Lectures on Applied Algebra II

audrey terras

July, 2011

Part I
Rings

1 Introduction

This part of the lectures covers rings and fields. We aim to look at examples and applications to such things as error-correcting codes and random number generators. Topics will include: definitions and basic properties of rings, fields, and ideals, homomorphisms, irreducibility of polynomials, and a bit of linear algebra of finite dimensional vector spaces over arbitrary fields. Our favorite rings are the ring $\mathbb{Z}$ of integers and the ring $\mathbb{Z}_n$ of integers modulo n.

Rings have 2 operations satisfying the axioms we will soon list. We denote the 2 operations as addition $+$ and multiplication $*$ or $\cdot$. The identity for addition is denoted 0. It is NOT assumed that multiplication is commutative. If multiplication is commutative, then the ring is called commutative. A field F is a commutative ring with an identity for multiplication (called $1 \neq 0$) such that F is closed under division by non-zero elements. Most of the rings considered here will be commutative. We will be particularly interested in finite fields like $\mathbb{Z}_p$, where p =prime. You must already be friends with the field $\mathbb{Q}$ of rational numbers; i.e., fractions with integer numerator and denominator. And you know the field $\mathbb{R}$ of real numbers from calculus; i.e., limits of Cauchy sequences of rationals. We are not supposed to say the word "limit" in these lectures as this is algebra. So we will not talk about constructing the field of real numbers.

Historically, much of our subject came out of number theory and the desire to prove the Fermat Last Theorem by knowing about factorization into irreducibles in rings like $\mathbb{Z}[\sqrt{-m}] = \left\{a + b\sqrt{-m} \mid a, b \in \mathbb{Z}\right\}$, where m is a non-square integer. For example, it turns out that, when $m = 5$, we have 2 different factorizations: $2 \cdot 3 = \left(1 + \sqrt{-5}\right)\left(1 - \sqrt{-5}\right)$. So the fundamental theorem of arithmetic is false for $\mathbb{Z}[\sqrt{-5}]$.

Algebraic number theory concerns fields like $\mathbb{Q}$ and rings like $\mathbb{Z}$ as well as fields like $\mathbb{Q}[\sqrt{-m}] = \left\{a + b\sqrt{-m} \mid a, b \in \mathbb{Q}\right\}$, where m is a square-free positive integer, and the corresponding ring of algebraic integers (which is $\mathbb{Z}[\sqrt{-m}]$ only when $m \equiv 2$ or $3 (\mathrm{mod}\, 4)$). The definition of ring of integers in an algebraic number field makes the ring of integers associated to $\mathbb{Q}[\sqrt{-m}]$ in the remaining case somewhat larger by allowing denominators of 2. (Thus you, in fact, have unique factorization into primes for the ring of integers in $\mathbb{Q}[\sqrt{-3}]$, despite the confusion about this on the web. That is $\frac{1+\sqrt{-3}}{2}$ is an integer in $\mathbb{Q}[\sqrt{-3}]$.) See any book on algebraic number theory; e.g., P. Samuel, *Algebraic Theory of Numbers*, or K. Ireland and M. Rosen, *A Classical Introduction to Modern Number Theory.*

Assuming that such factorizations were unique, Lamé thought that he had proved Fermat's Last Theorem in 1847. Dedekind fixed up arithmetic in such rings by developing the arithmetic of ideals, which are certain sets of numbers from the ring soon to be defined here. One then had (at least in rings of integers in algebraic number fields) unique factorization of ideals as products of prime ideals, up to order. Of course, Lamé's proof of Fermat's last theorem was still invalid (lame).

The favorite ring of the average human mathematics student is the field of real numbers $\mathbb{R}$. A favorite finite field for a computer is $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, where p=prime. Of course you can define $\mathbb{Z}_n$, for any positive integer n, but you only get a ring and not a field if n is not a prime. We considered $\mathbb{Z}_n$ as a group under addition in Part I. Now we view it as a ring with 2 operations, addition and multiplication.

Finite rings and fields were really invented by Gauss (1801) and earlier Euler (1750). Galois and Abel worked on field theory to figure out whether nth degree polynomial equations are solvable with formulas involving only radicals $\sqrt[m]{a}$. In fact, finite fields are often called "Galois fields." Dedekind introduced the German word *Körper* for field in 1871. David Hilbert introduced the term "ring" for the ring of integers in an algebraic number field in his Zahlbericht in 1897. Earlier Dedekind

had called these things "orders." The concept of ring was soon generalized. The relationship between groups and fields - Galois theory - was worked out by many mathematicians after Galois.

It would perhaps shock many pure mathematics students to learn how much algebra is part of the modern world of applied math. - both for good and ill. Google's motto: "Don't be evil," has not always been the motto of those using algebra. Of course, the Google search engine itself is a triumph of modern linear algebra.

Section 12 concerns random number generators from finite rings and fields. These are used in simulations of natural phenomena. In prehistoric times like the 1950s sequences of random numbers came from tables like that published by the Rand corporation. Random numbers are intrinsic to Monte Carlo methods. These methods were named for a casino in Monaco by J. von Neumann and S. Ulam in the 1940s while working on the atomic bomb. Monte Carlo methods are useful in computational physics and chemistry (e.g., modeling the behavior of galaxies, weather on earth), engineering (e.g., simulating the impacts of pollution), biology (simulating the behavior of biological systems such as a cancer), statistics (hypothesis testing), game design, finance (e.g., to value options, the analyze derivatives - the very thing that led to the horrible recession/depression of 2008), numerical mathematics (e.g., numerical integration, stochastic optimization).

In Section 13 we will show how the finite field with 2 elements and vector spaces over this field lead to error correcting codes. These codes are used in CDs and in the transmission of information between a Mars spacecraft and NASA on the earth. Section 14 concerns (among other things) the construction of Ramanujan graphs which can provide efficient communication networks.

Section 15 gives applications of the eigenvalues of matrices to Googling.

Section 16 gives applications of elliptic curves over finite field to cryptography.

Figure 1 comes from making an $m \times m$ matrix of values of $x^2 + y^2 (\mathrm{mod}\, m)$ for $x, y \in \mathbb{Z}/n\mathbb{Z}$. Then Mathematica does a `ListDensityPlot` of the matrix. There is a movie of such things on my website letting m vary from 3 to 100 or so.

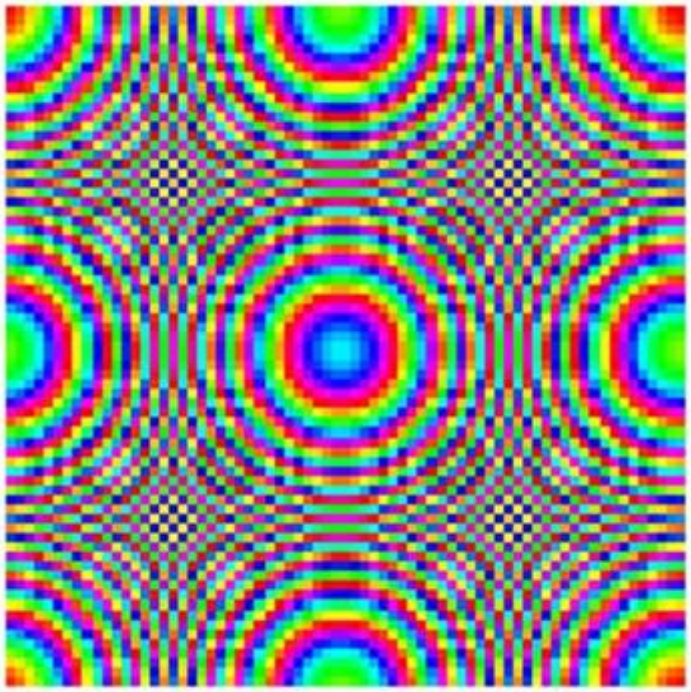

Figure 1: The color at point $(x, y) \in \mathbb{Z}_{163}^2$ indicates the value of $x^2 + y^2 (\mathrm{mod}\, 163)$.

A more complicated finite field picture is that of Figure 2. It is associated with 2×2 matrices with elements in the finite field $\mathbb{Z}_{11}$. We will explain it in Section 14.

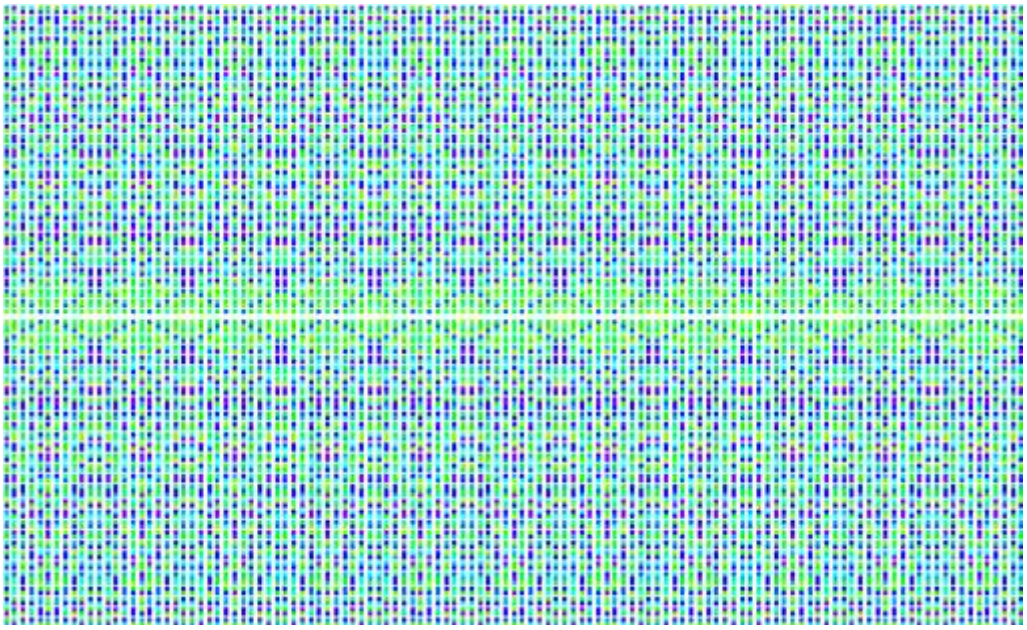

Figure 2: Points $(x, y) \in \mathbb{Z}_{11^2}^2$−the x-axis have the same color if $z = x + y\sqrt{\delta}$ are equivalent under the action of non-singular 2×2 matrices $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with entries in $\mathbb{Z}_{11}$. The action of g on z is by fractional linear transformation $z \to (az + b)/(cz + d) = gz$. Here δ is a fixed non-square in the field $\mathbb{F}_{121}$ with 121 elements.

Much of abstract ring theory was developed in the 1920s by Emmy Noether. Discrimination against both women and Jews made it hard for her to publish. The work became well known thanks to Van der Waerden's 2 volumes titled *Modern Algebra*. Van der Waerden wrote this after studying with Emmy Noether in 1924 in Göttingen. He had also heard lectures of Emil Artin in Hamburg earlier.

The abstract theory of algebras (which are special sorts of rings) was applied to group representations by Emmy Noether in 1929. This has had a big impact on the way people do harmonic analysis, number theory, and physics.

We should also mention Von Neumann rings of operators from a paper of J. Von Neumann in 1929. These are rings of operators acting on Hilbert spaces, which are (complete) infinite dimensional vector spaces with an inner product. The subject is intrinsic to modern analysis.

The rush to abstraction of 20th century mathematics has had some odd consequences. One of the results of the abstract ring theory approach was to create such an abstract version of Fourier analysis that few can figure out what is going on. A similar thing has happened in number theory where the abstract notion of adelic group representations has replaced the theory of automorphic forms for discrete groups of matrices acting on real symmetric spaces. See Terras, *Harmonic Analysis on Symmetric Spaces and Applications*, I,II for the classical version. On the other hand modern algebra has often made it easier to see the forest for the trees by simplifying computations, removing subsubscripts, doing calculations once in the general case rather than many times, once for each example.

The height of abstraction was achieved in the algebra books of Nicolas Bourbaki (really a group of French mathematicians). I am using the Bourbaki notation for the fields of real numbers, complex numbers, rational numbers, and the ring of integers. But Bourbaki seems to have disliked pictures as well as applications. I don't remember seeing enough examples or history either when I attempted to read Bourbaki's *Algebra* as an undergrad. Cartier in an interview for the *Math. Intelligencer* (Vol. 9, No. I, 1998) said: "The Bourbaki were Puritans, and Puritans are strongly opposed to pictorial representations of truths of their faith."

We will attempt to be as non-abstract as possible in these notes and will seek to draw pictures in a subject where few pictures ever appear.

References:

L. Dornhoff and F. Hohn, *Applied Modern Algebra*; J. Gallian, *Contemporary Abstract Algebra*; W. J. Gilbert and W. K. Nicholson, *Modern Algebra with Applications*; G. Birkhoff and S. Maclane, *A Survey of Modern Algebra*; I. Herstein, *Topics in Algebra*; A. Terras, *Fourier Analysis on Finite Groups and Applications.*

2 Rings

Our favorite ring for error-correcting codes will be $\mathbb{Z}_2$ or $\mathbb{Z}_p$, where p is a prime. Other favorites are the ring of integers $\mathbb{Z}$, the field of real numbers $\mathbb{R}$, the field of complex numbers $\mathbb{C}$, the field of rational numbers $\mathbb{Q}$.

Definition 1 *A **ring** R is an abelian group under $+$ with an associative multiplication satisfying left and right **distributive laws**:*

$$a(b+c) = ab + ac \text{ and } \quad (a+b)c = ac + bc, \text{ for all } \quad a, b, c \in R.$$

We will call the identity for addition 0. Multiplication in a ring need not be commutative. If it is, we say that the ring is a **commutative ring**.

Also, the ring need not have an identity for multiplication (except that some people do require this; e.g. M. Artin, *Algebra*). If the ring does have such an identity, we say it's a **ring with (2-sided) identity** for multiplication and we call this identity 1 so that $1 \cdot a = a \cdot 1 = a$, $\forall a \in R$. Some people (e.g., Gallian, *Contemporary Abstract Algebra)* call the identity for multiplication a "unity." I find that too close to the word unit which means that the element has a multiplicative inverse.

The identity for multiplication must be unique by the same argument that worked for groups. Normal people might want to assume that 1 and 0 are distinct as well. Otherwise $\{0\}$ is a ring with identity for multiplication. That must be the silliest ring with identity for multiplication. However, it looks like some people do call this a ring with identity for multiplication. What can I say? The terminology is not set in stone yet. The subject is still alive. However, I will normally assume that $1 \neq 0$.

The preceding examples $\mathbb{Z}_p$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, were all commutative rings.

Some authors drop the requirement that multiplication be associative and consider non-associative rings. Imagine the problems if you have to keep the parentheses in your products because $(ab)c \neq (ab)c$. We will not consider such rings here.

Example 1. A Non-commutative Ring.

Consider the ring $\mathbb{R}^{2\times 2} = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \middle| a,b,c,d \in \mathbb{R} \right\}$, with addition defined by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} + \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} a+a' & b+b' \\ c+c' & d+d' \end{pmatrix}$$

and multiplication defined by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \cdot \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} aa'+bc' & ab'+bd' \\ ca'+dc' & cb'+dd' \end{pmatrix}.$$

Example 2. A Ring without an Identity for Multiplication. $2\mathbb{Z}$, the ring of even integers.

Proposition 2 ***(Properties of Rings).*** *Suppose that R is a ring. Then, for all $a,b,c \in R$, we have the following facts.*
1) $a \cdot 0 = 0 \cdot a = 0$, where 0 is the identity for addition in R.
2) $a(-b) = (-a)b = -(ab)$ (where, as in Part I, $a + (-a) = 0$).
3) $(-a)(-b) = ab$.
4) $a(b-c) = ab - ac$.
5) If R has an identity for multiplication (which we call 1) then $(-1)a = -a$, $(-1)(-1) = 1$.

Proof. First recall that, since R is a group under addition, the identity, 0, is unique as are additive inverses $-a$ of elements a.

1) Using the fact that 0 is the identity for addition in R as well as the distributive laws, we have

$$0 + a \cdot 0 = a \cdot 0 = a \cdot (0+0) = a \cdot 0 + a \cdot 0.$$

Upon subtracting $a \cdot 0$ from both sides of the equation, we see that $0 = a \cdot 0$. You can make a similar argument to see that $0 \cdot a = 0$.

2) We have

$$a(-b) + ab = a(-b+b) = a \cdot 0 = 0 \Longrightarrow a(-b) = -(ab).$$

What ring axioms are being used at each point? We leave it as an **exercise** to finish the proof of 2).

