Dwight School
New York, New York

Group Theory and Modern Cryptography

To what extent does group theory provide a mathematical backdrop for modern cryptography?

340923 words Mathematics

Author : Jacob Bruner
Advisor: Daniel Bjelis

Contents

1 Foreword

This paper is for research purposes only. Although the information presented is as close to accurate as possible, one should never implement a cryptographic system themselves unless they know exactly what they're doing. **I am not liable for any damages caused in testing any cryptographic concepts referenced in this paper.** Cryptography's greatest weakness is human-error, and virtually *all* breaks in cryptographic security result from poor implementation.[1]

2 Introduction and Aim

2.1 Topic and Research Question

For my Extended Essay, I will be exploring the relationship between Group Theory, a mathematical field generalizing symmetry and codifying number systems, and modern cryptography. Cryptography has seen a number of advancements in the computer-age, most notable of which is the use of 'elliptic curves.' In this paper, I intend to show how a group-theoretical perspective on cryptography helps understand elliptic curve cryptosystems. My guiding question is: *to what extent does group theory provide a mathematical backdrop for modern public-key cryptography?*

2.2 Motivating Idea

It's undeniable that our modern-day world is reliant on cryptography. Every time a phone sends a text, a browser connects to a server, an email gets sent off, a monetary transaction is made, and much much more, our devices are, unbeknownst to us, performing many hundreds of math operations to ensure our data are 'encrypted.' But what does 'encryption' mean? Let's introduce some definitions. 'Encryption' is the process of disguising a message to be, loosely speaking, hidden to all *except* the intended recipient. This is the process of converting a 'plaintext' message into a jumbled 'ciphertext', which can be readily shared without risk of the sensitive message leaking. Converting a plaintext message (typically a string/list of characters) into a ciphertext is known as an 'enciphering' or 'encrypting' transformation. Likewise the reverse operation of recovering the plaintext message from a ciphertext is known as the *deciphering transformation.*[2] If we denote the plain and cyphertext $\mathcal{P}$ and $\mathcal{C}$ respectively and the enciphering map f and its inverse f^{-1} we obtain the following diagram:

[1] Harold Thimbleby. "Human factors and missed solutions to Enigma design weaknesses". In: *Cryptologia* 40.2 (2016), pp. 177–202. DOI: 10.1080/01611194.2015.1028680.

[2] Neal Koblitz. *A Course in Number Theory and Cryptography.* 2nd ed. Graduate Texts in Mathematics No. 114. Springer, 1994. ISBN: 9781461264422, p. 54.

$$\mathcal{P} \xrightarrow{f} \mathcal{C} \xrightarrow{f^{-1}} \mathcal{P}$$

2.3 The Fundamentals of Symmetric Cryptography

The intuitive way to implement this cryptosystem has the two transacting parties agree upon the nature of the map f in secret, beforehand. This might mean meeting up with a friend in-person to establish the common secret, f and f^{-1}, so that you could encrypt and decrypt each other's emails—fending off any prying eyes accessing their emails. This type of system has a special name, *"Symmetric Key Cryptography"*, reflecting the fact that both parties have the same shared secret (foreshadowing). Important historical examples include the 'Caeser Cipher' (supposedly invented by *the* Julius Caeser), where f is a shift operation that maps each letter to a new one a number of places ahead or behind. For instance, the two parties might decide beforehand the map f: *shift each character forward in the alphabet 3 letters*, implying the inverse map f^{-1}: *shift each character back 3 letters*. This might look like so: [0]

$$\mathcal{P} \sim A, B, C, D, E, F, ...Y, Z$$
$$\mathcal{C} \sim D, E, F, G, H, I, ...B, C$$

And, for example, if you wanted to encrypt the message $\mathcal{P}$ = "HELLO", you would obtain the ciphertext $\mathcal{C}$ = "KHOOR" which you would promptly send off for your friend to decode with the inverse, subtract-three-letters map. Note that, even in this simple example, repeated letters, word length, and other syntactic information provide a lot of information about the nature of the plaintext. Although this is a trivial example, schemes which, loosely speaking, encode any information about the syntax (or related notions) of the plaintext are usually highly vulnerable to a technique called *differential cryptanalysis*,which, at a high level, measures the change in cyphertext given a change in input, primarily targeting published symmetric-key protocols.[3]

To see what a caeser cipher looks like mathematically, we start encoding each letter as a number from 0, which is A, to 25, which is Z. Of course this depends on what alphabet one uses, if one chooses to include spaces, numbers, punctuation etc. We can represent the operation that takes a letter and maps it n places ahead with addition. Importantly, this operation

[0]The informed reader will notice that Ceaser Ciphers are really just linear transformations in disguise, provoking thought into whether more advanced techniques could be used. For instance, we could consider groups of two letters, called *digraphs*, resulting in f being a 2x2 invertible matrix, encoding the shift in a two-by-two matrix. Similarly, we could consider invertible *affine* transformations of the form $f : \vec{x} \to A\vec{x} + \vec{b}$, where A encodes a certain scaling factor.

[3]Koblitz, *A Course in Number Theory and Cryptography*, p. 56.

must 'wrap around' back to zero if you try to exceed 'z' in the alphabet. This process, known as *'modular arithmetic'*, is like circling a clock, where after reaching twelve, the hour hand wrap back around to 1, but we start from 0 instead of 1. So, in our case, shifting 'z' by 3 letters looks like so: $25 + 1 \bmod 26 \equiv 0$, which reads "25 plus 1 *is congruent to* 0 *modulo* (or *mod*) 26." With this in mind, we obtain for each letter $p \in \mathcal{P}$:

$$f(p) = p + n \bmod 26$$

Representing a shifting of each letter in the plaintext by n places in the alphabet. Now clearly this isn't a very sophisticated cryptographic scheme... For instance, performing a frequency analysis and comparing the most commonly occurring letters to those of the English alphabet, easily breaking these types of cyphers, broadly referred to as 'substitution ciphers.'[4] Modern schemes typically employ more resistant techniques, namely where changing just one letter of plaintext often yields a completely different ciphertext, making modern techniques all but impervious to frequency analysis. (Unless someone *really* screwed up an implementation)

