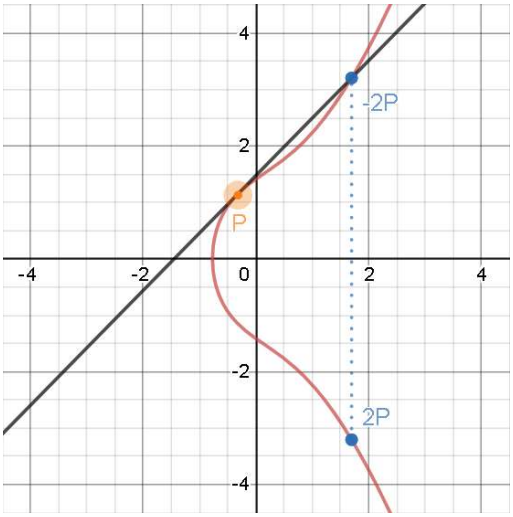

Figure 3.2 – Point Multiplication

This is shown graphically on Figure 3.2. Notice, the black line is tangent to the elliptical curve.

Starting by taking the derivative of both sides: $\frac{d}{dx}(y^2) = \frac{d}{dx}(x^3 + ax + b)$

LHS: $\frac{d}{dx}(y^2)$

Applying the chain rule: $\frac{dv(u)}{dx} = \frac{dv}{du} \cdot \frac{du}{dx}$

$v = u^2,\ u = y \quad | \quad \frac{dv}{du} = 2u, \frac{du}{dx} = \frac{d}{dx}(y)$

$$\frac{d}{dx}(y^2) = \frac{d}{du}(u^2)\frac{d}{dx}(y)$$

After substituting u: $\frac{d}{dx}(y^2) = 2y\frac{d}{dx}(y)$

RHS: $\frac{d}{dx}(x^3 + ax + b)$

$$\frac{d}{dx}(x^3 + ax + b) = 3x^2 + a + 0$$

Equating both sides: $2y\frac{d}{dx}(y) = 3x^2 + a$

Solving for $\frac{d}{dx}(y), we\ get\ that$: $\frac{d}{dx}(y) = \frac{3x^2 + a}{2y}$ $Equation(3.4)$

Now we can find the equation of the tangent line at point P(-0.33, 1.14):

$$y - y_P = m(x - x_P)$$

Substituting the values of the curve and Point P: $y - 1.14 = \frac{3(-0.33)^2 + 2}{2(1.14)}\big(x - (-0.33)\big)$

We get that the equation of the tangent at Point P is: $y = 1.02x + 1.48$

We now have only to find point -2P by finding the intercept, by doing the same steps as shown in Addition, but instead of x_Q, we use the same point x_P again. This means we can use the Equations 3.2 and 3.3 and substitute x_P instead of x_Q.

Therefore, we get that: $x_R = s^2 - 2x_P$ $Equation(3.5)$

And $y_R = y_P + s(x_R - x_P)$ $Equation(3.6)$

Substituting in the values of s, x_P and x_R,we find Point P+P $= -2P(1.69,3.2)$. We then reflect that point on the x-axis to find point 2P = $(1.69, -3.2)$. Comparing our result to Figure 3.2 we see that our answer is correct.

Until now, it was hard to see how these Group Operations were applicable to encryption, and now that we understand the basic operations, we can look at Scalar Multiplication. This is the actual operation used in ECC in the Diffie-Helman Key exchange protocol which is a technology that allows us to set up an encrypted connection with another party over an unsecure channel.

For now, we have used parameters a, b, field parameter K, and introduced the point at infinity O. All of these are public, and everyone can see them, yet when used correctly, they allow for encrypted communication. This is thanks to Scalar Multiplication which takes advantage of the Discrete

Logarithm problem (DLP) to create a "one-way function". The DLP says that it is easy to compute $Q = kP$ if we know either Q or P and k. However, given P and Q, it is computationally infeasible to compute k. This is why it is called a one-way function, easy to compute in one way, infeasible to compute backwards.

Scalar Multiplication

Scalar multiplication is simply repeated Addition of a point P, n times such that P+…+P= kP where k is the number of times the point is added to itself. This is done by finding R using Addition and using it as P and repeating the steps k times. Furthermore, to make this process more secure, rather than defining field K (mentioned at the beginning) as $\mathbb{R}$, we use $K(\mathbb{Z}/p\mathbb{Z})$.