3) First note that $-(-a) = a$ since $a + (-a) = 0$. Then, by part 2) and the associative law for multiplication:

$$(-a)(-b) = -(a(-b)) = -(-ab) = ab.$$

4) We have, using the distributive laws and part 2)

$$a(b-c) = ab + a(-c) = ab - ac.$$

5) **Exercise.** ■

Next we will define subring in an analogous way to the way we defined subgroup in Part I. You should be able to make the definition yourself without looking at what follows. Just don't forget to say the subring is non-empty.

Definition 3 *Suppose that R is a ring. If S is a non-empty subset of R which is a ring under the same operations as R, we call S a* ***subring*** *of R.*

Proposition 4 ***(Subring Test).*** *A non-empty subset S of a ring R is a subring of R iff S is closed under subtraction and multiplication.*

Proof. The 1-step subgroup test from Part I implies that S is a subgroup of R under addition. Moreover S must be abelian under $+$ since R is. Since S is closed under multiplication, we are done because the associative law for multiplication, plus the distributive laws follow from those in R. ■

Example 1. $\{0\}$ is a subring of any ring R.

To see this, apply the subring test. First note that $-0 = 0$ and thus $0 - 0 = 0 + 0 = 0$. Also $0 \cdot 0 = 0$, by multiplication rule 1).

Example 2. $S = \{0, 3, 6, 9 (\mathrm{mod}\, 12)\} = \{3x \mid x \in \mathbb{Z}_{12}\}$ is a subring of $\mathbb{Z}_{12}$.

To see this, use our subring test. Then $3x - 3y = 3(x-y)$ and $(3x)(3y) = 3(3xy)$ are both in S.

Example 3. $n\mathbb{Z}$ is a subring of $\mathbb{Z}$.

Example 4. The Gaussian integers $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$ is a subring of $\mathbb{C}$. Here $i = \sqrt{-1}$.

Again we use the subring test.

$$\begin{aligned} (a+bi) - (c+di) &= (a-c) + (b-d)i \in \mathbb{Z}[i], \\ \text{and} \quad (a+bi)(c+di) &= (ac-bd) + i(ad+bc) \in \mathbb{Z}[i], \end{aligned}$$

since $a, b, c, d \in \mathbb{Z}[i]$ implies $a - c$, $b - d$, $ac - bd$, $ad + bc \in \mathbb{Z}[i]$.

Example 5. The real numbers $\mathbb{R}$ form a subring of the complex numbers $\mathbb{C}$.

Example 6. The ring $\mathbb{Z}$ is a subring of $\mathbb{Z}[\sqrt{-5}] = \left\{ a + b\sqrt{-5} \;\middle|\; a, b \in \mathbb{Z} \right\}$.

Exercise. a) Show that $2\mathbb{Z} \cup 5\mathbb{Z}$ is not a subring of $\mathbb{Z}$.

b) Show that $2\mathbb{Z} + 5\mathbb{Z} = \{2n + 5m | n, m \in \mathbb{Z}\} = \mathbb{Z}$.

c) Show that $2\mathbb{Z} \cap 5\mathbb{Z} = 10\mathbb{Z}$.

Exercise. Consider the set

$$R = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \middle| a, b, c \in \mathbb{Z} \right\}.$$

Assume that addition is componentwise and multiplication is the usual matrix multiplication. Prove or disprove: R is a subring of the ring $\mathbb{Z}^{2x2}$ of all 2×2 matrices with integer entries.

Definition 5 *Suppose R is a ring with identity for multiplication, (which we call 1). The **units** in R are the invertible elements for multiplication:*

$$R^* = \{a \in R \mid \exists b \in R \text{ such that } ab = 1 = ba\} = \textit{units of } R.$$

If $ab = 1 = ba$, write $b = a^{-1}$.

Proposition 6 *If R is a ring with identity, the set of units R^* forms a group under multiplication.*

Proof. We need to check 4 things.

1) R^* is closed under multiplication.

2) The associative law holds for multiplication.

3) R^* has an identity for multiplication.

4) If $a \in R^*$, then $a^{-1} \in R^*$; i.e., R^* is closed under inverses.

To prove 2), you just need to recall that the associative law holds in R.

To prove 3), just note that $1 \cdot 1 = 1$.

To prove 4), let $a \in R^*$. Then there is a^{-1} in R such that $aa^{-1} = a^{-1}a = 1$. But then $a = \left(a^{-1}\right)^{-1}$ and thus $a^{-1} \in R^*$.

To prove 1), suppose $a, b \in R^*$. Then we have a^{-1} and b^{-1} in R and so $(ab)\, b^{-1}a^{-1} = abb^{-1}b^{-1} = 1$. Similarly $b^{-1}a^{-1}\,(ab) = 1$. It follows that $ab \in R^*$ with inverse $b^{-1}a^{-1}$. ■

One moral of the preceding proof is that in a non-commutative ring, $(ab)^{-1} = b^{-1}a^{-1}$. We knew this already from part I.

Example 1. $\mathbb{Z}^* = \{1, -1\}$.

To see this, just note that if n and $\frac{1}{n}$ are both in $\mathbb{Z}$, then n must be 1 or -1. Otherwise, $|n| > 1$ and $0 < \frac{1}{|n|} < 1$. This contradicts an exercise at the end of Section 3 of Part I.

Example 2. $\mathbb{Z}_n^* = \{a (\mathrm{mod}\, n) \mid \gcd(a, n) = 1\}$.

See Section 11 of Part 1.

Example 3. $\mathbb{Z}[x]$ =**ring of polynomials in 1 indeterminate x with integer coefficients**.

The elements of $\mathbb{Z}[x]$ have the form $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, where $a_j \in \mathbb{Z}$. If $a_n \neq 0$, we say that the **degree** of f is $n = \deg f$. The zero polynomial is not usually said to have a degree (unless you want to say it has degree $-\infty$).

To add two of these polynomials, if degree f is n and degree of g is $m \leq n$, put in some extra terms for g with coefficients that are 0, if necessary. Then you just add coefficients of like powers of x; i.e.,

$$\begin{aligned} f(x) &= a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \\ g(x) &= b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0 \end{aligned}$$

give

$$f(x) + g(x) = (a_n + b_n)x^n + (a_{n-1} + b_{n-1})x^{n-1} + \cdots + (a_1 + b_1)x + a_0 + b_0.$$

Multiplication is more complicated but you have known how to do this since high school. We know that we want the operation to be associative and distributive. So suppose

$$\begin{aligned} f(x) &= a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \\ g(x) &= b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0 \end{aligned}$$

$$\begin{aligned} f(x)g(x) &= \left(a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0\right)\left(b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0\right) \\ &= a_n b_m x^{m+n} + a_n b_{m-1}x^{m+n-1} + \cdots + a_n b_1 x^{n+1} + a_n b_0 x^n \\ &\quad + \left(a_{n-1}x^{n-1} + \cdots + a_1 x + a_0\right)\left(b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0\right) \\ &= a_n b_m x^{n+m} + (a_n b_{m-1} + a_{n-1}b_m)x^{n+m-1} + \cdots + \left(\sum_{i+j=k} a_i b_j\right) x^k + \cdots + (a_1 b_0 + a_0 b_1)x + a_0 b_0. \end{aligned}$$

The sum and product are still in $\mathbb{Z}[x]$. Checking the other ring properties is a bit tedious. The zero polynomial has all its coefficients equal to 0. The additive inverse of $f(x)$ has as its coefficients the negatives of the corresponding coefficients of $f(x)$. The multiplicative identity is the constant polynomial $f(x) = 1$. Checking the associative law for multiplication is the worst.

Assuming that no polynomial in the formula below is the zero polynomial, we have

$$\deg(fg) = \deg f + \deg g. \tag{1}$$

Exercise. Complete the proof that $\mathbb{Z}[x]$ is a commutative ring with identity for multiplication. Do your arguments work if you replace $\mathbb{Z}$ by any commutative ring R with identity for multiplication?

Exercise. What is the analog of formula (1) for $\deg(f+g)$? Hint: Consider an inequality rather than an equality.

Question. What is the group of units of $\mathbb{Z}[x]$?

To answer this, you need formula (1). This means that if $fg = 1$, $\deg f + \deg g = 0$. The only way that can happen is if $\deg f = \deg g = 0$. Thus

$$(\mathbb{Z}[x])^* = \mathbb{Z}^* = \{1, -1\}.$$

The group of units of the polynomial ring is the same as the group of units of the ring of integers.

Of course, you can still consider $\frac{1}{f(x)}$ but instead of a polynomial you get an infinite series. For example, the geometric series is

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n.$$

Moreover, this is only a convergent series if $|x| < 1$. But algebra is not suppose to deal with convergence and limits. Instead an algebraist would view this as a "**formal power series**" with coefficients in some ring R; as an element of $R[[x]]$, whose elements look like $\sum_{n=0}^{\infty} a_n x^n$, with $a_n \in R$.

Exercise. Find the group of units in $\mathbb{R}[x]$, the ring of polynomials with real coefficients.

Exercise. Check that the set $C(\mathbb{R})$ consisting of all continuous real valued functions on the real line forms a commutative ring if you define $(f+g)(x) = f(x) + g(x)$ for all $x \in \mathbb{R}$, and $(fg)(x) = f(x)g(x)$, $\forall\ x \in \mathbb{R}$. Here we assume that f, g are in $C(\mathbb{R})$. Does this ring have an identity for multiplication?

Exercise. Find units in the ring $\mathbb{Z}^{2\times 2}$ of 2×2 matrices with integer entries and the usual matrix operations.

3 Integral Domains and Fields are Nicer Rings

Next we want to consider rings that are more like the ring of integers.

Definition 7 *If R is a commutative ring, we say $a \neq 0$ in R is a **zero-divisor** if $ab = 0$ for some $b \in R$ such that $b \neq 0$.*

Example. In $R = \mathbb{Z}_6$, both 2 and 3 $(\mathrm{mod}\, 6)$ are zero divisors since $2 \cdot 3 \equiv 0 (\mathrm{mod}\, 6)$.

Definition 8 *If R is a commutative ring with identity for multiplication and no zero divisors, we say that R is an* ***integral domain****.*

I'm thinking zero divisors are "bad." and thus integral domains are "good." Of course I'm also thinking $\mathbb{Z}_6$ is pretty nice and it is clearly not an integral domain. Hmmmm. I must be thinking $\mathbb{Z}_5$ is way nicer than $\mathbb{Z}_6$.

Example 1. $\mathbb{Z}$ is an integral domain as are $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{Q}$.
Example 2. $\mathbb{Z}_n$ is not an integral domain if n is not a prime.

To see this, note that if n is not prime, then $n = ab$, where $0 < a, b < n$. But then neither a nor b can be congruent to 0 mod n and thus a and b are both zero divisors.
Example 3. $\mathbb{Z}_p$ is an integral domain if p is a prime.

To see this note that $ab \equiv 0 (\mathrm{mod}\, p) \Longleftrightarrow p$ divides ab. Then, by Euclid's Lemma from Section 5 of Part I, this means p divides either a or b. So either a or b is congruent to $0 (\mathrm{mod}\, p)$.

The following Lemma shows that cancellation is legal in an integral domain R even though inverses of non-0 elements may not exist in R.

Lemma 9 ***(Cancellation Law in an Integral Domain).*** *Suppose that R is an integral domain. If $a, b, c \in R$, $a \neq 0$, and $ab = ac$, then $b = c$.*

Proof. Since $ab = ac$, we see that $0 = ab - ac = a(b - c)$. Since $a \neq 0$ and R has no zero divisors, it follows that $b - c = 0$. Thus $b = c$. ■
Exercise. Show that the ring of matrices $\mathbb{Z}^{2\times 2}$, with the usual addition and multiplication, is not an integral domain.

Integral domains R are nice, but maybe not nice enough. Suppose we want to have $a^{-1} \in R$ for any $a \in R - \{0\}$. Then we want a field. Of course you can construct a field out of an integral domain by imitating the construction of $\mathbb{Q}$ out of $\mathbb{Z}$, but that is another story to be told in Section 9.

Definition 10 *A* ***field*** *F is a commutative ring with identity for multiplication such that any non-zero element $a \in F$ has a multiplicative inverse $a^{-1} \in F$.*

It follows from this definition that if F is a field, then the group of units $F^* = F - \{0\}$, which is as big as the unit group could be. Yes, there is no way it is ever legal to divide by 0.

Proposition 11 *1) Any field F is an integral domain.*
2) Any finite integral domain D is a field.

Proof. 1) If $a, b \in F$ such that $ab = 0$ and $a \neq 0$, then $b = a^{-1}ab = 0$. So F has no zero divisors.

2) We just need to show that $D - \{0\}$ is a group under multiplication. It is clearly closed and satisfies the associative law. So we just need to show that it is closed under inverse. We proceed as in the proof of the finite subgroup test in Section 12 of Part I. That is we look at $\langle a \rangle = \{a, a^2, a^3, ...\}$ for $a \in D - \{0\}$. Since $\langle a \rangle$ is finite, we know that $a^i = a^j$ for some pair i, j with $i > j$. But then $a^{i-j} = 1$. It follows then that $a^{-1} = a^{i-j-1}$. ■

Example 1. $\mathbb{Z}$ is not a field as the only units in $\mathbb{Z}$ are 1 and -1.
Example 2. $\mathbb{Z}_p$ is a field iff $p =$prime.
Example 3. $\mathbb{Q} =$ rational numbers $= \left\{ \frac{a}{b} | a, b \in \mathbb{Z}, b \neq 0 \right\}$ is a field.
Example 4. The set of real numbers $\mathbb{R}$ is also a field as is the set of complex numbers $\mathbb{C}$.

So we could view the finite field $\mathbb{Z}_p$ for $p =$prime, as an analog of the field $\mathbb{R}$ of real numbers. But the picture of $\mathbb{R}$ is a continuous line without holes, while our picture of $\mathbb{Z}_p$ is a finite circle of points.
Question. Are there other finite fields?

Answer. Yes, you can imitate the construction that gives the complex numbers $\mathbb{C}$.

$$\mathbb{F}_9 = \mathbb{Z}_3[i] = \{a + bi \mid a, b \in \mathbb{Z}_3\}, \quad \text{where } i^2 = -1.$$

The order of $\mathbb{F}_9$ is 9 since there are 3 choices of a and 3 choices of b in $a + ib$. You add and multiply in $\mathbb{F}_9$ just as you would in the complex numbers, except that every computation is modulo 3. Why is it a field? Certainly you get a ring.

To see that $\mathbb{F}_9$ is a field, we need to see that if $a+ib$ is a non-zero element of $\mathbb{F}_9$, then it has an inverse in $\mathbb{F}_9$. Use the same argument that works for $\mathbb{C}$. That is,

$$\frac{1}{a+ib} = \frac{1}{a+ib}\frac{a-ib}{a-ib} = \frac{a-ib}{a^2+b^2} = \frac{a}{a^2+b^2} + i\frac{-b}{a^2+b^2}.$$

Why is $a^2+b^2 \neq 0$? This is a little harder to prove than it was if $a, b \in \mathbb{R}$. We know that $a+ib \neq 0 \Longrightarrow$ either a or b is not 0 in $\mathbb{Z}_3$. Suppose a is not 0 in $\mathbb{Z}_3$. Thus $a \equiv 1$ *or* $-1 (\mathrm{mod}\, 3)$. So $a^2 \equiv 1 (\mathrm{mod}\, 3)$. Then $b^2 \equiv 0$ or $1 (\mathrm{mod}\, 3)$. It follows that $a^2+b^2 \equiv 1$ or $2 (\mathrm{mod}\, 3)$. Thus a^2+b^2 is not $0 (\mathrm{mod}\, 3)$. Since $\mathbb{Z}_3$ is a field, we know that $\frac{1}{a^2+b^2} \in \mathbb{Z}_3$.

Note that for any element $z \in \mathbb{F}_9$, $3z = z+z+z = 0$. This happens because $z = x+iy$, with $x, y \in \mathbb{Z}_3$ and $3z = 3x + i3y = 0$. We say that $\mathbb{F}_9$ has characteristic 3.

Definition 12 *The* ***characteristic*** *of a ring R is the smallest $n \in \mathbb{Z}^+$ such that $nx = \underbrace{x+x+\cdots+x}_{n \text{ times}} = 0$, for all $x \in R$. If no such n exists, we say that R has* ***characteristic 0***.

Lemma 13 *Suppose that R is a ring with identity 1 for multiplication. Then we have the following facts.*
1) If the additive order of 1 is not finite, then the characteristic of R is 0.
2) If the additive order of 1 is n, then the characteristic of R is n.