For instance, AES encryption, part of the modern web standard, is an example of a '*block substitution cypher*,' which are beyond the scope of this paper. At a high level, it combines techniques similar to our Caesar Cipher with certain affine (think scaling and transforming) maps performed on blocks of plaintext. It's certainly more complicated than that, but essentially boils down to our system plus some advanced techniques making it resistant to cryptanalysis. (Like permutations, combinations, text look-up-tables and more.) In general, Symmetric-key cryptography (*viz.* predetermined, shared secret) is well understood. For instance, Claude Shannon (*the* definitive father of information theory) proved mathematically that the so-called *'one-time-pad'* encryption technique was completely and utterly unbreakable. In the general case, he showed that, if a random-generated key is at least as long as the plaintext (at least specifying a unique, random[0] character for each plaintext character and in the same alphabet), then performing a caeser cypher shift on each *individual* plaintext character by the value specified by the random character yields a cryptosystem that is mathematically impenetrable. Or equivalantly, performing a random modular addition on each character of the plaintext, with the shared secret being the sequence of random shifts.[5] Although modern systems seek smaller key

[4]Koblitz, *A Course in Number Theory and Cryptography*, p. 54.

[0]There is a paramount distinction between *random* and *psuedorandom*, the latter of which is vastly easier to implement on a computer. True randomness has to be derived from a non-computer source (for the most part). Commonly implemented approximates include taking the least-significant-digit of a mouse position, or the frequency of keypresses, etc. A one-time-pad–like scheme generated from a psuedorandom source *is* breakable, although usually not without some effort.

[5]Claude E. Shannon. *Communication Theory of Secrecy Systems.* Vol. 28. The Bell

sizes for performance reasons, this worst-case scenario should demonstrate the strength of symmetric-key algorithms in general. The 'one-time-pad' gets its name from its use in WWII when the KGB would distribute palm-sized pads with these one-time-keys and a table to ease in conversion. Such pads were often made of flammable materials to be burned with no trace.[6]

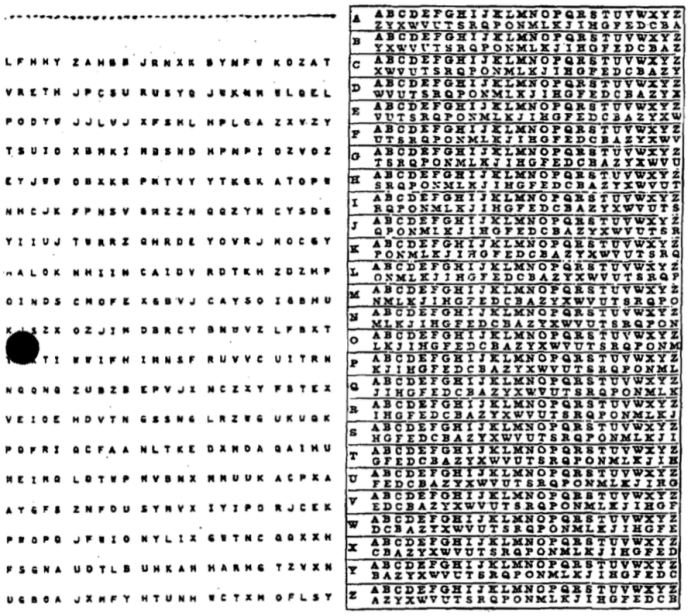


Figure 1: Format of a one-time-pad used by the NSA[7]

3 What is 'Public-Key' Cryptography

3.1 The 'Public-Key' Paradox

In the modern world, it's impractical to require that every shared secret be determined ahead of time. If a user wants to connect to a twitter server over a secure connection, how could symmetric-key encryption be employed? More generally, if two computers want to establish a connection for the first time, is there any way they could do so in an encrypted matter? The intuitive answer might be no, since how could one pass a shared secret without any man-in-the-middle being able to obtain that same key. But this defys the

System Technical Journal, 1949, pp. 656–715. DOI: 10.1002/j.1538-7305.1949.tb00928.x.

[6]Robert Edward Lewand. "The Perfect Cipher". In: *The Mathematical Gazette* 94.531 (2010), pp. 401–411. ISSN: 00255572. URL: http://www.jstor.org/stable/25759724 (visited on 05/21/2022).

ubiquity of encryption on the internet—every time I connect to a unfamiliar website, I still see a padlock on my browser. How could this be?

3.2 One-way Operations

Consider the intuitive fact that some operations are more difficult to do in reverse than fowards—certain 'one-way functions.' For instance, if I mixed two different-colored paints together and asked whether you, *a priori*, could deduce the two initial colors given the end result, would you be able to? Although its quite easy to check whether any two colors combine to match the end result, there isn't an easy operation that takes the end result and returns the initial colors. This happens to be true for a number of operations.

The foremost 'one-way operation,' employed to this day, is that of multiplying large numbers with minimal divisors. This makes sense from a practical standpoint; breaking up a large number into its factors usually boils down to computationally expensive trial divisions, to use a jargon, its *asymptotic time-complexity is exponential*, meaning that the time required to reach an answer scales exponentially with the size of the number. Algorithms like the general-number-field-Sieve quicken this slightly,[8] but regardless it is vastly faster to compute multiplication, the time complexity of which scales *at worst* with the square of the number.[9]

3.3 Diffie-Hellman Key Exchange

The Diffie-Hellman protocol was the first asymmetric protocol,[10] using the above intuition as a starting point. The end result of the key exchange is a commonly shared number, which can then function as the key for them to establish symmetric encryption.

Between two parties, Alice A and Bob B, the protocol first has them agree on a common 'generator' integer g, and this can be done over an eavesdropped channel. Each party chooses another *secret* integer in private: their private key. Say, Alice chooses a and Bob chooses b. Then, each party exponentiates the generator by their private key. So Alice and Bob compute g^a and g^b respectively. Each then sends this result to the other over the insecure channel. Finally, they both exponentiate the received numbers by

[8] Carl Pomerance. "A Tale of Two Sieves". In: *Notices of the AMS* 43.12 (1996), 1473–1485, p. 1482.