This is done using a Generator Point defined as $G \in \mathbb{Z}/p\mathbb{Z}$. Which generates a Cyclical Group, meaning every any point in this group can be reached by scalar multiplication.

Now that we know all of the parameters, let's list them to understand how ECC works in practice.

Domain Parameters k(field modulo k), a, b(curve parameters), G(generator point) are public, meaning everyone using the same encryption agrees on these parameters and they are known.

Field modulo k simply means the value of y is the remainder after it has been divided by k. For example, 2 mod 5 evaluates to 2 and 5 mod 3 also evaluates to 2, given the remainder is two. This increases the level of encryption by creating a many-to-one function so it is difficult to work back to find the original values used in the encryption.

Now we can finally look at how ECC is done in practice.

Use the same equation as before but defining it in the field K mod 17 we get:

$$y^2 = x^3 + 2x + 2 \ (\text{mod} 17)$$

For simplicity, we will use Generator point G = (5,1).

Now we use the equations found in Multiplication to get that 2P. By substituting our parameters into the Equation 3.4 we get that:

$$\text{s}=\frac{3(5^2)+2}{2(1)} = 77 \times 2^{-1} = 9 \times 9 = 13(mod 17)$$

This might seem rather confusing so we should break it down.

77 mod 17 = 9, simply because remainder = 9 when 77 is divided by 17.

$2^{-1} = 9$ seems less obvious, but we know that, $2 \times 2^{-1} = 1$ and to get 1 in mod 17 we must divide by 18, given that $18\ mod\ 17 = 1$ then we do $18 \times 2^{-1} = 9$ therefore $2^{-1} = 9$ in the field mod 17.

Then we use the Equations 3.5 and 3.6 to find x_{2P} and y_{2P} which gives us:

$$x_{2P} = 13^2 - 2(5) = 16 - 10 = 6\ mod\ 17$$

$$y_{2P} = 13(5-6) - 1 = -13 - 1 = -14 = 3\ mod\ 17$$

Notice how $-14 = 3\ mod\ 17$given the first multiple of 17 smaller than -14 is -17 and the difference (remainder) is 3.

To generate the cyclical group, this process is repeated to find 3P, 4P and so on. This cyclical Group has already been calculated in the [1][1]. And shows in Figure 4.1:

2P = (5,1)+(5,1)=(6,3)	5P = (9,16)
3P = 2P+P = (10,6)	6P = (16,13)
4P = (3,1)	And so on… (Check Appendix for full table)

Figure 4.1 Cyclical Group

Now that we understand the math behind elliptical curves and curve arithmetic's, and have calculated a cyclical group, we can explore how this key-exchange encryption works in practice.

Applying Elliptical Curves in ECC encryption

Given two parties Alice and Bob who want to do a key exchange using ECC, they will each have their own private key α and β. Let's take $\alpha = 3$ and $\beta = 2$. Which they will never share.

Alice and Bob now compute public keys by computing $A = 3\text{P} = (10,6)$ and $B = 2\text{P} = (6,3)$.

Now they will exchange A and B. Note that this exchange is not encrypted and is public. Then Alice and Bob will use the other parties public key and compute $\alpha\text{B} = 3(2P) = 6P$ and $\beta A = 2(3P) = 6P$. Now they are both at the same point $6P = (16,13)$ without ever giving away their private keys $\{\alpha, \beta\}$.

This kind of encryption is known as asymmetric encryption, after they have found a shared x-coordinate, in our example 16, they can start using symmetric encryption. This is because if a third party, ex. Oscar, was intercepting this communication, they would know all the parameters and the public keys $A(10,6), B(6,3)$, however they will never know the private keys of Bob and Alice.

To understand why Oscar can never find this share point, we can first define that as point T where:

$$T_{AB} = \alpha\beta P = \alpha\text{B} = \beta\text{A}$$

Keeping in mind Oscar knows A, B and G, to solve this equation and find T_{AB}, Oscar would have to find the solution to either:

$$\alpha = \log_P A \ \ or\ \beta = \log_P B$$

which is called the discrete logarithm problem, and as of now, there is no way for solving this equation in a feasible amount of time, therefore there is no way for Oscar to find the point T_{AB}.