Proof. 1) **Exercise**.
2) Clearly the characteristic must be divisible by n. On the other hand, if $\underbrace{1+\cdots+1}_{n \text{ times}} = 0$, then $\forall x \in R$

$$\left(\underbrace{1+1+\cdots+1}_{n \text{ times}}\right)x \;=\; \underbrace{x+x+\cdots+x}_{n \text{ times}} \;= 0.$$

This means that the characteristic is $\leq n$. It follows that the characteristic must equal n. ■

Example 1. $\mathbb{Z}$,$\mathbb{Q}$,$\mathbb{R}$,$\mathbb{C}$ all have characteristic 0. To see this, by the preceding Lemma, you just need to note that no finite sum of ones can equal 0 in these rings.

Example 2. $\mathbb{Z}_p$ has characteristic p by the Lemma since that is the additive order of 1 in $\mathbb{Z}_p$.

Example 3. $\mathbb{F}_9$ has characteristic 3 since that is the additive order of 1, again using the preceding Lemma. $\mathbb{F}_9$ has 9 elements.

Example 4. $\mathbb{Z}_p[x]$, the ring of polynomials in 1 indeterminate with coefficients in $\mathbb{Z}_p$, has characteristic p. $\mathbb{Z}_p[x]$ has infinitely many elements.

Example 5. $\mathbb{Z}_5[i] = \{\, a+bi \mid a, b \in \mathbb{Z}_5\}$, where $i^2+1=0$, is not a field, since $(2+i)(2-i) = 0$.

In 1964 D. Singmaster asked in the *American Math. Monthly* how many rings of order 4 are there? The solution was given by D. M. Bloom (11 rings of order 4, of which 3 have a multiplicative identity). See the website on small rings from an abstract algebra class of Gregory Dresden at the Math. Dept. of Washington and Lee University. We list these rings of order 4, though we have yet to define direct sum of rings, quotient rings, and isomorphic rings. You should be able to guess what these things are from your knowledge of direct sum of groups and quotient groups.

The Rings of Order 4 (Up to Isomorphism)

1) $\mathbb{Z}_4$,
2) the ring of matrices with entries in $\mathbb{Z}_4$:

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, a = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 3 & 0 \end{pmatrix} \right\}.$$

This is additively generated by a and $a^2 = 0$.
3) the ring of matrices with entries in $\mathbb{Z}_4$:

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, a = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}, \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} \right\}.$$

This is additively generated by a and we have $a^2 = 2a$.

4) the ring of matrices with entries in $\mathbb{Z}_4$:

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, a = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \right\}.$$

5) As a subring of $\mathbb{Z}_2 \oplus \mathbb{Z}_4$, $\{(0,0),(0,2),(1,0),(1,2)\}$.

6) A quotient ring of the ring of polynomials over $\mathbb{Z}_4$: $\mathbb{Z}_4[x]/\langle 2x, x^2+x \rangle$

Here $\langle 2x, x^2+x \rangle$ means the subring of $\mathbb{Z}_4[x]$ generated by $2x$ and x^2+x.

7) ring direct sum: $\mathbb{Z}_2 \oplus \mathbb{Z}_2$.

8) A quotient ring of the ring of polynomials over $\mathbb{Z}_2$: $\mathbb{Z}_2[x]/\langle x^2+1 \rangle = \{0, 1, x, 1+x\}$

9) A field with 4 elements: $\mathbb{Z}_2[x]/\langle x^2+x+1 \rangle = \{0, 1, x, 1+x\}$

This differs from ring 8 since we will see that this ring is a field while ring 8 has zero divisors, since $x^2+1 = (x+1)^2$ as elements of $\mathbb{Z}_2[x]$.

Ring 9 can also be viewed as a matrix ring with coefficients in $\mathbb{Z}_2$:

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, 1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, x = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \right\}.$$

10) a matrix ring with coefficients in $\mathbb{Z}_2$:

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, a = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, b = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \right\}.$$

11) a matrix ring with coefficients in $\mathbb{Z}_2$:

$$\left\{ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, a = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, b = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \right\}.$$

Exercise. Find the characteristics of the rings of order 8. Then tell which are commutative, which have a multiplicative identity, which are integral domains, which are fields?

Lemma 14 *Suppose that R is an integral domain. Then the characteristic of R is either a prime or 0*

Proof. If the additive order of 1 is not finite, the characteristic of R is 0, by the preceding Lemma.

Suppose the additive order of 1 is $n \in \mathbb{Z}^+$. We must show that n is prime. We do a proof by contradiction.

Otherwise $n = ab$ for some integers a, b with $0 < a, b < n$. This means $0 = a \cdot b = (a \cdot b) \cdot 1 = (a \cdot 1) \cdot (b \cdot 1)$. Since R has no zero divisors, it follows that either $a \cdot 1 = 0$ or $b \cdot 1 = 0$. But this contradicts the minimality of $n = ab$. It follows that n is prime. ■

Question. Which of the following 5 ring examples are fields?

$\mathbb{Z}$, $\mathbb{Z}[i]$, where $i^2 = -1$, $\mathbb{Z}[x]$ =polynomials with integer coefficients, $\mathbb{Z}[\sqrt{-5}]$, $\mathbb{Z}_p$, for prime p

Answer. Only the last example is a field. In all other cases $\frac{1}{2}$ is not in the ring even though 2 is.

We define subfield just as we define subgroup or subring.

Definition 15 *A subset F of a field E is a **subfield** if it is a field under the operations of E. We also say that E is a field extension of F.*

Proposition 16 ***(Subfield Test).*** *Suppose that E is a field and $F \subset E$. Then F is a subfield of E iff $\forall a, b \in F$ with $b \neq 0$, we have $a - b$ and $ab^{-1} \in F$.*

Proof. Just use the 1-step subgroup test on F to see that it is an additive subgroup of E and then use the same test again to see that $F - \{0\}$ is a multiplicative subgroup of $E - \{0\}$. The distributive laws are automatic. ■

The following Lemma gives a formula in characteristic $p \neq 0$ which some calculus students seem to believe is true in the real numbers. But that would say most of the terms in the binomial theorem somehow vanish miraculously.

Lemma 17 *Suppose that R is an integral domain of (necessarily prime) characteristic p. Then $\forall x, y \in R$, we have $(x+y)^p = x^p + y^p$.*

Proof. By the binomial theorem (whose proof works in any integral domain), we have

$$(x+y)^p = \sum_{k=0}^{p} \binom{p}{k} x^k y^{p-k}.$$

To finish this proof, we must show that the prime p divides $\binom{p}{k}$ if $k = 1, 2, ..., p-1$. This follows from the fact that the binomial coefficient is an integer which is represented by the fraction:

$$\binom{p}{k} = \frac{p(p-1)\cdots 2\cdot 1}{k(k-1)\cdots 2\cdot 1}.$$

Since p clearly divides the numerator, we just need to show that p does not divide the denominator. But that is true since p divides no factor in the denominator. This means that the binomial coefficients that are not congruent to 0 mod p are only those of the $k = 0$ and $k = p$ terms. ■

Exercise. a) Which of the following rings are integral domains? Give a brief explanation of your answer.

b) Same as a), replacing "integral domains" with "fields."

i) $\mathbb{Z}[i] = \{a + bi | a, b \in \mathbb{Z}\}$, where $i^2 = -1$.

ii) $\mathbb{Z}/12\mathbb{Z}$.

iii) $\mathbb{Z}_2^{2\times 2}$, 2×2 matrices with coefficients in $\mathbb{Z}_2$.

iv) $\mathbb{Z}_{11}$.

v) $\mathbb{Z} \oplus \mathbb{Z}$.

vi) $\mathbb{Q}$ =rational numbers.

vii) $C(\mathbb{R}) = \{$continuous real valued functions $f : \mathbb{R} \to \mathbb{R}\}$ with addition and multiplication defined as usual in calculus pointwise; i.e., $\forall x \in R$, $(f+g)(x) = f(x) + g(x)$ and $(fg)(x) = f(x)g(x)$.

Exercise. a) List all the zero divisors in the 7 rings from the preceding problem except that you should replace vii) $C(\mathbb{R})$ with $C^{pw}(\mathbb{R})$ =the piecewise continuous functions on $\mathbb{R}$ (i.e., we allow a finite number of removable or jump discontinuities).

b) List all the units in the rings R of part a); i.e. find R^*.

c) What is the relation between the zero divisors and the units of R, if any?

4 Building New Rings from Old: Quotients and Direct Sums of Rings

We need to build quotient rings in the same way that we constructed $\mathbb{Z}/n\mathbb{Z}$. We will also imitate the construction of quotient groups in Section 17 of Part I. To create a quotient group using a subgroup H of a group G, we needed H to be a normal subgroup. It turns out we will need a similar notion for the subring S of ring R. That is, we will need S to be an ideal in R.

Definition 18 *A non-empty subset I of a ring R is an **ideal** iff I is an additive subgroup of R such that $ra \in I$ and $ar \in I$, $\forall r \in R$ and $\forall a \in I$.*

Example. $n\mathbb{Z}$ is an ideal in $\mathbb{Z}$.

We call n the **principal ideal generated by n** and write $n\mathbb{Z} = \langle n \rangle$.

Definition 19 *Given a ring R and an element $a \in R$, the (2-sided) **ideal generated by a,** denoted $\langle a \rangle$, consists of elements ra and ar for all $r \in R$. Similarly, the ideal $\langle S \rangle$ generated by a subset S of R is the smallest ideal containing S.*

Exercise. a) Suppose that R is a commutative ring with identity. Let $S \subset R$. Show that $\langle S \rangle = \{ra + sb \mid \forall r, s \in R, \forall a, b \in S\}$ is an ideal.

b) Suppose $R = \mathbb{Z}$. Show that if $a, b \in \mathbb{Z}$, then $\langle a, b \rangle = \langle \gcd(a, b) \rangle$.

Ideals were introduced by R. Dedekind in 1879. The main use for them in number theory is to get a substitute for prime numbers - the prime ideals we are about to define. This allows one to have unique factorization of ideals in rings of integers of algebraic number fields into products of prime ideals, though the unique factorization fails for actual algebraic integers in a number field like $\mathbb{Q}(e^{2\pi i/n})$. The concept of ideal was further developed by D. Hilbert and E. Noether.

To construct the quotient ring R/I if I is an ideal in the ring R, we create the set of additive **cosets** $[a] = a + I = \{a + r \mid r \in R\}$. Once again, you can view these cosets as equivalence classes for the equivalence relation defined on elements

$a, b \in R$ by $a \sim b$ iff $a - b \in I$. **Exercise.** Prove this last statement. We should note that some authors (e.g., Gallian) call R/I a "factor ring."

Then we add and multiply cosets as usual:

$$[a] + [b] = [a + b], \qquad [a] \cdot [b] = [ab]. \tag{2}$$

Theorem 20 *Suppose that I is a subring of the ring R. Then, with the definitions just given, R/I is a ring iff I is an ideal in R.*

Proof. $\Longrightarrow$ If I is an ideal in R, we need to see that the operations defined in formula (2) make R/I a ring. Once we have checked that the operations are well defined, everything else will be easy. To check the operations make sense, suppose that $[a] = [a']$ and $[b] = [b']$. Then we must show that $[a + b] = [a' + b']$ and $[ab] = [a'b']$. In fact, we have already checked the additive part in Section 17 of Part I, since I is automatically a normal subgroup of the additive group of R. Why?

So let's just check the multiplicative part. We need to prove that $ab - a'b' \in I$. To do this, we recall proofs of the formula for the derivative of a product. That means we should add and subtract $a'b$ (or ab'). This gives

$$ab - a'b' = ab - a'b + a'b - a'b' = (a - a')\, b + a'(b - b').$$

Since both $a - a'$ and $b - b'$ are in the ideal I, it follows that $(a - a')\, b$ and $a'(b - b')$ are both in I. But then the sum must be in I and we're done.

So now we know addition and multiplication in R/I are well defined. From the fact that R is a ring, it is easy to see that R/I must be a ring too. The identity for addition in R/I is $[0]$. The additive inverse of $[a]$ is $[-a]$. The associative laws in R/I follow from the laws in R as do the distributive laws.

$\Longleftarrow$ Conversely, if R/I is a ring, the multiplication of cosets defined in formula (2) must be well defined. Since $[0] = I$, for any $a \in R$, we have $[0]\,[a] = [a]\,[0] = [0]$. This means $IR \subset I$ and $RI \subset I$. Of course I must also be closed under addition and subtraction as $[0] \pm [0] = [0]$ in R/I. Thus I must be an ideal in R/I. ■

Example. Consider the ring $\mathbb{R}[x]$ of polynomials in the indeterminate x. An example of an ideal in this ring is

$$I = \langle x^2 + 1 \rangle = \left\{ f(x)(x^2 + 1) \;\middle|\; f(x) \in \mathbb{R}[x] \right\}.$$

Question: What is $\mathbb{R}[x]/\langle x^2 + 1 \rangle$?

Answer: We can identify this quotient with the ring $\mathbb{C}$ of complex numbers. To give some evidence for this statement, let $\Theta = [x] = x + I = x + \langle x^2 + 1 \rangle$ in $\mathbb{R}[x]/I$. Then $\Theta^2 + 1 = [x]^2 + [1] = \left[x^2 + 1\right] = [0]$. This means $\Theta^2 = -1$ in our ring $\mathbb{R}[x]/I$. So Θ behaves like i in $\mathbb{C}$.

In order to prove our statement identifying $\mathbb{C}$ and $\mathbb{R}[x]/\langle x^2 + 1 \rangle$, we need to study polynomial rings a little more and we need to define what we mean by isomorphic rings. See Section 5. In particular, we need the analog of the division algorithm for polynomial rings like $\mathbb{R}[x]$. Once we have that, we can identify cosets $[f(x)]$ in $\mathbb{R}[x]/I = \mathbb{R}[x]/\langle x^2 + 1 \rangle = \mathbb{R}[x]/\left(x^2 + 1\right)\mathbb{R}[x]$ with cosets of the remainders $[r(x)]$ upon dividing $f(x)$ by x^2+1; i.e., $f(x) = \left(x^2 + 1\right) q(x) + r(x)$, where $\deg r < 2$ or $r(x) = 0$. So the remainders look like $a + bx$, with $a, b \in \mathbb{R}$. This means elements of $\mathbb{R}[x]/I$ have the form $[a + bx] = [a] + [b][x] = a + b\Theta$, which we can identify with a complex number $a + bi$, for $a, b \in \mathbb{R}$.

Our Goal. Replace $\mathbb{R}$ in the preceding construction with a finite field $\mathbb{Z}_p$. Then replace $x^2 + 1$ with any irreducible polynomial mod p. Then apply the result to error correcting codes in Section 13. For example, take $p = 2$. Since $x^2 + 1 = (x + 1)^2$ in $\mathbb{Z}_2[x]$, we know that $x^2 + 1$ is not an irreducible polynomial in $\mathbb{Z}_2[x]$. An irreducible polynomial is our analog of a prime in the ring $\mathbb{Z}_2[x]$. We have already seen an irreducible polynomial in $\mathbb{Z}_2[x]$. It was $x^2 + x + 1$. Why? It has no roots mod 2 and thus cannot have degree 1 factors as we will show in our section on polynomial rings. This will imply that $\mathbb{Z}_2[x]/\langle x^2 + x + 1 \rangle$ is a field with 4 elements $\{[0], [1], [x], [x + 1]\}$ where $[x]^2 + [x] + 1 = 0$.

The following theorem says that every ideal in the ring of integers is principal. We will have a similar theorem later about rings like $\mathbb{R}[x]$.

Theorem 21 *Any ideal I in the ring $\mathbb{Z}$ of integers is principal; i.e., $I = \langle n \rangle = n\mathbb{Z}$ for some $n \in \mathbb{Z}$. In fact, if $I \neq \{0\}$, we can choose n to be the smallest positive element of I.*

Proof. The case $I = \{0\} = \langle 0\rangle$ is clear. Otherwise $I \neq \{0\}$ and we let n be the least positive element of I. Then $\langle n\rangle \subset I$. Suppose that $a \in I$. The division algorithm says $a = nq + r$, with $0 \leq r < n$. Since I is an ideal and $r = a - nq$, we know that $r \in I$. But n is the least positive element of I. Therefore $r = 0$. This implies $I \subset \langle n\rangle$. So $I = \langle n\rangle$. ■

Question: Suppose I is an ideal in R, a commutative ring with identity for multiplication. When is the quotient ring R/I an integral domain?

Answer: When I is a **prime ideal** meaning that $ab \in I \Longrightarrow$ either a or $b \in I$.

Proof. First note that R/I automatically has all the properties of an integral domain except for the lack of zero divisors. It inherits these properties from R. For example, the identities for addition and multiplication are $[0]$ and $[1]$, respectively. We get a zero divisor in R/I iff there are $a, b \in R$ such that $[a][b] = [0]$ but $[a] \neq [0]$ or $[b] \neq [0]$. This means $ab \in I$ but $a \notin I$ or $b \notin I$. ■

Example. Which ideals $\langle n\rangle$ in $\mathbb{Z}$ are prime ideals? The answer is the ideals $\langle p\rangle$ with p a prime.