[9] Viktor Bunimov. "Area and time efficient modular multiplication of large integers". In: *Proceedings IEEE International Conference on Application-Specific Systems, Architectures, and Processors. ASAP 2003* (2003). DOI: 10.1109/ASAP.2003.1212863.

[10] Although this isn't a true "public-key" algorithm, rather its a non-authenticated key exchange, DH can be easily modified to send arbitrary messages in a scheme known as 'Elgamal encryption,' which is true "public-key" cryptography.
(Taher Elgamal. "A Public-Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms". In: *IEEE Transactions on Information Theory* 31.4 [1985], pp. 469–472. DOI: 10.1109/TIT.1985.1057074)

their private key. For example, Alice would receive g^b over the network, then compute$(g^b)^a$. By the commutativity of integer multiplication, multiplying g by itself b times *then* a times is the same as multiplying g by itself a times *then* b times. Explicitly, with Alice and Bob's private-keys in amber and blue respectively:

$$(g^a)^b = g^{ab} = (g^b)^a$$

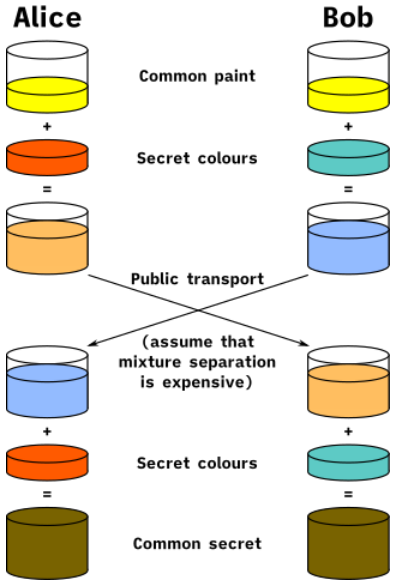


Figure 2: Diffie-Hellman protocol depicted visually, with mixing and separating paint replacing multiplication and factorization. (public domain)

The one caveat to the above description, however, is that these multiplications are performed *modulo* n for some integer n. As discussed before, this means numbers wrap around 'like a clock'; if g^a exceeds n, the remainder of $g^a \div n$ is returned. This should make sense, because computers have limited memory and because exponentiation by a (large) number yields a incomprehensibly large result—too large to send over a network or perform computation with. Additionally, computers can perform modular exponentiation *very* performantly by leveraging tricks in binary.[11]

Our relation from before looks like:

$$(g^a \bmod p)^b \bmod = g^{ab} \bmod p = (g^b \bmod p)^a \bmod p.$$

Importantly, though, Alice and Bob both agree on a common result $g^{ab} \bmod p$.

[11] Bunimov, "Area and time efficient modular multiplication of large integers".

In the real world, this type of Diffie-Hellman transaction is used by countless devices every second to allow secure connections over the internet. Browsers have a predetermined set of generators from the National Institute of Standards and Technology (NIST) which they use to establish a shared secret with a web-server which is then used to implement symmetric encryption to facilitate fast data transfer. Every time you are to connect to a `https://...` website, with the padlock in the corner, the client (you) and the server are performing this type of key exchange. Although, modern implementations might use a different 'number system' to perform the repeated multiplications...

3.4 A Proof that: $(a \bmod p)^k \bmod p = a^k \bmod p$

Here I offer a short proof verifying the ideas above. Given the linearity of modularity over multiplication:

$$a \times b \bmod n = (a \bmod n) \times (b \bmod n) \bmod n$$

I will show that $(a \bmod p)^k \bmod p = a^k \bmod p$ with induction.

Basis case, $P(1)$ or $k = 1$, holds since,

$$a^1 \bmod p = (a \bmod p)^1 \bmod p \qquad \text{(basis)}$$

Assume $P(k)$ holds for all k like so:

$$a^k \bmod p = (a \bmod p)^k \bmod p \qquad \text{(P(k))}$$

Now, multiplying $\bmod p$ both sides by $(a \bmod p)$ and applying the given:

$$\begin{aligned} a^k \bmod p \times a \bmod p \quad \bmod p &= (a \bmod p)^k \bmod p \times a \bmod p \quad \bmod p \\ (a^k \times a) \bmod p &= ((a \bmod p)^k (a \bmod p)) \bmod p \\ a^{k+1} \bmod p &= (a \bmod p)^{k+1} \bmod p \qquad \text{(P(k+1))} \end{aligned}$$

Hence, since $P(1)$ holds and $P(k) \Rightarrow P(k+1)$, the proposition holds for all positive integers. ■

4 An Introduction to Group Theory

Transitioning from the methods described earlier, I will now explore how many of these concepts fit into a group-theoretical perspective and how these tools can illuminate a path to inventing stronger cryptosystems.

4.1 The Group Axioms

A group is a set G, equipped with a binary operation mapping two elements to another of the form $* : \ G * G \to G$ such that the following conditions hold:[12]

Associativity
For all $a, b, c \in G, \ a * (b * c) = (a * b) * c$
Identity
There exists an $e \in G$ such that $a * e = e * a = a$ for all $a \in G$
Inverse
Each $x \in G$ has an inverse $x^{-1} \in G$ with $x * x^{-1} = x^{-1} * x = e$ [13]

Under this axiomatic definition of a group, a few canonical examples might come to mind.

For instance, the integers form a group under addition $(\mathbb{Z}, +)$. It's worth verifying for yourself that these do fit the definition of a group. Our identity element is 0, since $0 + a = a + 0 = a$ for all $a \in \mathbb{Z}$. Addition of integers is clearly associative. And, every number has a unique inverse, $a^{-1} := -a$.

Similarly, the real numbers (excluding 0) might form a group under multiplication, write $(\mathbb{R} \setminus \{0\}, \times)$ or $\mathbb{R}^{\times} \setminus \{0\}$. We verify it has a two-sided identity, 1; that its operation is associative; and that every element has an inverse $a^{-1} := \frac{1}{a}$, since $a \times \frac{1}{a} = \frac{1}{a} \times a = 1, \ \forall a \in (\mathbb{R} \setminus \{0\}, \times)$.