---

[1] Christof Paar, Jan Pelzl. "Understanding Cryptography", Accessed from http://swarm.cs.pub.ro/~mbarbulescu/cripto/Understanding%20Cryptography%20by%20Christof%20Paar%20.pdf on the 5th December 2019

Note that our example had small parameters for ease of calculation, therefore solving for α and β would not be difficult however, in reality the curves used have much larger prime numbers as parameters. For example, in Bitcoin, the curve used to generate keys is secp256k1 which is defined as $y^2 = x^3 + 7 \bmod p$, where p= $2^{256} - 2^{32} - 2^9 - 2^8 - 2^7 - 2^6 - 2^4 - 1$ which is a very large prime number. Notice how the parameters a and b of the curve are still small, however p is very large.

This is not the case in all elliptical curves, curve NIST P-224 was invented by the National Institute of Standards and Technology (NIST) and uses large numbers for a and b parameters, however according to safecurves.cr.yp.to , this curve is not secure enough to be used in ECC. It is defined as:

$y^2 = x^3 - 3x +$
18958286285566608000408668544493926415504680968679321075787234672564 mod p,
where $p = 2^{224} - 2^{96} + 1$.

Conclusion

As we can see that the magnitude of a and b is not as important, however the point Generator P must be large to make a curve secure. Which, looking at the Discrete Logarithm Problem, is logical, given that P is a variable in this equation and the larger it is, the more factors there are to compute to find a possible solution to the problem. While a and b are not in the DLP and they are simply used in the computing of the points, privately by the two parties. It is more important that the curve has the properties needed in ECC such as that it has a non-zero discriminant, and uses a field modulus of a large number, than its parameters being large.

Therefore, looking at ECC in terms of its application in computer science and networking, having the parameters a and b as small as possible will make calculating the slope, points and lines easier and more efficient, using less computational power. This is beneficial as those calculations are done by parties on their own devices which will often have little computing power (ex. phone).

On the other hand, the Generator Point P, must be large to increase the difficulty of the Discrete Logarithm Problem. Point P can be large as the cyclical group can be calculated by supercomputers and server farms with a lot of computational power as the cyclical group will be the same for everyone using the curve and so the private devices can use the solutions of these supercomputers to find the points.

Additionally, as computational power increased, the curves used in ECC simply increased their Generator Points P and size of private keys.

However, with the recent advances in quantum computing, which brings the possibility to potentially solve Discrete Logarithm problems in seconds, ECC and other technologies that use the Discrete Logarithm problem could soon face challenges and new forms of encryption will have to be developed. As increasing the difficulty of the problem will not be enough to keep the encryption safe, and our data private.

Bibliography

[1] Christof Paar (2009). “Understanding Cryptography A Textbook for Students and Practitioners”, Retrieved from http://swarm.cs.pub.ro/~mbarbulescu/cripto/Understanding%20Cryptography%20by%20Christof%20Paar%20.pdf on 10th October 2019

[2] https://safecurves.cr.yp.to/ Accessed on 10th October 2019

[3] Leijen, Daan (December 3, 2001). "Division and Modulus for Computer Scientists" (PDF). Retrieved from https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/divmodnote.pdf on 15th October 2019.

[4] https://hackernoon.com/what-is-the-math-behind-elliptic-curve-cryptography-f61b25253da3 Accessed on 1th October 2019

[5] Entrust inc. (2014), “Zero to ECC in 30 Minutes” Retrieved from https://www.entrust.com/wp-content/uploads/2014/03/WP_Entrust_Zero-to-ECC_March2014.pdf on 7th October 2019

Appendix

Full Cyclical Group of the $y^2 \equiv x^3 + 2 \cdot x + 2 \bmod 17.$ curve

$2P = (5,1) + (5,1) = (6,3)$
$3P = 2P + P = (10,6)$
$4P = (3,1)$
$5P = (9,16)$
$6P = (16,13)$
$7P = (0,6)$
$8P = (13,7)$
$9P = (7,6)$
$10P = (7,11)$
$11P = (13,10)$
$12P = (0,11)$
$13P = (16,4)$
$14P = (9,1)$
$15P = (3,16)$
$16P = (10,11)$
$17P = (6,14)$
$18P = (5,16)$
$19P = \mathcal{O}$

Retrieved from [1] pg. 246.