Proof. First note that $ab \in \langle n\rangle = n\mathbb{Z}$ is equivalent to saying n divides ab.

If n is not a prime, then $n = ab$ with $1 < a, b < n$. It follows that $ab \in \langle n\rangle$, but n cannot divide either a or b. Thus $\langle n\rangle$ cannot be a prime ideal in $\mathbb{Z}$.

If p is a prime and $ab \in \langle p\rangle$, then p divides ab. Euclid's lemma from Part I, Section Section 5, tells us that then p must divide either a or b. Thus either a or b must be in $\langle p\rangle$ and $\langle p\rangle$ is a prime ideal. ■

But we really want the answer to the following question.

Question: Suppose I is an ideal in R, a commutative ring with identity for multiplication. We assume $1 \neq 0$ in R. When is the quotient ring R/I a field?

Answer: When I is a **maximal** ideal. This means that if A is an ideal of R such that $I \subset A \subset R$, then either $A = I$ or $A = R$.

Proof. First note that R/I automatically inherits all the properties of a field from R except closure under inverse for non-zero elements. In particular, $[0] = I$ is the identity for addition in R/I and $[1] = 1 + I$ is the identity for multiplication.

Suppose that R/I is a field. If A is an ideal such that $I \subset A \subset R$ but $A \neq I$, then we need to show that $A = R$. Since $A \neq I$, there is an element $a \in A - I$. This means $[a] \neq [0]$ in R/I. Since R/I is a field, there exists $[b] \in R/I$ such that $[a][b] = [1]$. This means $ab - 1 \in I$. Thus $1 = ab - c$ for some $c \in I$. But then, because I is an ideal containing a and c, it follows that $1 \in I$. Therefore, for any $r \in R$, we have $r = 1 \cdot r \in I$ and $I = R$. Thus I is a maximal ideal.

Now suppose that I is maximal. We need to show that R/I is a field. Suppose $[a] \neq [0]$ in R/I. We need to find $[a]^{-1}$. Look at the ideal B generated by I and a. That is $B = \{c + ra \mid c \in I, r \in R\} = \langle I, a\rangle$. **Exercise.** Show that B is an ideal. Then $I \subset B \subset R$. We know that $I \neq B$. Since I is maximal, it follows that $B = R$. But then $1 \in B$. So $1 = c + ra$ for some $c \in I, r \in R$. This implies $[r][a] = [1]$. So $[r] = [a]^{-1}$ and we are done. ■

Example 1. Which ideals $\langle n\rangle$ in $\mathbb{Z}$ are maximal ideals? The answer is that the prime ideals in $\mathbb{Z}$ are the maximal ideals in $\mathbb{Z}$. We know this since we proved finite integral domains are fields in Proposition 11.

Exercise. Prove the prime ideals in $\mathbb{Z}$ are the maximal ideals in $\mathbb{Z}$ directly by showing $\mathbb{Z}/n\mathbb{Z}$ is a field iff n=prime.

Example 2. What are the maximal ideals in $\mathbb{Z}_{12}$?

First we show that all ideals I in $\mathbb{Z}_{12}$ are principal. To prove this, consider the the corresponding ideal $\widetilde{I}$ in $\mathbb{Z}$, which is $\widetilde{I} = \{m \in \mathbb{Z} \mid [m] \in I\}$. **Exercise.** Show that $\widetilde{I}$ is an ideal in $\mathbb{Z}$. Now, we know any ideal in $\mathbb{Z}$ is principal. Thus $\widetilde{I} = \langle n\rangle = n\mathbb{Z}$ for some $n \in \mathbb{Z}$. But then $I = n\mathbb{Z}_{12}$.

Next suppose that u is a unit in $\mathbb{Z}_{12}$. Then we can show that $un\mathbb{Z}_{12} = n\mathbb{Z}_{12}$. For the fact that $r = u^{-1}$ exists in $\mathbb{Z}_{12}$ implies $n\mathbb{Z}_{12} = u^{-1}un\mathbb{Z}_{12} \subset un\mathbb{Z}_{12}$. There is no problem seeing the reverse inclusion $un\mathbb{Z}_{12} \subset n\mathbb{Z}_{12}$. This is a general fact about principal ideals, by the way.

Now we can list all the ideals in $\mathbb{Z}_{12}$. They are:

$$\begin{aligned}
\langle 0\rangle &= \{0\},\ \langle 1\rangle = \langle 3\rangle = \langle 5\rangle = \langle 7\rangle = \langle 11\rangle = \mathbb{Z}_{12},\ \langle 2\rangle = \langle 10\rangle = 2\mathbb{Z}_{12} = \{0, 2, 4, 6, 8, 10 (\mathrm{mod}\, 12)\},\\
\langle 3\rangle &= \langle 9\rangle = 3\mathbb{Z}_{12} = \{0, 3, 6, 9 (\mathrm{mod}\, 12)\},\ \langle 4\rangle = \langle 8\rangle = 4\mathbb{Z}_{12} = \{0, 4, 8 (\mathrm{mod}\, 12)\},\\
\langle 6\rangle &= 6\mathbb{Z}_{12} = \{0, 6 (\mathrm{mod}\, 12)\}.
\end{aligned}$$

The poset diagram for the ideals in $\mathbb{Z}_{12}$ is in Figure 3.

It follows from Figure 3 that the maximal ideals in $\mathbb{Z}_{12}$ are $\langle 2\rangle$ and $\langle 3\rangle$.

Exercise. Explain the equalities in the formulas for the ideals of $\mathbb{Z}_{12}$; e.g. why is it that $\langle 8\rangle = \langle 4\rangle$?

Exercise. Show that any ideal in the ring $\mathbb{Z}_n$ is principal.

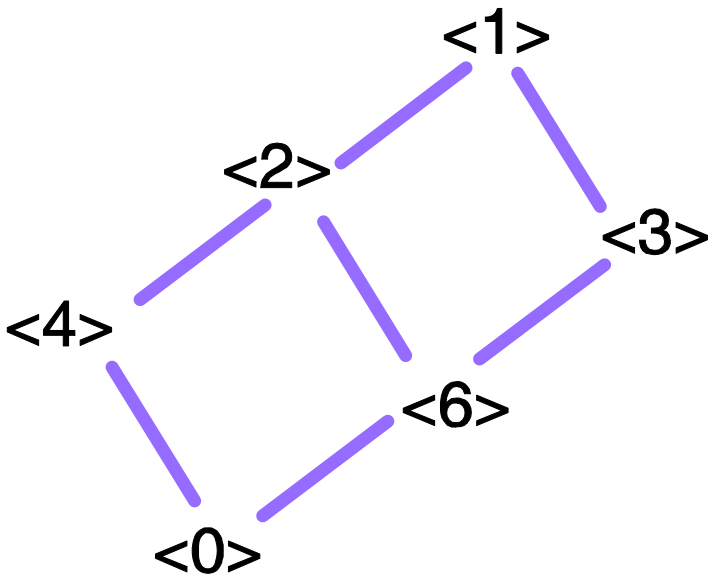

Figure 3: poset diagram of the ideals in $\mathbb{Z}_{12}$

Exercise. Suppose that R is an integral domain. Show that for $a, b \in R$, we have $aR = \langle a \rangle = \langle b \rangle = bR$ iff $a = ub$, where u is a unit in R.
Exercise. Find all the maximal ideals in $\mathbb{Z}_{15}$.

Definition 22 *To create the* ***direct sum*** $R \oplus S$ *of 2 rings R and S, we start with the Cartesian product $R \times S$ and define the ring operations componentwise. That is, for (a, b) and $(r, s) \in R \times S$, we define $(a, b) + (r, s) = (a + r, b + s)$ and $(a, b)(r, s) = (ar, bs)$.*

Exercise. Show that the preceding definitions make $R \oplus S$ a ring.
Exercise. Is $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ a field, an integral domain? Same question for $\mathbb{Z}_2 \oplus \mathbb{Z}_3$.
Exercise. Find the characteristics of the following rings: $\mathbb{Z}_2 \oplus \mathbb{Z}_2$, $\mathbb{Z}_2 \oplus \mathbb{Z}_4$, $\mathbb{Z}_2 \oplus \mathbb{Z}_3$.
Exercise. Find a subring of $R = \mathbb{Z} \oplus \mathbb{Z}$ that is not an ideal. **Hint.** Look at $S = \{(a, b) | a + b \text{ is even}\}$.
Exercise. If A and B are ideals in a commutative ring R, define the **sum** $A + B = \{a + b | a \in A, b \in B\}$ and the **product** $AB = \left\{ \sum_{i=1}^{n} a_i b_i \;\middle|\; a_i \in A, b_i \in B \right\}$.
a) Show that $A + B$ and AB are ideals of R.
b) Show that $A + B = R$ implies $AB = A \cap B$.
c) Suppose $R = \mathbb{Z}$. If $A = \langle a \rangle$ and $B = \langle b \rangle$, show that $A + B = \langle \gcd(a, b) \rangle$. If $A + B = \langle 1 \rangle$, show that $AB = \langle ab \rangle = A \cap B$.

5 Polynomial Rings

Suppose that F is a field. We consider the ring $F[x]$ of polynomials in one indeterminate, x, and coefficients in F.

Beware: Don't confuse polynomials and functions. For example, in $\mathbb{Z}_3[x]$ the polynomials $f(x) = x^2 + x + 1$ and $g(x) = x^4 + x + 1$ represent the same function even though the 2 polynomials are different. To see this, plug in the elements of $\mathbb{Z}_3$.

$f(0) = 1$	$f(1) = 0$	$f(-1) = 1$
$g(0) = 1$	$g(1) = 0$	$g(-1) = 1$

This sort of thing has to happen, since the number of functions $T : \mathbb{Z}_3 \to \mathbb{Z}_3$ is $3^3 = 27$, while the ring $\mathbb{Z}_3[x]$ is infinite.

Hopefully we did the exercise in Section 2 showing that $R[x]$ is a commutative ring with multiplicative identity, if R is a commutative ring with multiplicative identity. If not, do that **exercise** now! The bad part is the associative law for multiplication. It might help to look at the following:

$$\left(\sum_{k=0}^{K} a_k x^k\right)\left(\left(\sum_{m=0}^{M} b_m x^m\right)\left(\sum_{n=0}^{N} c_n x^n\right)\right) = \sum_{r=0}^{K+M+N} x^r \left(\sum_{\substack{k+m+n=r \\ 0\le k\le K \\ 0\le m\le M \\ 0\le n\le N}} a_k b_m c_n\right).$$

If $R = \mathbb{Z}_3$, we add and multiply as in the following examples.

$(x^2 + 2x + 2) + \left(x^3 + x + 2\right) = x^3 + x^2 + 1$ (since $3 \equiv 0 (\mathrm{mod}\, 3)$ and $4 \equiv 1 (\mathrm{mod}\, 3)$);

$(x^2 + 2x + 1) \cdot \left(x^3 + 2\right) = x^5 + 2x^4 + x^3 + 2x^2 + x + 2.$

We learned to do this in the dim dark past by making the following table:

$$\begin{array}{r} x^2 + 2x + 1 \\ x^3 + 2 \\ \hline 2x^2 + 4x + 2 \\ x^5 + 2x^4 + x^3 \\ \hline x^5 + 2x^4 + x^3 + 2x^2 + x + 2 \end{array} \quad (\text{since } 4 \equiv 1 (\mathrm{mod}\, 3))$$

Recall that units R^* of a ring R are the invertible elements for multiplication in R. When $R = \mathbb{Z}$, the only units are 1 and -1. When the ring is $\mathbb{Z}[x]$, it turns out the units are the same as for $\mathbb{Z}$, as we saw in Section 2. We get the same result when F is a field, i.e.,

$$(F[x])^* = F^* = F - \{0\}. \tag{3}$$

The proof is the same as that in Section 2 for $\mathbb{Z}$.

Exercise. Prove formula (3).

Now we want to imitate what we said about the integers in Section 5 of Part I of these notes. We will have analogs of primes, the division algorithm, the Euclidean algorithm, and the fundamental theorem of arithmetic for the ring $F[x]$, where F is any field. Pretty amazing!

Assumption. For the rest of this section F is a field.

First we define the polynomial analog of prime.

Definition 23 *A polynomial $f(x)$ of degree > 0 is **irreducible** iff $f(x) = g(x)h(x)$ for $g, f \in F[x]$ implies either g or h has degree 0.*

Now we want to get rid of the units in a factorization as we did in $\mathbb{Z}$ by allowing only positive non unit integers to be called primes assuming they could not be factored non-trivially. To get rid of units in $F[x]$, we look at **monic** polynomials; i.e., polynomials with leading coefficient (i.e., coefficient of the highest power of x) equal to 1. So a monic irreducible polynomial is the analog of a prime in $F[x]$.

Example. Irreducible Polynomials of Low Degree in $\mathbb{Z}_2[x]$.

degree 1 polynomials: $x, x + 1$ Both irreducible.

degree 2 polynomials: $x^2,\ x^2+1 = (x+1)^2,\ x^2+x = x(x+1),\ x^2+x+1$. The 1st 3 polynomials are clearly reducible. What about $x^2 + x + 1$? Does x or $x + 1$ divide $x^2 + x + 1$? The answer is: No! For we have $x^2 + x + 1 = x(x + 1) + 1$. This means that if we had $x^2 + x + 1 = xq(x)$, then x would divide $1 = xq(x) - x(x + 1)$. But that is impossible, as $0 = \deg(1) = \deg(x\{q(x) - x - 1\}) \ge 1$. A similar argument shows that $x + 1$ cannot divide $x^2 + x + 1$.

This means that **x^2+x+1 is the only irreducible polynomial of degree 2 in $\mathbb{Z}_2[x]$.**

degree 3 polynomials: $x^3,\ x^3 + 1,\ x^3 + x = x(x^2 + 1),\ x^3 + x + 1,\ x^3 + x^2 = x^2(x + 1),\ x^3 + x^2 + 1,\ x^3 + x^2 + x = x(x^2 + x + 1),\ x^3 + x^2 + x + 1$. Which of the polynomials $x^3 + 1,\ x^3 + x + 1,\ x^3 + x^2 + 1,\ x^3 + x^2 + x + 1$ are irreducible? To

answer this question more rapidly it helps to know that $x-a$ divides a polynomial $f(x)$ iff $f(a)=0$. Here we are assuming $a \in \mathbb{Z}_2$ and $f(x) \in \mathbb{Z}_2[x]$. We will prove this in a few pages as a corollary of the division algorithm.

The polynomial $f(x)$ of degree 3 will be reducible iff it has a factorization $f(x)=g(x)h(x)$ with $g(x), h(x) \in \mathbb{Z}_2[x]$, such that $\deg g \neq 0$ and $\deg h \neq 0$. But then $3 = \deg g + \deg h$ implies that either g or h has degree 1. This means that f is reducible iff $f(a)=0$ for some $a \in \mathbb{Z}_2$.

So we need to test $x^3+1,\ x^3+x+1,\ x^3+x^2+1,\ x^3+x^2+x+1$ for roots in $\mathbb{Z}_2$. However the only possible root is 1, since we have already eliminated the polynomials with 0 as a root. The polynomials with an even number of terms will have 1 as a root in $\mathbb{Z}_2[x]$.

This implies that **there are only 2 irreducible degree 3 polynomials in $\mathbb{Z}_2[x]$: x^3+x+1 and x^3+x^2+1.**

Exercise. Find the degree 4 irreducible polynomials in $\mathbb{Z}_2[x]$.

Exercise. Find the irreducible polynomials of degrees 1 and 2 in $\mathbb{Z}_3[x]$.

In order to do the same things for rings of polynomials $F[x]$, when F is a field, that we did for the ring $\mathbb{Z}$ of integers, we will need a division algorithm. The division algorithm works just as it did in high school or wherever it was. In fact, it really works the same way it did for the integers in elementary school (as in Section 5 of Part I).

Example 1. In $\mathbb{Z}_2[x]$, we have the following computation.

$$\begin{array}{rr|l} & x^3 & +1 \\ \hline x^2+x+1 & x^5+x^4+x^3 & +x^2+x+1 \\ \hline & x^5+x^4+x^3 & \\ \hline & & x^2+x+1 \\ \hline & & x^2+x+1 \\ \hline & & 0 \\ \hline \end{array}$$

As a result, we have $x^5+x^4+x^3+x^2+x+1 = \left(x^2+x+1\right)\left(x^3+1\right)$. The remainder is 0.