To help intuition,consider that the integers $(\mathbb{Z}, \times)$ *do not* form a group under multiplication, since inverses are not suitably defined, i.e., there is no integer a such that $a \times 2 = 2 \times a = 1$.

4.2 The Case for Commutativity and Abelian Groups

One key property that we didn't require in our group axioms is *commutativity*. At first glance, this might seem like an obvious condition, since many of the introductory examples do obey commutativity. But most of the insight that comes from group theory is from the study of non-commutativity. This distinction happens to be so important that commutative groups have a special name: *abelian groups*.[14] Likewise, groups that violate this condition are sometimes called *non-abelian groups*. In general, non-abelian groups can be thought of as corresponding to symmetries that change the backdrop for another symmetry to occur. For instance, the group of symmetries on a

[12] Dan Saracino. *Abstract Algebra*. 2nd ed. 2008. ISBN: 1577665368, p. 16.

[13] Note that uniqueness of inverses is not given in the axioms for a group, since it follows from the Identity and Associativity requirement and the existence of a (not strictly unique) inverse. If b, c are left and right inverses respectivly of an element a with identity 1, then considering $c = 1 * c = (b * a) * c = b * (a * c) = b * 1 = b$, gives us $b = c$ as required.

[14] Richard M. Foote David S. Dummit. *Abstract Algebra*. 3rd ed. Wiley, 2004. ISBN: 9780471433347, p. 17.

square, D_8, is non-abelian (meaning non-commutative). If you labeled the verticies of a square, you'd find performing a rotation and than a reflection is not, in general, the same as performing a reflection than a rotation[0]. Despite this, many of the examples we have and will look at happen to be abelian. For instance, the additive and multiplicative (sans 0) groups of $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$ and $\mathbb{C}$, are 'abelian.' To highlight the importance of commutativity, consider that many properties of electric charge, as described by the Standard Model (in Quantum Field Theory), follow directly from the commutativity of the circle group $U(1)$ or T.[15] But this is far beyond the scope of basic group theory.

4.3 Applying Group Theory to Cryptography: Cyclic Groups

Much of the groups that are of importance to cryptography are abelian groups. For intuition on why this may be the case, consider the Diffie-Hellman example before. Both parties combined a series of steps involving their generator and private keys. If the underlying multiplicative group were not abelian, $(g^a)^b$ doesn't necessarily equal $(g^b)^a$. Similarly, in the case of a one-time-pad, if our modular addition weren't commutative, our results would be unpredictable. This provokes the question, were these cases abelian groups?

The group of integers under addition *modulo* n is denoted $(\mathbb{Z}/n\mathbb{Z}, +)$ or just $\mathbb{Z}/n\mathbb{Z}$ and is called the 'cyclic group of order n.' In an earlier example, we performed operations in $(\mathbb{Z}/26\mathbb{Z}, +)$ so as to not surpass the alphabet. The notation here is actually meaningful. Without delving too far into 'quotient groups' and 'normal subgroups,' the slash corresponds to 'quotienting' out the integers by p times the integers, creating a group where 1 is indistinguishable from $p+1$ and indeed $2p+1, 3p+1, -p+1, -2p+1...$, called an *equivalence relation.*[16] This formalizes the impercise notions of 'going around a clock face.'

Contrastingly, performing multiplication in a modular fashion forms the group $(\mathbb{Z}/n\mathbb{Z}, \times)$ or $(\mathbb{Z}/n\mathbb{Z})^\times)$. This group behaves a little differently and consists of the elements $\{0, 1..., n-1\}$ that are relatively prime to n, i.e., share no nontrivial divisors. In this multiplicative group, it takes more care to define inverses than in the more familiar examples. Inverses can be computed using the Euclidean Algorithm or by leveraging the symmetries in modular multiplication. Essentially, in $(\mathbb{Z}/n\mathbb{Z})^\times)$, the multiplicative inverse of a number n is a number m such that,

$$nm \equiv 1 \mod p$$

[0] If we were working with the presentation of D_8, $\langle a, x \mid a^4 = x^2 = e,\ xax^{-1} = a^{-1} \rangle$, we would say that x and a don't commute—corresponding to the reflection and rotation generators respectively.

[15] **weinberg**.

[16] Saracino, *Abstract Algebra*, p. 104.

Now, proving that the elements are only those relatively prime to n, consider that the invertible elements of $(\mathbb{Z}/n\mathbb{Z})^\times$) (for any q) are exactly those that are relatively prime to q. The proof is by contradiction:

If $\gcd(n, q) = d$ (i.e., they share a nontrivial divisor) with $d \neq 1$, assume there is some m for which $nm \equiv 1 \mod p$. Expanding this out,

$$\begin{aligned} nm &\equiv 1 \mod p \\ nm - 1 &\equiv 0 \mod p \\ \Rightarrow\ & p|nm - 1 \\ \Rightarrow\ & d|nm - 1 \\ \Rightarrow\ & d|nm - 1 - nm \text{ (since } d|nm) \\ \Rightarrow\ & d|1 \end{aligned}$$

As 1 is not divisible by anything except 1, our assumption was false and n cannot have an inverse and is not $\in (\mathbb{Z}/n\mathbb{Z})^\times$).[17] ■

Notice that for prime p, every number $< p$ is relatively prime to p, so the order of the group $(\mathbb{Z}/n\mathbb{Z})^\times$ is $p-1$, which you might notice is the same as the order of $(\mathbb{Z}/n\mathbb{Z})^+$. This fact will be useful in a bit.

4.4 Discrete Logarithm Problem

In this section, I want to explore the formal statement of 'why' certian operations are difficult to reverse, although In a Diffie-Hellman key exchange over $(\mathbb{Z}/q\mathbb{Z})^\times$, Alice and Bob were using the map f,

$$f(n) = g^n \mod q$$

or, written in terms of their respective groups,

$$f : \mathbb{Z} \to (\mathbb{Z}/q\mathbb{Z})^\times$$

This f is a function between groups (a *group homomorphism*) from $\mathbb{Z}$ *onto* a ***subset*** of $(\mathbb{Z}/q\mathbb{Z})^\times$ surjectively, namely the subset generated by g^n which I will denote H. The inverse mapping here (the 'computationally difficult' part), isn't defined for the entirety of dom f, since f takes an infinite group $\mathbb{Z}$ to a subset H of the finite group $(\mathbb{Z}/q\mathbb{Z})^\times$.

Roughly sketching the 'why' behind what is to follow, it can be shown that any finite abelian group must consist of cyclic groups as building blocks, cyclic groups being the integers under addition modulo some n. [18] Hence, our subset H (which one can prove forms a group as well, a 'subgroup') must be one of these cyclic groups. We can then write this sort-of inverse map f^{-1} as,

[17] Koblitz, *A Course in Number Theory and Cryptography*, p. 19.

[18] By the (aptly named) Fundamental Theorem of Finitely Generated Abelian Groups. (Saracino, *Abstract Algebra*, p. 134)

$$\log_g : H \to (\mathbb{Z}/m\mathbb{Z}, +)$$

Ascertaining this 'hidden' cyclic subgroup H is known to be a difficult problem, and this difficulty of computing the original integer n is infeasible since $\log_g$ does not map back to all the integers: just some additive group of the integers modulo some m. This is known as the *discrete logarithm problem.*[19] Like discussed before with division, there is no known algorithm that can solve a discrete logarithm with time-complexity that scales polynomially (on a non-quantum computer). It remains an open question in theoretical computer-science whether such a polynomial-time algorithm exists, though all of modern cryptography rests on this principle.

5 Group Law on an Elliptic Curve

5.1 Definition of an Elliptic Curve

In a complete turn of events (that will hopefully make sense in the end), I want to explain the notion of *Elliptic Curves.* Elliptic curves are a family of algebraic[20] curves defined as the solution set of a polynomial equation. Namely, an elliptic curve is the set of solutions (x, y) to an equation of the form:[21] [22]

$$y^2 = x^3 + ax + b, \text{ for constants } a, b.$$

Importantly, the behavior of this curve is dependent on the ground 'field' on which one defines it. We will look at the familiar case of the reals shortly, but this is an important distinction, hinting towards the possibility of a discrete context, or field, to come. Field, in mathematics, refers to a structure where any two elements have a well-defined notion of addition, subtraction, multiplication and division (except by a 0 element). The most important examples are the rationals $\mathbb{Q}$, the reals $\mathbb{R}$, and the complex numbers $\mathbb{C}$. Not the integers since, take, for example: $1, 2 \in \mathbb{Z}$, but $\frac{1}{2} \notin \mathbb{Z}$. We will begin with the case of an elliptic curve over the reals.

5.2 Elliptic Curves over the Reals

In 3, we see the general shape of an elliptic curve, usually with a sort-of bulbous head on the left. This complicated shape arises from square rooting a general depressed cubic of the form $y = x^3 + px + q$. Without the x^2 term, this cubic is guaranteed to have only one inflection point at $x = 0$,

[19] Koblitz, *A Course in Number Theory and Cryptography*, p. 97.

[20] Specifically a dimension one algebraic variety since it admits a group structure.

[21] Koblitz, *A Course in Number Theory and Cryptography*, p. 10.

[22] The general form of an elliptic curve $y^3 + a_1xy + a_2y = x^3 + a_3x^2 + a_4x + a_5$ can always be transformed in a series of scalings and rotations to the reduced form preserving the group structure as long as the field is of characteristic not 2 or 3.

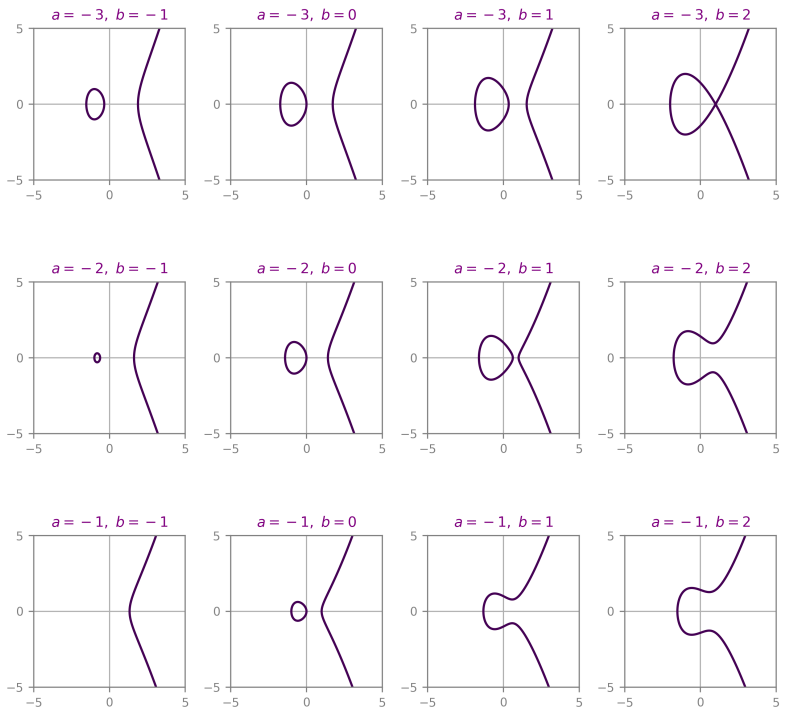


Figure 3: A few selected elliptic curves plotted with matplotlib over the reals with a from -3 to 0 and b from -1 to 2.

since its second derivative $y'' = 6x$ does not depend on p or q. Since square roots preserve the inflection points, we can see the connected graphs in 3 all have inflections at zero. Furthermore, the process of square-rooting increases values from $0 < y < 1$ and decreases values when $y > 1$ with a fixed point at 1. Over $\mathbb{R}$, square roots are limited to inputs greater than 0, but since we take both branches ($\pm$) of the root, we see reflection symmetry across the y-axis (this will be important in a moment). To illustrate the relationship:

5.3 Geometric secant-tangent construction

It turns out, Elliptic curves are special because one can define a reasonable notion of addition with points on the curve. The one caveat is that we

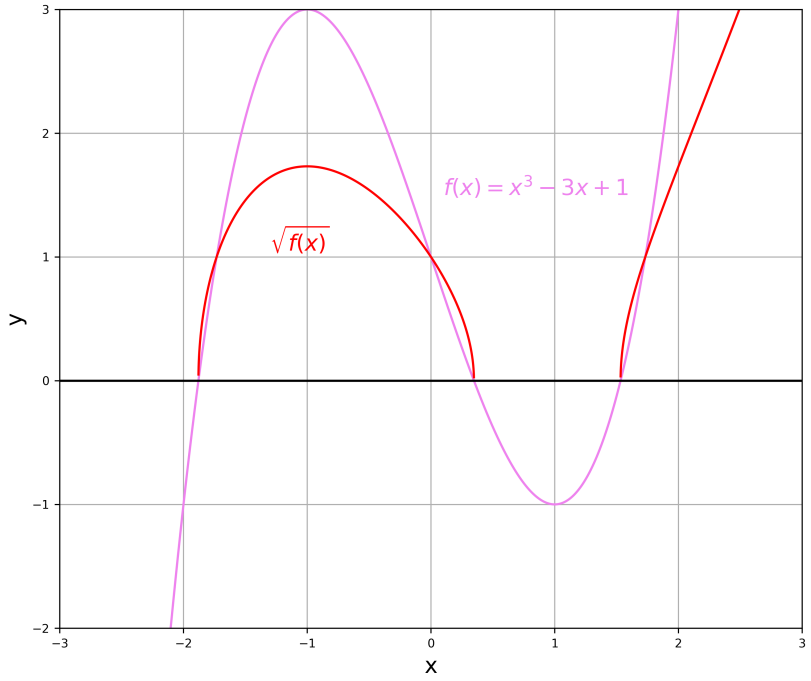


Figure 4: A depiction of how one branch of an elliptic curve arises from square-rooting a depressed cubic.

require a point 'at infinity' I to do so. Formally, this means were working in the projective real plane $\mathbb{RP}^2$ or $\mathbb{R}^2 \cup \{\infty\}$. But with this, we obtain the property that any secant drawn through two points on the curve must intersect another third point on the curve or this point at infinity I. Thus we will define the addition of two points P and Q to be this third point reflected across the y-axis.[23]

[23]Note, without reflecting across the y-axis, this operation would not be associative. It also makes sense with the notion that three colinear points sum to the point-at-infinity $A + B + C = I$ and hence, $A + B = -C$.

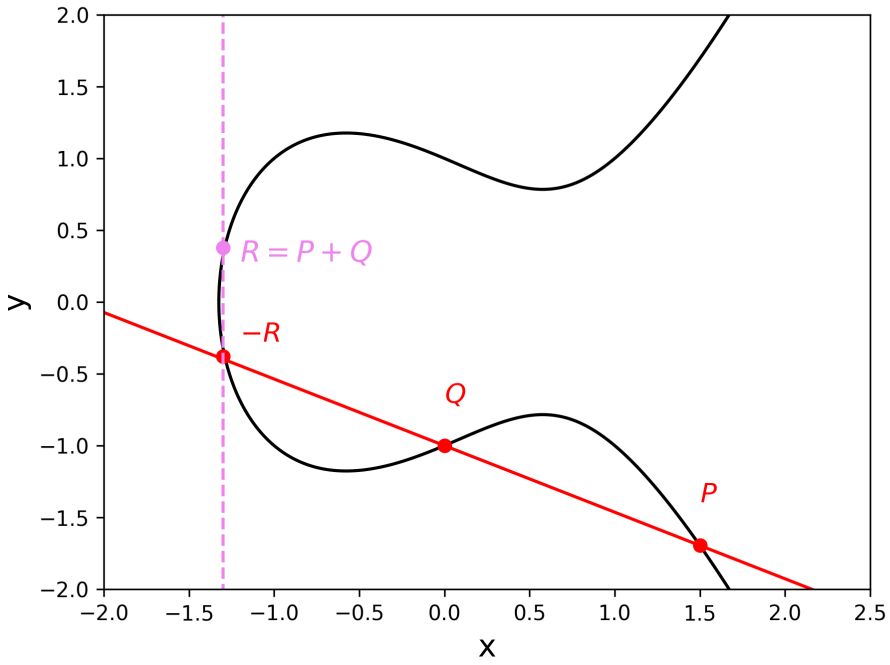


Figure 5: Point addition of two points P and Q, defined by reflecting the third point of intersection across the y-axis, on $y^2 = x^3 - x + 1$.

5.4 Addition of two points

We can make this idea rigorous by using some geometric tools. Our third point $-R$ is defined as the intersection of our elliptic curve $y^2 = x^3 + ax + b$ and some secant line $y = mx + l$. Given the points $P, Q = (x_p, y_p), (x_q, y_q)$ we can write down the slope:

$$m = \frac{y_p - y_q}{x_p - x_q}$$

Proceeding with the simultaneous equations

$$y^2 = x^3 + ax + b \tag{1}$$
$$y = mx + l \tag{2}$$

Substituting (2) into (1),

$$\begin{aligned} (mx + l)^2 &= x^3 + ax + b \\ m^2x^2 + 2mlx + l^2 &= x^3 + ax + b \\ \Rightarrow 0 &= x^3 - (m^2)x - (2ml)x + (b - l^2) \end{aligned} \tag{3}$$

Now, this cubic has three real solutions, namely the roots of P, Q and R respectively. Writing these as factors:

$$(x - x_p)(x - x_q)(x - x_r) = x^3 - (m^2)x^2 - (2ml)x + (b - l^2)$$

The next step is to expand the left side out and equate coefficients.