Example 2. In $\mathbb{Z}_3[x]$, we have the following computation.

$$\begin{array}{rr|l} & 2x & +1 \\ \hline 2x+1 & x^2+x & +2 \\ \hline & x^2+2x & \\ \hline & 2x & +2 \\ \hline & 2x & +1 \\ \hline & & 1 \\ \hline \end{array}$$

This says $x^2+x+2 = (2x+1)(2x+1)+1$. The remainder is 1. Note that we are definitely using the fact that $\mathbb{Z}_3$ is a field and $2^{-1} \equiv 2 (\mathrm{mod}\, 3)$. That is, $2 \cdot 2 \equiv 1 (\mathrm{mod}\, 3)$.

Theorem 24 ***(The Division Algorithm for Polynomial Rings)*** *Suppose that F is a field. Given $f(x)$ and $g(x) \in F[x]$ with $g(x)$ not the zero polynomial, there are polynomials $q(x)$ (the quotient) and $r(x)$ (the remainder) in $F[x]$ such that $f(x)=g(x)q(x)+r(x)$ and $\deg r < \deg g$ or r is the zero polynomial.*

Proof. Sketch (Induction on $\deg f$).

If f has degree 0, or if f is the zero polynomial, then the remainder will be f. Also, if $\deg g = 0$, the result is trivial as g is then a unit in the ring $F[x]$. So we assume $\deg g > 0$ from now on.

Induction Step. Now assume the theorem true if $\deg f \leq m-1$. Suppose that

$$f(x) = b_m x^m + \cdots \quad \text{and} \quad g(x) = a_n x^n + \cdots, \quad \text{with} \quad a_n \neq 0 \text{ and } b_m \neq 0.$$

We may assume $m \geq n$ or we can take $r=f$. Then we start the process by choosing the 1st term of $q(x)$ to be $a_n^{-1} b_m x^{m-n}$ so that $h(x) = f(x) - a_n^{-1} b_m x^{m-n} g(x)$ has degree less than $\deg f = n$.

$$\begin{array}{rr|l} & a_n^{-1} b_m x^{m-n} & + \cdots \\ \hline g(x) = a_n x^n + \cdots & f(x) = b_m x^m + \cdots & \\ \hline & b_m x^m + \cdots & \\ \hline & 0 & h = \text{ lower degree polynomial than } f \\ \hline \end{array}$$

This gets the induction going. The induction hypothesis allows us to divide h by g and we're done. ■

Exercise. Fill in the details of the preceding proof.

Exercise. Prove the uniqueness of the polynomials q and r in the division algorithm.

Corollary 25 *Suppose that F is a field, $f(x) \in F[x]$, and $a \in F$. Then $f(a) = 0 \iff f(x) = (x-a)q(x)$ for some $q(x) \in F[x]$.*

Proof. $\Longrightarrow$ By the division algorithm, $f(x) = (x-a)q(x) + r(x)$, where $\deg r < 1$ or $r = 0$. It follows that r must be a constant in F. Thus $f(a) = (a-a)q(a) + r = r$. If $f(a) = 0$, then r is the 0 polynomial and $f(x) = (x-a)q(x)$.

$\Longleftarrow$ Hopefully this is clear. ■

Corollary 26 *Suppose $f \in F[x]$ and $\deg f = n$. Then f has at most n roots in F counting multiplicity. This means that we count a not just once but k times if $(x-a)^k$ exactly divides $f(x)$ and $k > 1$ (meaning that $(x-a)^k$ divides $f(x)$ and $(x-a)^{k+1}$ does not divide $f(x)$).*

Proof. By the preceding corollary, $f(a) = 0$ implies $f(x) = (x-a)q(x)$ and $\deg f = n = 1 + \deg q$. Thus $\deg q = n - 1$. So we finish the proof by induction on the degree of f. ■

Corollary 27 *Every ideal in $F[x]$ is principal, when F is a field.*

Proof. Let I be an ideal in $F[x]$. If $I = \{0\} = \langle 0 \rangle$, we're done. Otherwise, let $f(x)$ be an element of I of minimal degree. Then we claim $I = \langle f \rangle$. To prove this, suppose $h \in I$. Then the division algorithm says there exist $q, r \in F[x]$ such that $h = qf + r$, with $\deg r < \deg f$ or $r = 0$. Then $r = h - qf \in I$ since $h, f \in I$. This contradicts the minimality of the degree of f unless $r = 0$. Then $h \in \langle f \rangle$ and $I = \langle f \rangle$. ■

Corollary 28 ***(Irreducible Polynomials correspond to Fields).*** *The following are equivalent in $F[x]$, when F =field.*

a) $\langle f(x) \rangle$ is a maximal ideal.

b) $f(x) \in F[x]$ is irreducible.

c) $F[x]/\langle f(x) \rangle$ = field.

Proof. We know from Section 4 that a)$\Longleftrightarrow$c). So let's show a)$\Longleftrightarrow$b).

a)$\Longrightarrow$b) Assume $\langle f \rangle$ is a maximal ideal. If $f = g \cdot h$, for $g, h \in F[x]$, such that neither g nor h is a unit, then $\langle f \rangle \subset \langle g \rangle \subset \langle 1 \rangle = F[x]$ and $\langle f \rangle \subset \langle h \rangle \subset \langle 1 \rangle = F[x]$ and none of the inclusions are equality. Why? Recall the exercise in Section 4 that said $\langle a \rangle = \langle b \rangle \Longleftrightarrow b = au$, for some unit u in $F[x]$. This contradicts the maximality of $\langle f \rangle$.

b)$\Longrightarrow$a) Suppose f is irreducible. We know by the preceding Corollary that every ideal in $F[x]$ is principal. So any ideal containing $\langle f \rangle$ must have the form $\langle g \rangle$, for some $g \in F[x]$. If $\langle f \rangle \subset \langle g \rangle \subset F[x]$, then $f = g \cdot h$ for some $h \in F[x]$. But the irreducibility of f says that either g or h is a unit. If g is a unit then $\langle g \rangle = F[x]$. If h is a unit, then $\langle f \rangle = \langle g \rangle$. Thus $\langle f \rangle$ is maximal. ■

Exercise. a) In $\mathbb{Z}_3[x]$ show that the polynomials $f(x) = x^4 + x$ and $g(x) = x^2 + x$ determine the same function mapping $\mathbb{Z}_3$ into $\mathbb{Z}_3$.

b) In $\mathbb{Z}_7[x]$ find the quotient and remainder upon dividing $f(x) = 5x^4 + 3x^3 + 1$ by $g(x) = 3x^2 + 2x + 1$.

c) Find all degree 2 irreducible polynomials in $\mathbb{Z}_3[x]$ with lead coefficient equal to 1.

Integral domains with a division algorithm are called **Euclidean domains**. Thus the ring of polynomials over a field F is an Euclidean domain. As such it has similar properties to $\mathbb{Z}$. One defines the greatest common divisor $d = \gcd(f, g)$ for $f, g \in F[x]$, to be the unique monic polynomial which divides both f and g such that any common divisor h of f and g must divide d. Again there is an Euclidean algorithm to compute d. One has the analog of the theorem in Section 5 of Part I saying $d = uf + vg$ for some $u, v \in F[x]$ and the Euclidean algorithm can be used to find u and v.

Exercise. Prove the preceding statements about $\gcd(f, g)$ for $f, g \in F[x]$ by imitating the proofs that worked in $\mathbb{Z}$.

Exercise. Suppose x is a non-0 element of a finite field F with n elements. Show that $x^{n-1} = 1$,

Exercise. a) Consider $\mathbb{Z}_5[i] = \{a + bi \mid a, b \in \mathbb{Z}_5\}$, where $i^2 = -1$. Show that this ring is not a field.

b) Consider $\mathbb{Z}_7[i] = \{a + bi \mid a, b \in \mathbb{Z}_7\}$, where $i^2 = -1$. Show that this ring is a field.

c) Can you develop a more general version of this problem for $\mathbb{Z}_p[i]$ where p is an odd prime according to whether p is congruent to 1 or 3 (mod 4)?

Exercise. Show that the ideal $\langle a(x), b(x) \rangle$ in $F[x]$ is $\langle \gcd(a(x), b(x)) \rangle$.

Exercise. a) Show that if an ideal A of ring R contains an element of the unit group R^*, then $A = R$.

b) Show that the only ideals of a field F are $\{0\}$ and F itself.

Example. A Field with 8 elements is $\mathbb{F}_8 = \mathbb{Z}_2[x]/\langle x^3 + x + 1 \rangle$.

To see this, you just have to recall a few of the facts that we proved here. First we know from earlier in this Section that x^3+x+1 is irreducible in $\mathbb{Z}_2[x]$. Second, we showed that $\langle f\rangle$ is maximal in $\mathbb{Z}_2[x]$ iff f is irreducible in $\mathbb{Z}_2[x]$, And we know from Section 4 that $\mathbb{Z}_2[x]/I$ is a field iff the ideal I is maximal. Thus $\mathbb{Z}_2[x]/\langle x^3+x+1\rangle$ is a field.

How do we know that our field has 2^3 elements? The answer is that a coset $[f]$ in $\mathbb{Z}_2[x]/\langle x^3+x+1\rangle$ is represented by the remainder of f upon division by x^3+x+1. The remainder has degree ≤ 2 and thus has the form ax^2+bx+c, where $a,b,c\in\mathbb{Z}_2$. Moreover 2 polynomials g,h of degree ≤ 2 cannot be congruent modulo x^3+x+1 unless they are equal. Why? Congruent means the difference $g-h$ is a multiple of x^3+x+1. The only way a degree ≤ 2 polynomial like $g-h$ can be a multiple of a degree 3 polynomial is if $g-h$ is really the 0 polynomial. Thus g must equal h.

The preceding is analogous to what happens in $\mathbb{Z}/163\mathbb{Z}$. The elements $[m]$ are represented by $[r]$, where r is the remainder of m upon division by 163.

We can think $\Theta=[x]$ in $\mathbb{Z}_2[x]/\langle x^3+x+1\rangle$. This means Θ is a root of $x^3+x+1=0$. We are saying that $\mathbb{F}_8=\left\{a\Theta^2+b\Theta+c \mid a,b,c\in\mathbb{Z}_2\right\}$. Moreover we can view $\mathbb{F}_8$ as a vector space over the field $\mathbb{Z}_2$. A basis for this vector space is $\left\{1,\Theta,\Theta^2\right\}$. See Section 10 for more information on vector spaces over finite fields.

If we express the elements of $\mathbb{F}_8$ in the form $a\Theta^2+b\Theta+c$, for $a,b,c\in\mathbb{Z}_2$, it is easy to add the elements but hard to multiply. Thus it is useful to show that the multiplicative group of $\mathbb{F}_8$ is cyclic - a fact that can be proved for any finite field. It turns out the a generator in this case is Θ. This is a fact that does not always hold for a finite field since only **primitive polynomials** $k(x)$ over the base field $\mathbb{Z}_p$ will have this property that a root of $k(x)$ is a generator of $(\mathbb{Z}_p[x]/\langle k(x)\rangle)^*$.

Next we create a table of powers of $\Theta=[x]$ in $\mathbb{Z}_2[x]/\langle x^3+x+1\rangle$. We know that $\Theta^3+\Theta+1=0$. This implies that $\Theta^3=-\Theta-1=\Theta+1$ since $-1=+1$ in $\mathbb{Z}_2$. Next we note that $\Theta(a_0+a_1\Theta+a_2\Theta^2)=a_0\Theta+a_1\Theta^2+a_2\Theta^3=a_0\Theta+a_1\Theta^2+a_2(\Theta+1)=a_2+(a_0+a_2)\Theta+a_1\Theta^2$.

So multiplication by Θ sends the coefficients (a_0,a_1,a_2) to (a_2,a_0+a_2,a_1). This is what is called a "**feedback shift register**." So now it is easy to make a table of powers of Θ. The jth row will list the coefficients (a_0,a_1,a_2) of $\Theta^j=a_0+a_1\Theta+a_2\Theta^2$.

$\Theta^j=a_0+a_1\Theta+a_2\Theta^2$	a_0	a_1	a_2
Θ	0	1	0
Θ^2	0	0	1
Θ^3	1	1	0
Θ^4	0	1	1
Θ^5	1	1	1
Θ^6	1	0	1
$\Theta^7=1$	1	0	0

This shows that the multiplicative group of units $\mathbb{F}_8^*$ is a cyclic group of order 7. We call x^3+x+1 a primitive polynomial in $\mathbb{Z}_2[x]$ for this reason. Figure 4 shows the picture of the feedback shift register corresponding to this polynomial. You can use primitive polynomials to construct other feedback shift registers. It is a finite state machine that will cycle through 2^n-1 states if $f(x)$ is a primitive polynomial of degree n in $\mathbb{Z}_2[x]$. The states are really the rows of the table of powers of Θ for Θ a root of $f(x)$, and this is really the unit group of the finite field $\mathbb{Z}_2[x]/\langle f(x)\rangle$. The successive states of the registers are given in the preceding table.

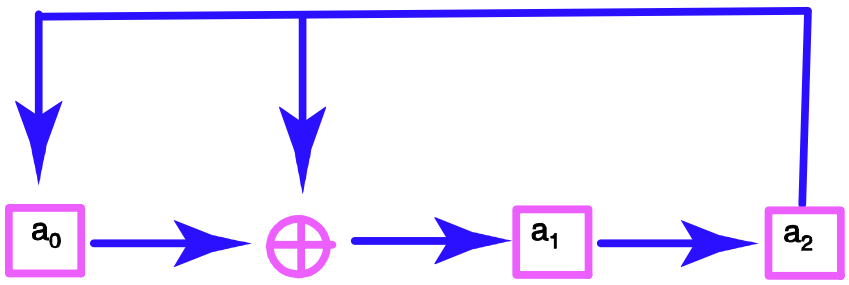

Figure 4: A feedback shift register diagram corresponding to the finite field $\mathbb{Z}_2[x]/\langle x^3+x+1\rangle$ and the multiplication table given in the text.

Feedback shift registers are of interest in generating pseudo-random numbers and in cryptography. There are applications in digital broadcasting, communications, and error-correcting codes.

Exercise. a) Show that x^2-2 is an irreducible polynomial in $\mathbb{Z}_5[x]$.

b) Show that the factor ring $\mathbb{Z}_5[x]/ < x^2-2 >$ is a field with 25 elements.

c) Show that the field in part b) can be identified with $\mathbb{Z}_5[\sqrt{2}] = \{a+b\sqrt{2} | a,b \in \mathbb{Z}_5\}$ = the smallest field containing $\mathbb{Z}_5$ and $\sqrt{2}$.

d) What is the characteristic of the field with 25 elements in parts b and c?

Exercise. Identify $\mathbb{F}_{25} \cong \mathbb{Z}_5[x]/\left\langle x^2-2\right\rangle$ as the set $\mathbb{Z}_5[\sqrt{2}]$ as in the preceding problem. Set $\theta = \sqrt{2}$. Find the table of powers $\Theta^j = a_0 + a_1\Theta$, where $\Theta^2 = 2$, in a similar manner to the table we created for $\mathbb{Z}_2[x]/\left\langle x^3+x+1\right\rangle$. Do these powers give the whole unit group in $\mathbb{Z}_5[x]/ < x^2-2 >$? This would say that the polynomial x^2-2 is primitive in $\mathbb{Z}_5[x]$? What is the feedback shift register diagram corresponding to this polynomial? How many states does it cycle through before it repeats?

Hint. The order of the unit group $\mathbb{F}_{25}^*$ is 24 while the order of 2 is 4 and thus the order of $\sqrt{2}$ is 8.

Exercise. a) Show that the ideal $< x >$ in $\mathbb{Z}_3[x]$ is maximal.

b) Prove that $\mathbb{Z}_3$ is isomorphic to $\mathbb{Z}_3[x]/ < x >$.

Exercise. a) Show that $\mathbb{Z}_3[x]/ < x^2+x+2 >$ is a field $\mathbb{F}_9$ with 9 elements which can be viewed as the field $\mathbb{Z}_3[\theta] = \{a+b\theta | a,b \in \mathbb{Z}_3\}$, where $\theta^2+\theta+2=0$.

b) Imitating the table in above, compute the powers of θ from part a).

c) Is the multiplicative group $\mathbb{F}_9^*$ in the field of part a) cyclic?

d) Draw the corresponding feedback shift register diagram as in Figure 4.

Exercise. Find an infinite set of polynomials $f(x) \in \mathbb{Z}_3[x]$ such that $f(a) = 0$ for all $a \in \mathbb{Z}_3$.

6 Ring Homomorphisms

We need to discuss the ring analog of group homomorphism from Part I.

Definition 29 *Suppose that R and S are rings. Then $T : R \to S$ is a **ring homomorphism** iff T preserves both ring operations; i.e.,*

$$T(a+b) = T(a) + T(b) \quad \text{and} \quad T(ab) = T(a)T(b), \quad \text{for all} \quad a,b \in R.$$

*If, in addition, T is 1-1 and onto, we say that T is a **ring isomorphism** and write $R \cong S$.*

For most purposes, we can identify isomorphic rings, just as we can identify isomorphic groups.

Example. $\pi : \mathbb{Z} \to \mathbb{Z}_n$ defined by $\pi(a) = a (\mathrm{mod}\, n)$ is a ring homomorphism and is onto but not 1-1. This example is easily generalized to $\pi : R \to R/I$, where I is any ideal in a ring R.

Application. Test for Divisibility by 3.

Any integer has a decimal expansion which we write $n = a_k a_{k-1} \cdots a_1 a_0$, for $a_j \in \{0,1,2,3,4,5,6,7,8,9\}$, where this means that $n = a_k 10^k + a_{k-1}10^{k-1} + \cdots + a_1 10 + a_0$. Then we have 3 divides $n = a_k a_{k-1} \cdots a_1 a_0$ iff 3 divides $a_k + a_{k-1} + \cdots + a_1 + a_0$ =the sum of the digits of n.

Proof. Look at the homomorphism $\pi : \mathbb{Z} \to \mathbb{Z}_3$ defined by $\pi(a) = a (\mathrm{mod}\, 3)$. Then, since $\pi(10) \equiv 1 (\mathrm{mod}\, 3)$, we have

$$\begin{aligned} \pi(n) &= \pi\left(a_k 10^k + a_{k-1}10^{k-1} + \cdots + a_1 10 + a_0\right) \\ &= \pi(a_k)\,\pi(10)^k + \pi(a_{k-1})\,\pi(10)^{k-1} + \cdots + \pi(a_1)\,\pi(10) + \pi(a_0). \\ &\equiv a_k + a_{k-1} + \cdots + a_1 + a_0 (\mathrm{mod}\, 3). \end{aligned}$$

■

Example. Does 3 divide 314159265358979323846? We compute the sum of the digits to be 103 and then the sum of those digits is 4 which is not divisible by 3. So the answer is "No."

Exercise. Does 3 divide 271828182845904523536 ?

Exercise. Create a similar test to see whether 11 divides a number? Then use your theorem to see whether 11 divides the numbers in the 2 preceding exercises.

The following theorem shows that ring homomorphisms have analogous properties to group homomorphisms.

Theorem 30 ***(Properties of Ring Homomorphisms).*** *Suppose that S and S' are rings and $T : S \to S'$ is a ring homomorphism. Let 0 be the identity for addition in S and 1 the identity for multiplication in S, if S has an identity for multiplication. Then let $0'$ be the identity for addition in S'. Then we have the following facts.*

1) a) $T(0) = 0'$.

b) If S has an identity for multiplication and $T(1) \neq T(0)$, then $T(1)$ is the identity for multiplication in the image ring $T(S)$. Here the image of S is $T(S) = \{T(s) \mid s \in S\}$.

2) The image $T(S)$ is a subring of S'. If S is a field, then $T(1) \neq T(0)$ implies that the image $T(S)$ is a field.

3) Define the ***kernel*** *of T to be $\ker T = T^{-1}(0') = \{x \in S \mid T(x) = 0'\}$. Then $\ker T$ is an ideal in S. And T is 1-1 iff $\ker T = \{0\}$.*

4) ***(First Isomorphism Theorem).*** *$S/\ker T \cong T(S)$.*

Proof. 1) a) follows from the corresponding fact about groups, since S and S' are groups under addition.

1) b) To do this, we must think a little since S, S' are not groups under multiplication unless something weird happens like $S = S' = \{0\}$. Also we need part 2) to know that the image $T(S)$ is a ring. Then, if $T(1) \neq T(0)$ and $a \in S$, we have

$$\begin{aligned} T(1)T(a) &= T(1 \cdot a) = T(a), \\ T(a)T(1) &= T(a \cdot 1) = T(a). \end{aligned}$$

It follows that $T(a)$ is the identity for multiplication in $T(S)$.

2) The image $T(S)$ is an additive subgroup of S' from results of Section 18 of Part I. To see that $T(S)$ is closed under multiplication, note that $T(a)T(b) = T(ab) \in T(S)$, for all $a, b \in S$. The associative law for multiplication and distributive laws follow from those for S; e.g., $T(a)\left(T(b) + T(c)\right) = T(a(b+c)) = T(ab+ac) = T(a)T(b) + T(a)T(c)$. Then if S is a field, from part b) we know that $T(1)$ is the identity for multiplication in $T(S)$. Since $S^* = S - \{0\}$ is a group under multiplication, we can use results from Part I, Section 18 to see that $T(S^*) = T(S)^*$ is a group under multiplication.

3) We know that $\ker T$ is an additive subgroup of S' by results of Section 18 of Part I. To show $\ker T$ is an ideal we also need to show that if $a \in \ker T$ and $s \in S$, then sa and as are in $\ker T$. This is easy since $T(sa) = T(s)T(a) = T(s)0 = 0$ implies $sa \in \ker T$. The same argument works for as.

4) We imitate the proof of the 1st Isomorphism Theorem for groups in Part I, Section 18. As before, we define a map $\widetilde{T} : S/\ker T \to T(S)$ by setting $\widetilde{T}\left([a]\right) = \widetilde{T}(a + \ker T) = T(a)$. Then we need to show that $\widetilde{T}$ is a ring isomorphism.

$\widetilde{T}$ is **well defined** since if $[a] = a + \ker T = [b]$, then $b = a + u$, where $u \in \ker T$. This implies $\widetilde{T}(b) = T(b) = T(a+u) = T(a) + T(u) = T(a) + 0 = T(a) = \widetilde{T}(a)$.

$\widetilde{T}$ is **1-1** since $\widetilde{T}([a]) = \widetilde{T}([b])$ implies $T(a) = T(b)$ and thus $a - b \in \ker T$ and $[a] = [b]$.

$\widetilde{T}$ is **onto** since any element of $T(S)$ has the form $T(a)$ for some $a \in S$. Thus $\widetilde{T}\left([a]\right) = T(a)$.

$\widetilde{T}$ **preserves both ring operations** since for any $a, b \in S$, we have the following, using the definition of addition and multiplication in the quotient $S/\ker T$, the definition of $\widetilde{T}$, and the fact that T is a ring homomorphism:

$$\begin{aligned} \widetilde{T}\left([a] + [b]\right) &= \widetilde{T}\left([a+b]\right) = T(a+b) = T(a) + T(b) = \widetilde{T}\left([a]\right) + \widetilde{T}\left([b]\right), \\ \widetilde{T}\left([a] \cdot [b]\right) &= \widetilde{T}\left([a \cdot b]\right) = T(a \cdot b) = T(a) \cdot T(b) = \widetilde{T}\left([a]\right) \cdot \widetilde{T}\left([b]\right). \end{aligned}$$

■

Example. Define $\mathbb{Z}_3[i] = \{x + iy \mid x, y \in \mathbb{Z}_3\}$, where $i^2 = -1$. We can use the 1st isomorphism theorem to show that

$$\mathbb{Z}_3[i] \cong \mathbb{Z}_3[x]/\left\langle x^2 + 1 \right\rangle.$$

Note that the right-hand side is a field because $x^2 + 1$ is irreducible in $\mathbb{Z}_3[x]$ since -1 is not a square mod 3 means $x^2 + 1$ has no roots in $\mathbb{Z}_3$. The left-hand side can be shown directly to be a field by proving that it is possible to find the multiplicative inverse of any non-0 element.

First we define a ring homomorphism $T : \mathbb{Z}_3[x] \to \mathbb{Z}_3[i]$ by $T(f(x)) = f(i)$ for any polynomial $f(x) \in \mathbb{Z}_3[x]$. The map T is well-defined, preserves the ring operations and is onto. We leave this as an **Exercise**. For example, one must show that

$$T(f+g) = (f+g)(i) = f(i) + g(i) = T(f) + T(g)$$

and

$$T(f \cdot g) = (f \cdot g)(i) = f(i) \cdot g(i) = T(f) \cdot T(g).$$

Prove these facts first for $f(x) = ax^r$.

We claim $\ker T = \langle x^2+1 \rangle$. To see this, note 1st that $\langle x^2+1 \rangle \subset \ker T$, since $i^2+1=0$. To prove $\ker T \subset \langle x^2+1 \rangle$, let $g(x) \in \ker T$. By the division algorithm, we have $g(x) = (x^2+1)\,q(x) + r(x)$, where $\deg r < 2$ or $r = 0$. Since $g(i) = 0$, we see that $r(i) = 0$. But if $r \neq 0$, $\deg r = 0$ or 1, and we have $r(x) = ax + b$, with $a, b \in \mathbb{Z}_3$. But then $ai + b = 0$. If $a \neq 0$, this would say $i = -b/a \in \mathbb{Z}_3$, a contradiction to the fact that -1 is not a square in $\mathbb{Z}_3$. Thus r must be the 0-polynomial and $\ker T \subset \langle x^2+1 \rangle$ to complete the proof that $\ker T = \langle x^2+1 \rangle$.

It follows then from the 1st isomorphism theorem that $\mathbb{Z}_3[i] \cong \mathbb{Z}_3[x]/\langle x^2+1 \rangle$.

Exercise. Suppose that R, S are rings, A is a subring of R, B is an ideal of S. Let $T : R \to S$ be a ring homomorphism.

a) Show that for all $r \in R$ and all $n \in \mathbb{Z}^+$, we have $T(nr) = nT(r)$ and $T(r^n) = T(r)^n$.

b) Show that $T(A)$ is a subring of S.

c) Show that if A is an ideal in R and $T(R) = S$, then $T(A)$ is an ideal of S.

Exercise. Under the same hypotheses as in the preceding Exercise, prove:

a) $T^{-1}(B)$ is an ideal of R. Here the **inverse image** of B under T is $T^{-1}(B) = \{a \in A \mid T(a) \in B\}$. We are not assuming the inverse function of T exists.

b) If T is a isomorphism of R onto S, then T^{-1} is an isomorphism of S onto R.

Exercise. a) If we try to define $T : \mathbb{Z}_5 \to \mathbb{Z}_{10}$ by setting $T(x) = 5x$ we don't really have a well defined function. Explain.

b) Show that $T : \mathbb{Z}_4 \to \mathbb{Z}_{12}$ defined by $T(x) = 3x$ is well defined but does not preserve multiplication.

c) Show that every homomorphism $T : \mathbb{Z}_n \to \mathbb{Z}_n$ has the form $T(x) = ax$ for some fixed a in $\mathbb{Z}_n$ with $a^2 = a$.

Exercise. a) Show that the ring of complex numbers $\mathbb{C}$ is isomorphic to the factor ring $\mathbb{R}[x]/ < x^2+1 >$. Here $\mathbb{R}[x]$ is the ring of polynomials in 1 indeterminate x and real coefficients.

b) Show that complex conjugation $\phi(a+ib) = a - ib$, for a, b in $\mathbb{R}$ and $i^2 = -1$, defines a ring isomorphism from $\mathbb{C}$ onto $\mathbb{C}$.

c) Show that $\mathbb{C}$ is not isomorphic to $\mathbb{R}$.

d) Show that $\mathbb{C}$ is isomorphic to the ring $\left\{ \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \middle| a, b \in \mathbb{R} \right\}$, where the operations are the usual matrix addition and multiplication.

Exercise. a) Show that the only ring homomorphisms from the rationals $\mathbb{Q}$ to $\mathbb{Q}$ are the map $T(x) = 0$, $\forall x \in \mathbb{Q}$ and the identity $I(x) = x$, $\forall x \in \mathbb{Q}$. **Hint:** First look at the map on $\mathbb{Z}$.

b) Show that the only ring isomorphism ϕ mapping the reals $\mathbb{R}$ onto $\mathbb{R}$ is the identity map. **Hint:** First, recall that the positive reals are squares of non-zero reals and vice versa. Then recall that $a < b \iff b - a > 0$. Use this to show that $a < b$ implies $\phi(a) < \phi(b)$. Then suppose that $\exists a$ s.t. $\phi(a) \neq a$. Consider the 2 cases that $a < \phi(a)$ and $\phi(a) < a$. There is a rational number between a and $\phi(a)$. Use the fact that ϕ must be the identity on the rationals to get a contradiction.

7 The Chinese Remainder Theorem

An example of the Chinese remainder theorem can be found in a manuscript by Sun Tzu (or Sun Zi) from the 3rd century AD.

Theorem 31 ***(The Chinese Remainder Theorem for Rings).*** *Assume that the positive integers m, n satisfy $\gcd(m,n) = 1$. The mapping $\widetilde{T} : \mathbb{Z}_{mn} \to \mathbb{Z}_m \oplus \mathbb{Z}_n$ defined by $\widetilde{T}([s]) = T(s) = (s \pmod m, s \pmod n)$ is a ring isomorphism showing that $\mathbb{Z}_{mn}$ is isomorphic to the ring $\mathbb{Z}_m \oplus \mathbb{Z}_n$.*

Proof. First consider the mapping $T : \mathbb{Z} \to \mathbb{Z}_m \oplus \mathbb{Z}_n$ defined by $T(s) = (s \pmod m, s \pmod n)$. Then T is a ring homomorphism. Note that we already showed it is an additive group homomorphism in Part I, Section 19. To see that it preserves multiplication: let $a, b \in \mathbb{Z}$. Then, using the definition of T and the definition of multiplication in $\mathbb{Z}_m \oplus \mathbb{Z}_n$, we have:

$$T(a \cdot b) = (a \cdot b \pmod m, a \cdot b \pmod n) = (a \pmod m, a \pmod n) \cdot (b \pmod m, b \pmod n) = T(a) \cdot T(b).$$

It follows from the 1st isomorphism theorem that $\mathbb{Z}/\ker T$ is isomorphic to $T(\mathbb{Z})$. So we need to compute the kernel of T. This is, since $\gcd(m,n) = 1$,

$$\ker T = \{a \in \mathbb{Z} | a \equiv 0 (\bmod m) \text{ and } a \equiv 0 (\bmod n)\} = \{a \in \mathbb{Z} | m \text{ divides } a \text{ and } n \text{ divides } a\} = mn\mathbb{Z}.$$

The map $\widetilde{T} : \mathbb{Z}_{mn} \to \mathbb{Z}_m \oplus \mathbb{Z}_n$ is defined by $\widetilde{T}([s]) = T(s) = (s \,(\mathrm{mod}\, m)\,, s \,(\mathrm{mod}\, n))$. This map $\widetilde{T}$ is 1-1 since $\widetilde{T}^{-1}((0,0)) = \{0\}$ by a fact from Section 18 of Part I. By the pigeonhole principle, $\widetilde{T}$ must be onto because both $\mathbb{Z}_{mn}$ and $\mathbb{Z}_m \oplus \mathbb{Z}_n$ have mn elements. ■

In particular, the Chinese Remainder Theorem says that if $\gcd(m, n) = 1$, there is a solution $x \in \mathbb{Z}$ to the simultaneous congruences:

$$\begin{cases} x \equiv a(\mathrm{mod}\, m) \\ x \equiv b(\mathrm{mod}\, n). \end{cases}$$

When the theorem is discussed in elementary number theory books, only the onto-ness of the function is emphasized. Many examples like the following are given. This result and its generalizations have many applications; e.g., to rapid and high-precision computer arithmetic. See the next section or L. Dornhoff and F. Hohn, *Applied Modern Algebra;* D. E. Knuth, *Art of Computer Programming, II*; I. Richards, Number Theory in *Math. Today* edited by L. A. Steen; K. Rosen, *Elementary Number Theory;* or A. Terras, *Fourier Analysis on Finite Groups and Applications.*

Example. Solve the following simultaneous congruences for x, y:

$$\begin{cases} 3x \equiv 1(\mathrm{mod}\, 5) \\ 2x \equiv 3(\mathrm{mod}\, 7). \end{cases}$$

The 1st congruence has the solution $x \equiv 2(\mathrm{mod}\, 5)$ as one can find by trial and error. Then put $x = 2 + 5u$ into the 2nd congruence. This gives

$2x = 2(2+5u) \equiv 3(\mathrm{mod}\, 7)$ and thus $4+10u \equiv 3(\mathrm{mod}\, 7)$. This becomes $3u \equiv 6(\mathrm{mod}\, 7)$. One immediately sees a solution $u \equiv 2(\mathrm{mod}\, 7)$. This means that $u = 2+7t$. Plug this back into our formula for x and get $x = 2+5(2+7t) = 12+35t$. This means $x \equiv 12(\mathrm{mod}\, 35)$. You should check that it works.