$$\begin{aligned} LH &= x^3 - (x_p + x_q + x_r)x^2 + (x_p x_q + x_p x_r + x_q x_r)x - x_p x_q x_r \\ RH &= x^3 \qquad\qquad - (m^2)x^2 \qquad\qquad\qquad - (2ml)x + (b - l^2). \end{aligned}$$

This means, using the purple coefficients,

$$\begin{aligned} m^2 &= x_p + x_q + x_r \\ \therefore \quad x_r &= m^2 - x_p - x_q. \end{aligned} \tag{4}$$

Substituting the x coordinate of $-R$, x_r into the point-slope form of the line with $m = \frac{y_p - y_q}{x_p - x_q}$ we obtain,

$$\begin{aligned} y_r &= y_q + m(x_r - x_q) \\ &= y_p + m(x_r - x_p). \end{aligned}$$

Flipping our result across the y-axis to obtain R, we have (finally),[24]

$$\begin{aligned} P \ + \ Q \ &:= \ R \\ (x_p, y_p) + (x_q, y_q) &:= \left(m^2 - x_p - x_q, \quad -y_q - m(x_r - x_q)\right), \\ &\text{where } m = \tfrac{y_p - y_q}{x_p - x_q}. \end{aligned}$$

■

5.5 Inverse Points and Infinity

In the case of adding two points opposite each other over the y-axis, one says they intersect the third point "at infinity." In general, over a projective plane, we revise the parallel postulate to say "any two parallel lines share one point of intersection, namely the point at infinity." We see this below; like train tracks meeting the horizon, the two parallel, vertical lines meet at ∞ or I.

Because of this construction, I acts as a left- and right-sided identity element. Since, for any ($\mathbb{R}$-rational) point A with inverse $-A$ on the elliptic-curve,

$$\begin{aligned} A + (-A) &= I \\ A + (-A) + A &= I + A \\ \Rightarrow \quad A + I &= I + A = A \end{aligned} \tag{5}$$

[24] As long as $P \neq Q$ and $P \neq -Q$.

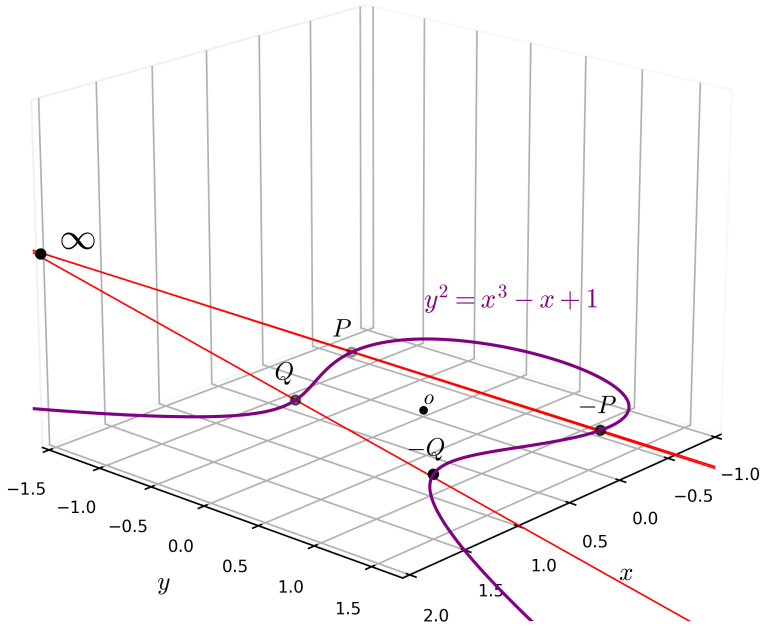


Figure 6: Parallel lines drawn through inverse points, all intersecting at point at infinity I over $y^2 = x^3 - x + 1$. As in, $P + (-P) = I = Q + (-Q)$.

5.6 Adding a point to itself

The case that was left out of the earlier pictorial representation of point addition is when a point is added to itself. Not to fear, however, since we can replace the secant through two points with a tangent through one. The resultant point is the resultant intersection flipped across the y-axis. The slope of this tangent can be found through implicit differentiation.

$$\begin{aligned} d(y^2) &= d(x^3 + ax + b) \\ 2ydy &= 3x^2dx + adx \\ \frac{dy}{dx} &= \frac{3x^2 + a}{2y}. \end{aligned}$$

Thus with,

$$y_r - y_p = m(x_r - x_p)$$

$$\text{where } m = \tfrac{3x_p^2+a}{2y_p},$$

we can use the same result as before with $P = Q$:

$$\begin{aligned} m^2 &= x_p + x_q + x_r \\ \Rightarrow \quad x_r &= m^2 - 2x_p \end{aligned} \tag{6}$$

and, using the point-slope form and flipping across the y-axis,

$$y_r = -y_p - m(x_r - x_p)$$

■

*A word on notation: the choice to use + to denote this binary operation is entirely arbitrary, and many authors use alternate notation to represent this ($\times, \oplus, \otimes, \star$ etc.). Using + hints that adding a point P to itself n times might be written nP, but for sake of consistency with the theory to come, I will write this P^n meaning $P + P + ... + P$, n-times. This will make sense in a moment.

6 Group structure of $E[K]$

6.1 Verifying the group axioms

By the higher powers of mathematical serendipity, this forms a group! A very special group at that. We can write this elliptic curve group as $E[\mathbb{R}]$ for the case in the reals. Or, $E[K]$ for any field K (again, a structure with $+, -, \times, \div$).

Now, we will (at least, heuristically) verify that the group axioms hold, we first see in eq (5) that we have a two-sided identity, I, given by the point at infinity. Hence,

$$P + I = I + P = I \text{ for all } P \in E[K]$$

Next, we prove the existence of an inverse for every element by noticing that our curve is symmetric about the y-axis. Similarly our identity element I is self-inverse. Hence,

$$\text{for all } P \in E[K], \text{ there exists an inverse } -P \text{ such that } P + (-P) = I$$

This operation is associative as well, but unfortunately a rigorous proof of this fact is unprecedentedly untame, especially through this geometric depiction of elliptic curves. With more powerful tools from algebraic geometry, this proof is almost trivial, following nicely from tracking values on points of a curve (namely $\text{Div}^0(E)$) and the relationship between poles and zeros, as stated by the Riemann-Roch theorem.[25] Lastly, this binary operation of

[25] Joseph H. Silverman. *The Arithmetic of Elliptic Curves*. Graduate Texts in Mathematics №106. Springer, 1986. ISBN: 9780387962030, p. 66.

adding two points is *commutative* as well, following geometrically from the secant construction or through the formulas.