There are many ways to understand the Chinese remainder theorem. The 1st step is to extend it to an arbitrary number of relatively prime moduli. If the positive integers m_i satisfy $\gcd(m_1, ..., m_r) = 1$, and $m = m_1 m_2 \cdots m_r$, then the rings $\mathbb{Z}_m$ and $\mathbb{Z}_{m_1} \oplus \cdots \oplus \mathbb{Z}_{m_r}$ are isomorphic under the mapping $f(x \bmod m) = (x \bmod m_1, ..., x \bmod m_r)$. We leave the proof of this as an **exercise**.

Let's look at the case $r = 2$ again. To create the isomorphism between $\mathbb{Z}_{15}$ and $\mathbb{Z}_3 \oplus \mathbb{Z}_5$, for example, you can make a big table of positive integers.

	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	1														
2		2													
3			3												
1				4											
2					5										
3						6									
1							7								
2								8							
3									9						
1										10					
2											11				
3												12			
1													13		
2														14	
3															15

Next we fill in the blanks upper left 3×5 part of the table by taking the 1st number left out which is 4 and moving it up 3 rows. Similarly we move 5 up 3 rows. The next number 6 must be moved up 3 rows and then moved left 5 columns.

	1	2	3	4	5
1	1			4	
2		2			5
3	6		3		

Continue in this way to complete the table.

	1	2	3	4	5
1	1	7	13	4	10
2	11	2	8	14	5
3	6	12	3	9	15

Since the isomorphism preserves addition and multiplication we can compute mod 17 by computing mod 3 and mod 5. This is not so impressive but it would work better if we took a lot of primes like $m = 2^5 \cdot 3^3 \cdot 5^2 \cdot 7^2 \cdot 11 \cdot 13 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 37 \cdot 41 \cdot 43 \cdot 47$ which is $> 3 \cdot 10^{21}$. Then computing mod m would be the same as computing mod m_i, for $m_1 = 2^5, m_2 = 3^3, ..., m_{15} = 47$. See Dornhoff and Hohn, *Applied Modern Algebra*, p. 238 for more information.

In another visualization, we see the additive group $\mathbb{Z}_{15}$ as a discrete circle by taking the Cayley graph $X(\mathbb{Z}_{15}, \{\pm 1 (\mathrm{mod}\, 15)\})$. See Figure 5. Now we can view the same group as a 2 dimensional product of the circles $\mathbb{Z}_3$ and $\mathbb{Z}_5$. This is a torus or doughnut graph. It is also the Cayley graph $X(\mathbb{Z}_{15}, \{5, 6, 9, 10 (\mathrm{mod}\, 15)\})$ in Figure 6.

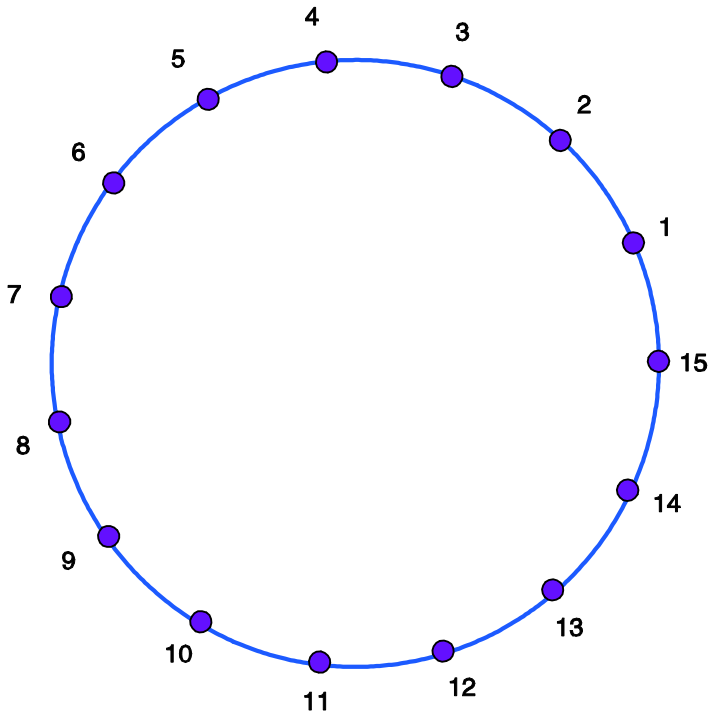

Figure 5: The Cayley graph $X(\mathbb{Z}_{15}, \{\pm 1 (\mathrm{mod}\, 15)\})$

Exercise. Draw analogous figures to Figures 5 and 6 for $\mathbb{Z}_{35}$.

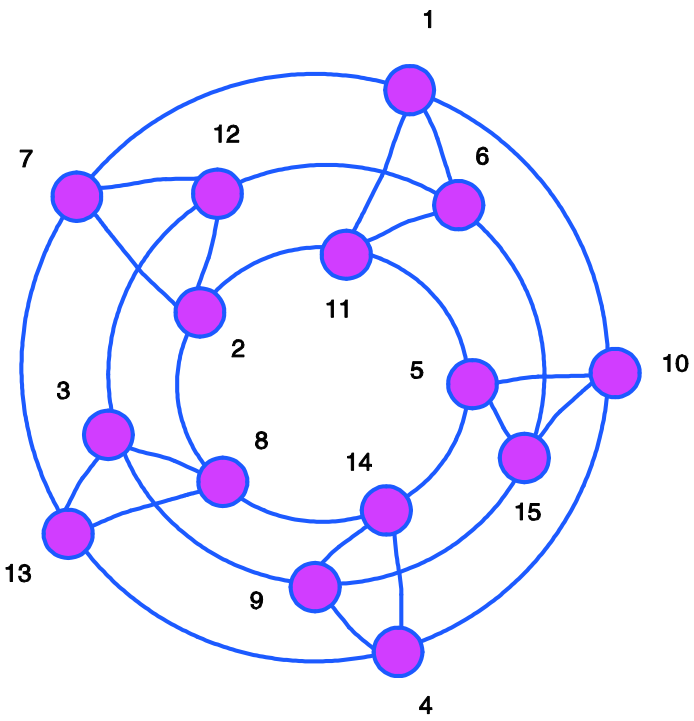

Figure 6: The Cayley graph $X(\mathbb{Z}_{15}, \{5, 6, 9, 10 (\mathrm{mod}\, 15)\})$

As we noted above, the most important thing for a number theorist is the onto-ness of the isomorphism $\widetilde{T} : \mathbb{Z}_{mn} \to \mathbb{Z}_m \oplus \mathbb{Z}_n$ for $\gcd(m, n) = 1$, defined by $\widetilde{T}([s]) = T(s) = (s \pmod m, s \pmod n)$. We want to discuss an old method to give an explicit formula for the solution of the simultaneous congruences that express this onto-ness. For example, suppose that $m = 3$ and $n = 5$, and we want to solve

$$\begin{cases} x \equiv a (\mathrm{mod}\, 3) \\ x \equiv b (\mathrm{mod}\, 5). \end{cases} \tag{4}$$

Then we first solve 2 sets of simultaneous congruences:

$$\begin{cases} 5u \equiv 1 (\mathrm{mod}\, 3) \\ u \equiv 0 (\mathrm{mod}\, 5) \end{cases} \qquad \text{and} \qquad \begin{cases} v \equiv 0 (\mathrm{mod}\, 3) \\ 3v \equiv 1 (\mathrm{mod}\, 5). \end{cases}$$

and set $x = 5au + 3bv (\mathrm{mod}\, 15)$. It is easily checked that x does solve the problem of formula (4).

Exercise. Find an analogous procedure to that of the last paragraph to solve

$$\begin{cases} x \equiv a (\mathrm{mod}\, 3) \\ x \equiv b (\mathrm{mod}\, 5) \\ x \equiv c (\mathrm{mod}\, 7). \end{cases}$$

What is the solution when $a = 2, b = 4, c = 1$?

Exercise. Show that $\mathbb{Z}_3[x]/\left\langle x^2 - 1 \right\rangle \cong \mathbb{Z}_3 \oplus \mathbb{Z}_3$.

There are also applications of the Chinese remainder theorem in RSA cryptography, secret sharing, the Fast Fourier transforms, and proving the Gödel incompleteness theorem in logic.

There are many puzzles related to the Chinese remainder theorem. Some are ancient. The following is a puzzle found in O. Ore, *Number Theory and its History*, pp. 118 ff.

"An old woman goes to market and a horse steps on her basket and crushes the eggs. The rider offers to pay for the damage and asks her how many eggs she had brought. She does not remember the exact number, but when she had taken them out 2 at a time there was 1 egg left. The same happened when she picked them out 3,4,5, and 6 at a time, but when she took them out 7 at a time they came out even. What is the smallest number of eggs she could have had?"

Exercise. Solve the preceding puzzle.

8 Comparisons Between $\mathbb{Z}$ and $F[x]$

Suppose that F is a field and $F[x]$ is the ring of polynomials in one indeterminate with coefficients in F. We want to create a table of comparisons between the ring of integers $\mathbb{Z}$ and the ring $F[x]$.

Table Comparing $\mathbb{Z}$ and $F[x]$

property	$\mathbb{Z}$	$F[x]$
infinite ring	yes	yes
integral domain	yes	yes
unit group	$\{1,-1\}$	F^*
division algorithm	$n = mq + r, 0 < r \leq m$	$f(x) = g(x)q(x) + r(x)$, $r = 0$ or $\deg r < \deg g$
divisibility	$m\|n \iff n = mq$, for some $q \in \mathbb{Z}$	$g(x)\|f(x) \iff f(x) = g(x)q(x)$, for some $q(x) \in F[x]$
prime	$p > 1$ s.t. $p = a \cdot b \Longrightarrow$ either a or b is a unit	$f(x)$ monic irreducible, $f(x) \neq$ constant polynomial, $f = g \cdot h$ implies either g or h is a unit
unique factorization into primes	$n \neq 0 \Longrightarrow n = \pm p_1 p_2 \cdots p_r$, $p_i =$ prime, factorization unique up to order	$f(x) \neq 0 \Longrightarrow f(x) = \text{unit} \cdot p_1(x) \cdots p_r(x)$, $p_i(x)$ monic irreducible, factorization unique up to order
every ideal principal	$I = \langle n \rangle$, n least positive element of I, if $I \neq \{0\}$	$I = \langle f \rangle$, f element of I of least degree if $I \neq \{0\}$
maximal ideal	$I = \langle p \rangle$, $p =$prime	$I = \langle f(x) \rangle$, $f(x)$ irreducible
$R/I =$ field when I maximal	$\mathbb{Z}/p\mathbb{Z}$=field when $p =$prime	$F[x]/\langle f(x) \rangle$ =field when f irreducible
Euclidean algorithm for $\gcd(a,b) = na + mb$ (Bezout's identity)	yes	yes
Euclid's Lemma	$p =$prime, $p\|ab \Longrightarrow p\|a$ or $p\|b$	$f(x)$ irreducible, $f(x)\|a(x)b(x) \Longrightarrow f(x)\|a(x)$ or $f(x)\|b(x)$

Exercise. **a)** Compute $h(x) = \gcd(x^2+1, x^4+x^3+x^2+1)$ in $\mathbb{Z}_2[x]$. Find polynomials $u(x)$ and $v(x)$ in $\mathbb{Z}_2[x]$ such that $h(x) = u(x)\left(x^2+1\right) + v(x)\left(x^4+x^3+x^2+1\right)$.

b) Factor x^3+1 as a product of monic irreducible polynomials in $\mathbb{Z}_7[x]$.

c) Write $x^4+x^3+x^2+1$ as a product of monic irreducible polynomials in $\mathbb{Z}_2[x]$.

Exercise. Prove the analog of Euclid's Lemma for $F[x]$, where F is a field. .

Exercise. a) Assume p=prime. How many irreducible polynomials of the form $f(x) = x^2 + ax + b$ are there in $\mathbb{Z}_p[x]$?

b) Show that for every prime p, there exists a field with p^2 elements.

Exercise. a) Find all zeros of $f(x) = 3x^2 + x + 4$ in $\mathbb{Z}_7$ by the process of substituting all elements of $\mathbb{Z}_7$.

b) Find all zeros of the polynomial $f(x)$ in part a) using the quadratic formula for $\mathbb{Z}_7$. Do your answers agree? Should they?

c) Same as a) for $g(x) = 2x^2 + x + 3$ over $\mathbb{Z}_5$.

d) Same as b) for $g(x)$ in part c).

Exercise. Assume that F is a field. Prove that every polynomial $f(x) \in F[x]$ factors as $f(x) = \text{unit} \cdot p_1(x) \cdots p_r(x)$, where $p_i(x) =$ monic irreducible, and this factorization is unique up to order.

Next let's consider a few facts about $F[x]$ and its quotients.

Proposition 32 ***(Irreducibility Test for Low Degree Polynomials).*** *Suppose F is a field, $f(x) \in F[x]$ and $\deg f = 2$ or 3. Then f is not irreducible iff $\exists c \in F$ such that $f(c) = 0$.*

Proof. We have a non-trivial factorization of f iff $f = gh$, for $g, h \in F[x]$, where either g or h has degree 1. This is true since $\deg f = \deg g + \deg h$ and $2 = 1+1$ or $3 = 2+1 = 1+2$ are the only possibilities for partitions of 2 or 3 into sums of 2 integers. It follows that we can take one of the factors, say $g(x)$, to be monic and linear; i.e., $g(x) = x - c$ for some $c \in F$. ■

Example. The preceding test fails for $\mathbb{Z}_6[x]$ since, for example, $f(x) = (2x+1)^2$ has no roots in $\mathbb{Z}_6$.

Proposition 33 ***(A Field with p^n elements, p =prime).*** *If p=prime and $f(x)$ is an irreducible polynomial in $\mathbb{Z}_p[x]$, then the quotient $\mathbb{F}_{p^n} = \mathbb{Z}_p[x]/\langle f(x)\rangle$ is a field with p^n elements, where $n = \deg f$.*

Proof. Since f is irreducible, the ideal $\langle f\rangle$ is maximal and thus $F[x]/\langle f(x)\rangle$ is a field by Corollary 28. To see that this field has p^n elements, we just need to see that the elements of $\mathbb{Z}_p[x]/\langle f(x)\rangle$ are represented by the remainders of $h \in \mathbb{Z}_p[x]$ upon division by f. These remainders have the form $r(x) = a_{n-1}x^{n-1} + \cdots + a_1x + a_0$, where $a_j \in \mathbb{Z}_p$. Moreover, the cosets in $\mathbb{Z}_p[x]/\langle f(x)\rangle$ of 2 distinct remainders cannot be the same. Otherwise f would divide the difference of the remainders. But $\deg f$ is greater than the degree of the difference of 2 remainders. Contradiction. How many such remainders are there? There are p possibilities for each a_j. Thus there are p^n remainders and thus p^n elements of $\mathbb{Z}_p[x]/\langle f(x)\rangle$. ■

Example. A field with 9 elements: $\mathbb{Z}_3[x]/\langle x^2+1\rangle = \mathbb{F}_9$.

If we view $\mathbb{Z}_3$ as an analog of the real numbers, this field may be viewed a a finite analog of the complex numbers. From the preceding Proposition, we know that $\mathbb{Z}_3[x]/\langle x^2+1\rangle$ is a field with 9 elements, once we know that x^2+1 is irreducible in $\mathbb{Z}_3[x]$. From Proposition 32, we know that x^2+1 is irreducible iff it has no roots in $\mathbb{Z}_3$. Consider $f(x) = x^2+1$ and plug in the elements of $\mathbb{Z}_3$. You get $f(0) = 1$, $f(1) = f(-1) = 2$. Thus x^2+1 is irreducible in $\mathbb{Z}_3[x]$.

The field $\mathbb{Z}_3[x]/\langle x^2+1\rangle$ is isomorphic to $\mathbb{Z}_3[i] = \{a+bi \mid a,b \in \mathbb{Z}_3\}$ =the smallest field containing $\mathbb{Z}_3$ and i such that $I^2 = -1$. We showed in the preceding paragraph that no element of $\mathbb{Z}_3$ satisfies the equation for i. To prove this, you can define an onto map $T : \mathbb{Z}_3[x] \to \mathbb{Z}_3[i]$ by $T(f(x)) = f(i)$. Then $\ker T = \langle x^2+1\rangle$ since if $g \in \ker T$, we have $g(x) = q(x)(x^2+1) + r(x)$, where $\deg r(x) < 2$ or $r = 0$. So $0 = g(i) = r(i)$ says $r = 0$, and then $g \in \langle x^2+1\rangle$. Conversely any element of $\langle x^2+1\rangle$ must be in $\ker T$. Thus, the 1st isomorphism theorem says $\mathbb{Z}_3[x]/\langle x^2+1\rangle$ is isomorphic to $\mathbb{Z}_3[i]$.