As a result of meeting these conditions, $E[K]$ forms an abelian group under our binary operation.

6.2 Finite Fields $\mathbb{F}_p$

This elliptic curve group action holds for other fields K as well. Most important to cryptography is elliptic curves over "finite" fields. A finite field, denoted $\mathbb{F}_p$ for p prime,[26] is a set of ordered pairs (x, y) with addition, subtraction and multiplication taken *modulo* p.

As I alluded to earlier, the condition that p be prime is so that every element $n < p$ has a multiplicative inverse, since, to have a multiplicative inverse, $\gcd(n, p) = 1$. Really, finite fields are the elements of $\mathbb{Z}/p\mathbb{Z}$ with addition and multiplication in $(\mathbb{Z}/p\mathbb{Z})^+$ and $(\mathbb{Z}/p\mathbb{Z})^\times$. The field axioms readily follow from the group axioms of the multiplicative and additive groups.

6.3 Elliptic curves over finite fields, $K = \mathbb{F}_p$

Using the same equations from before, except with operations taken *modulo* p, this elliptic curve still forms a group over a finite field, $E[\mathbb{F}_p]$. The elements of this group are the points,

$$E[\mathbb{F}_p] = \left\{(x, y) \mid y^2 \equiv x^3 + ax + b \bmod p\right\} \cup \{\infty\}.$$

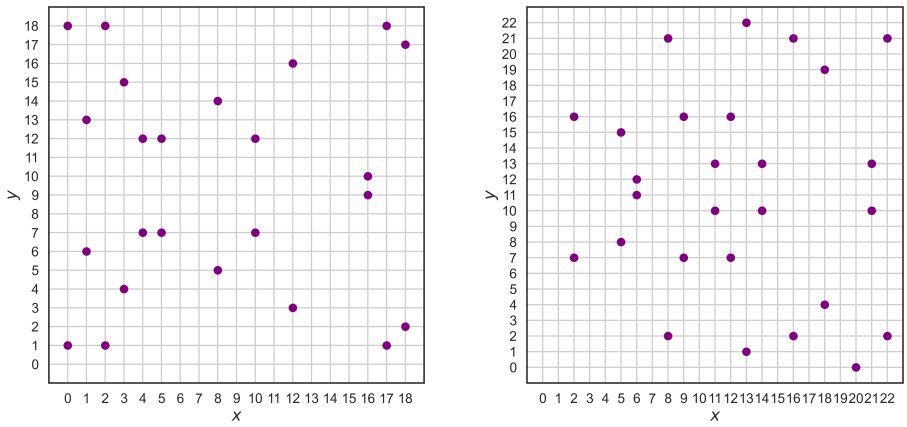
with the additional ∞ element serving as the identity.

Plotting examples of these groups over finite fields, the geometric intuition breaks down somewhat.

7 Conclusion

7.1 Limitations and Reflection

[26] Technically, p can be a prime power. One might write $\mathbb{F}_q$ where $q = p^r$.



(a) $y^2 = x^3 + 15x + 1$ over $\mathbb{F}_{19}$ (b) $y^2 = x^3 + 12x + 17$ over $\mathbb{F}_{23}$

Figure 7: Select elliptic curves plotted over finite fields. Notice the symmetry about $\frac{p}{2}$.

Bibliography

A History of U.S. Communication Security. Vol. 1. The David G. Boak Lectures. National Security Agency, 1973. URL: https://www.governmentattic.org/18docs/Hist_US_COMSEC_Boak_NSA_1973u.pdf.

Aschbacher, Michael. *The Status of the Classification of the Finite Simple Groups*. Vol. 51. Notices of the American Mathematical Society 7. American Mathematical Society, 2004, pp. 736–740. URL: https://www.ams.org/notices/200407/fea-aschbacher.pdf.

Bunimov, Viktor. "Area and time efficient modular multiplication of large integers". In: *Proceedings IEEE International Conference on Application-Specific Systems, Architectures, and Processors. ASAP 2003* (2003). DOI: 10.1109/ASAP.2003.1212863.

David S. Dummit, Richard M. Foote. *Abstract Algebra*. 3rd ed. Wiley, 2004. ISBN: 9780471433347.

Elgamal, Taher. "A Public-Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms". In: *IEEE Transactions on Information Theory* 31.4 (1985), pp. 469–472. DOI: 10.1109/TIT.1985.1057074.

Koblitz, Neal. *A Course in Number Theory and Cryptography*. 2nd ed. Graduate Texts in Mathematics No. 114. Springer, 1994. ISBN: 9781461264422.

Lewand, Robert Edward. "The Perfect Cipher". In: *The Mathematical Gazette* 94.531 (2010), pp. 401–411. ISSN: 00255572. URL: http://www.jstor.org/stable/25759724 (visited on 05/21/2022).

Pomerance, Carl. "A Tale of Two Sieves". In: *Notices of the AMS* 43.12 (1996), 1473–1485.

Saracino, Dan. *Abstract Algebra*. 2nd ed. 2008. ISBN: 1577665368.

Shannon, Claude E. *Communication Theory of Secrecy Systems*. Vol. 28. The Bell System Technical Journal, 1949, pp. 656–715. DOI: 10.1002/j.1538-7305.1949.tb00928.x.

Silverman, Joseph H. *The Arithmetic of Elliptic Curves*. Graduate Texts in Mathematics №106. Springer, 1986. ISBN: 9780387962030.

Thimbleby, Harold. "Human factors and missed solutions to Enigma design weaknesses". In: *Cryptologia* 40.2 (2016), pp. 177–202. DOI: 10.1080/01611194.2015.1028680.