Next we consider $\mathbb{Z}_p$, for p=prime to be an analog of the real numbers $\mathbb{R}$ and we ask, **is there an analog of the quadratic formula?** So consider the quadratic equation $ar^2 + br + c = 0$ for $a,b,c \in \mathbb{Z}_p$ and $a \neq 0$. Now we do the $\mathbb{Z}_p$ analog of completing the square, **as long as $p \neq 2$**. We can divide by a since $\mathbb{Z}_p$ is a field and obtain

$$r^2 + \frac{b}{a}r = \frac{-c}{a}.$$

Then for $p \neq 2$ we can add $\left(\frac{b}{2a}\right)^2$ to both sides and get

$$r^2 + \frac{b}{a}r + \left(\frac{b}{2a}\right)^2 = \frac{-c}{a} + \left(\frac{b}{2a}\right)^2.$$

Now the left hand side is a square and we have

$$\left(r + \frac{b}{2a}\right)^2 = \frac{-c}{a} + \left(\frac{b}{2a}\right)^2 = \frac{-4ac + b^2}{4a^2}.$$

Define the **discriminant** $D = b^2 - 4ac$ and obtain

$$\left(r + \frac{b}{2a}\right)^2 = \frac{D}{(2a)^2}.$$

So we take square roots of both sides and note that we may have to go to a larger field than $\mathbb{Z}_p$ to find $\sqrt{D}$. This gives:

$$r = \frac{-b \pm \sqrt{D}}{2a}.$$

This is the "same" quadratic formula we saw in junior high or high school (now maybe middle school). We have 2 cases:

Case 1 for Finite Fields. If $\sqrt{D} \in \mathbb{Z}_p$ then $r \in \mathbb{Z}_p$.

Case 2 for Finite Fields. If $\sqrt{D} \notin \mathbb{Z}_p$, we can view r as an element of our analog of the complex numbers $\mathbb{Z}_p[\sqrt{D}] \cong \mathbb{Z}_p[x]/\langle x^2 - D\rangle = \mathbb{F}_{p^2}$.

For the real numbers we also had 2 cases:

Case 1. $D = b^2 - 4ac \geq 0 \Longrightarrow$roots real.

Case 2. $D = b^2 - 4ac < 0 \Longrightarrow$roots complex and not real.

You may be wondering about the case $p = 2$. When $p = 2$, the quadratic formula does not make sense as $1/2$ makes no sense in $\mathbb{Z}_2$.

In the next section we recall a bit of linear algebra, just to make sure that you believe it works for any fields as well as for the field of real numbers.

9 Field of Quotients

Field of quotients or fractions or quotient field generalizes the idea from elementary school creating the rational numbers $\mathbb{Q}$ from the integers $\mathbb{Z}$. It also generalizes the construction of the field of rational functions $F(x)$ from the ring of polynomials $F[x]$ over a field F. Recall that we need to identify the fractions: $\frac{1}{3} = \frac{2}{6} = \frac{-2}{-6}$ or $\frac{3}{4} = \frac{75}{100}$. You need to recall how to add and multiply them too: $\frac{2}{3} + \frac{5}{7} = \frac{14+15}{21} = \frac{29}{21}$, $\frac{2}{3} \cdot \frac{5}{7} = \frac{10}{21}$.

Once you remember this, you should be able to generalize the idea to any integral domain. Thus you would make the following definition, state the next theorem, and do the ensuing exercise.

Definition 34 *Suppose that D is an integral domain. Then we can construct the **field of quotients** F by 1st creating a set S whose elements are the symbols $\frac{a}{b}$, where $a, b \in D$ and $b \neq 0$. An equivalence relation on S is given by saying $\frac{a}{b} \sim \frac{c}{d}$ iff $ad = bc$. Then define F to be the set of equivalence classes $\left[\frac{a}{b}\right]$ of S. Addition is defined by saying $\left[\frac{a}{b}\right] + \left[\frac{c}{d}\right] = \left[\frac{ad+bc}{bd}\right]$, and multiplication is defined by saying $\left[\frac{a}{b}\right]\left[\frac{c}{d}\right] = \left[\frac{ac}{bd}\right]$, for $a, b, c, d \in D$. In these definitions of addition and multiplication, we always assume $bd \neq 0$.*

Theorem 35 *The preceding definition creates a field F which contains a subring isomorphic to D.*

Exercise. Prove the preceding theorem by checking that $\sim$ is indeed an equivalence relation, then that the definitions of addition and multiplication of equivalence classes is independent of representative. Finally check that F satisfies field axioms.

The preceding exercise is worked out in many of the references.

The earliest known use of fractions (according to the all wise internet) goes back to 2800 B.C. in India (the Indus valley).

10 Vector Spaces

We recall only the basics of linear algebra. We will make much of this subject an exercise. You can find solutions in Dornhoff and Hohn, *Applied Modern Algebra*, for example. Or you could look at whatever book you used for this part of your calculus course and ask what remains true if we replace $\mathbb{R}$ by $\mathbb{Z}_p$ or some other field F. Most of the earlier chapters on Gaussian elimination, dimension, determinants work as before. So, for example, you might take the book *Linear Algebra and its Applications* by G. Strang (or the book with the same title by D. Lay) and convince yourself that all the results of the early chapters work for arbitrary fields.

The favorite calculations from linear algebra involve Gaussian elimination. It turns out that this is not due to Gauss at all but appears in a Chinese book parts of which were written as early as 150 B.C. **Gaussian elimination** allows one to put a matrix $A \in F^{m \times n}$ into echelon form using elementary row operations.

The **elementary row operations over the field F** are:
1) permute row i and row j;
2) multiply row i by a non-0 element of F times row i;
3) replace row i by row i plus an element of F times row j.

Row Echelon form means that
1) Rows with at least 1 non-0 element are above the rows of all 0's.
The 1st non-0 element of a non-0 row is called a **pivot**.
Below each pivot is a column of 0's;
2) Each pivot is to the right of the pivot in the row above.

You can put the matrix in **row-reduced echelon form** by normalizing all pivots to be 1 and then by putting 0's above all pivots.

Example. Suppose the field is $\mathbb{Z}_3$ and the matrix is

$$\begin{pmatrix} 2 & 1 & 0 & 2 \\ 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

We can replace (row 2) by (row2-2*row1) and do the same for (row 3) to get

$$\begin{pmatrix} 2 & 1 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 2 \end{pmatrix}.$$

Finally replace the (row 3) by (row3-row2) to get

$$\begin{pmatrix} 2 & 1 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

This matrix is in row-echelon form.

The row reduced-echelon form of the matrix is

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Exercise. Over the field $\mathbb{Z}_2$ put the following matrix (which we will see again in the section on error-correcting codes)

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

into row-echelon form and row-reduced echelon form.

Since the elementary row operations on the matrix of a homogeneous equation do not change the solution set, we get the theorem saying that a homogeneous system of m linear equations in n unknowns over a field F has a non-trivial solution if $m < n$. This can be seen by looking at the matrix version of the system of equations $Ax = 0$, with $A \in F^{m\times n}, x \in F^n$. One can do Gaussian elimination on the matrix A; i.e., elementary row operations over the field F are used to put A into a nicer (echelon) form. To help understand this, note that each elementary row operation on A corresponds to finding an $m \times m$ non-singular matrix U with entries in the field F and replacing A by UA. For example, the 1st operation we did to the matrix in the example above was to replace (row 2) by (row 2 -2row1). This is achieved by multiplying the matrices below:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 0 & 2 \\ 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Example. Suppose the field is $\mathbb{Z}_3$ and the matrix is that of the previous example

$$\begin{pmatrix} 2 & 1 & 0 & 2 \\ 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Then the corresponding system of equations is

$$\begin{cases} 2x_1 + x_2 + 2x_4 = 0 \\ x_1 + 2x_4 = 0 \\ x_1 = 0. \end{cases}$$

An equivalent system is that corresponding to the row reduced-echelon form matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

which is

$$\begin{cases} x_1 = 0 \\ x_2 = 0 \\ x_4 = 0. \end{cases}$$

The result is that $x_1 = x_2 = x_4 = 0$ and the other coordinate x_3 is arbitrary (i.e., free to be whatever it wants to be in $\mathbb{Z}_3$.)

Exercise. Over the field $\mathbb{Z}_2$ write down the homogeneous system of equations corresponding to the following matrix

$$G = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

From an earlier exercise, the row echelon form of G is $G' = (I_4 \ A)$. If $H = \begin{pmatrix} -A \\ I_3 \end{pmatrix}$, solve $xH = 0$, for $x \in \mathbb{Z}_2^7$. How does the set of such vectors x compare with the set of vectors $y = uG$, for $u \in \mathbb{Z}_2^4$?

Question. Does Gaussian elimination work over Euclidean domains D like $\mathbb{Z}$ and $F[x]$? The answer is that it does work well. In fact, you could even allow D to be a principal ideal domain.

Elementary Row Operations over a Euclidean Domain D

1) permute row i and row j;
2) replace row i by a unit in D^* times row i;
3) replace row i by row i plus an element of F times row j.

This allows one, for example, to put a square matrix of integers into the **Smith normal form,** meaning a diagonal matrix such that if d_i is the ith diagonal entry then d_i divides d_{i+1} for all i. The diagonal entries are called **elementary divisors** and are unique up to multiplication by units in D. This result in turn can be used to prove the fundamental theorem of finitely generated abelian groups stated below. It is also useful in computations of algebraic number theory and algebraic topology.

Theorem 36 ***(The Fundamental Theorem of Abelian Groups).*** *Any finitely generated abelian group G is isomorphic to direct sum $\mathbb{Z}_{a_1} \oplus \cdots \oplus \mathbb{Z}_{a_n} \oplus \mathbb{Z}^r$.*

Similarly if two $n \times n$ matrices A, B have entries in F, where F is a field, one can use the Smith normal form (over the polynomial ring $F[x]$) of $A - xI$ and $B - xI$ to decide whether A and B are **similar** matrices, meaning that $B = UAU^{-1}$ for some invertible matrix $U \in F^{n \times n}$. The Smith normal form is due to H. J. Smith (1826-1883).

Exercise. Put the following matrix of integers into its Smith normal form using only elementary row operations over $\mathbb{Z}$:

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

Exercise. Put the following 2 matrices in their Smith normal forms using only elementary row operations over $\mathbb{Z}_3[x]$:

$$\begin{pmatrix} x-1 & 1 \\ 0 & x-1 \end{pmatrix}, \begin{pmatrix} x-1 & 0 \\ 1 & x-2 \end{pmatrix}.$$

A Digression on Algebraically Closed Fields.

Some things in the later chapters of linear algebra texts (like eigenvalues) don't work well for fields like $\mathbb{R}$ but instead require the field to be a larger field like $\mathbb{C}$ where all polynomials factor completely into a product of degree 1 polynomials. This is the fundamental theorem of algebra which says that $\mathbb{C}$ is an algebraically closed field. A field F is **"algebraically closed"** means that all polynomials in $F[x]$ factor completely into a product of degree 1 polynomials from $F[x]$. Given 2 algebraic closures of $\mathbb{F}_p$, there is a field isomorphism from one algebraic closure to the other fixing every element of $\mathbb{F}_p$.

The history of proofs of the fundamental theorem of algebra is very interesting. The first proofs (given by d'Alembert in 1746 and C. F. Gauss in 1799) had flaws. Gauss later published 3 correct proofs. Some analysis is usually required and thus we will not prove the theorem here. My favorite proof uses Liouville's theorem from complex analysis. See Birkhoff and MacLane, *A Survey of Modern Algebra* for a topological proof. Another reference is Courant and Robbins, *What is Mathematics?*

Anyway, if we want to stick with finite fields, we have no analog of $\mathbb{C}$. For any finite field F, there will be an irreducible polynomial of degree $n > 1$ having coefficients in F.

Exercise. Prove this last statement. **Hint.** Look at the polynomial $1 + \prod_{a \in F} (x - a)$.

When I am feeling finite, the preceding paragraph makes me very sad. Of course, you can keep adding roots of polynomials to $\mathbb{F}_p$. This may be done in various ways. For example, PlanetMath.org envisions taking $\mathbb{F}_{p^\infty} = \bigcup_{n=1}^{\infty} \mathbb{F}_{p^{n!}}$ to get an algebraically closed field containing $\mathbb{F}_p$. The finite field $\mathbb{F}_{p^n}$ can be constructed as the splitting field of the polynomial $x^{p^n} - x$, as we shall see in Section 11.

Definition 37 *A set V (the **vectors**) is a **vector space** over a field F (the **scalars**) means that V is an abelian group under addition and that there is a function from $F \times V$ into V sending $(\alpha, v) \in F \times V$ to $\alpha \cdot v = av$ (multiplication by scalars) such that $\forall v, w \in V$ and $\forall \alpha, \beta \in F$ we have the following 4 properties.*
1) $\alpha(v + w) = \alpha v + \alpha w$;
2) $(\alpha + \beta)v = \alpha v + \beta v$;
3) $\alpha(\beta v) = (\alpha\beta)v$;
4) $1v = v$.

Example. The plane. $\mathbb{R}^2 = \{(x, y) \mid x, y \in \mathbb{R}\}$ is a vector space over the field $\mathbb{R}$. Here addition is componentwise; i.e.,

$$(x, y) + (u, v) = (x + u, y + v)$$

and multiplication by scalars is also componentwise; i.e.,

$$\alpha(x, y) = (\alpha x, \alpha y),$$

for all $x, y, \alpha \in \mathbb{R}$. If you replace $\mathbb{R}$ by any field F, you make F^2 into a vector space with analogous definitions of addition and multiplication by scalars.

Definition 38 *The **dimension** of a finite dimensional vector space V over the field F is the size of a basis B of V. A **basis** B of the finite dimensional vector space V over the field F means that B is a finite subset of V with 2 properties:*
1) ***B spans V****; i.e. every vector $v \in V$ is a linear combination of elements of B. That is, if $B = \{b_1, ..., b_n\}$,*

$$v = \sum_{i=1}^{n} \alpha_i b_i, \quad \text{for some scalars } \alpha_i \in F.$$

2) ***B is a set of linearly independent vectors****; i.e.,* $\sum_{i=1}^{n} \alpha_i b_i = 0$ *for some scalars $\alpha_i \in F$ implies all $\alpha_i = 0$.*

To show that the idea of dimension makes sense, one must prove that any 2 bases of a vector space have the same number of elements. We leave this as an **exercise. Hint.** See Birkhoff and MacLane, *A Survey of Modern Algebra* or Dornhoff and Hohn, *Applied Modern Algebra*. Or you can imitate the proof in Strang, *Linear Algebra and its Applications*, as follows. If, by contradiction, the sets $A = \{u_1, ..., u_m\}$ and $B = \{v_1, ..., v_n\}$ are 2 bases of the vector space V over the field F, and $m = |A| < n = |B|$, we can write every element of B as an F-linear combination of elements of A; and then obtain m equations in n unknowns expressing the linear dependence of the elements of B. Such a system has a non-trivial solution - a contradiction.

We assume that all vector spaces considered here are finite dimensional.

Example. The plane. $\mathbb{R}^2$ is 2-dimensional with basis $\{(1, 0), (0, 1)\}$. This is the standard basis. Another basis of $\mathbb{R}^2$ is $\{(1, 1), (2, 0)\}$.

Exercise. Prove the preceding statements and then the analog replacing $\mathbb{R}$ by $\mathbb{Z}_3$.

Exercise. Show that, for a matrix in row-echelon form, the set of non-0 rows is a linearly independent set.

Notation. We will view elements of F^n as row vectors (mostly). This means we need to write vM if $M \in F^{n \times m}$. The function $v \to vM$ composes badly with another matrix function since the function part, M, is on the right rather than the left. That's why I would really prefer to write column vectors. But they take up a lot of space on a page.

Definition 39 *A **subspace** W of a vector space V over the field F is a non-empty subset $W \subset V$ which is a vector space under the same operations as those of V.*

Example 1. The plane. $\mathbb{R}^2$ has as subspace $W = \{(x,0)|x \in \mathbb{R}\}$, the real line. Similarly $\mathbb{Z}_3^2$ has as a subspace $W = \{(x,0) \mid x \in \mathbb{Z}_3\}$.
Example 2. A subfield *F* of a field *E* is a vector subspace of E, considering E and F as vector spaces over any subfield of F.

We define the **span** of a set S of vectors in the vector space V over the field F to be

$$Span(S) = \{\text{finite } F - \text{linear combinations of vectors from } S\}.$$

Exercise. Prove that $Span(S)$ is indeed a vector subspace of V.
Exercise. Prove that elementary row operations do not change the span of the set of rows of a matrix.

The **(row) rank** of a matrix $A \in F^{m \times n}$ is defined to be the dimension of the span of the set of row vectors of A. Thanks to the following exercise, we just say "rank" rather than the row rank of a matrix.
Exercise. Prove that the row rank of a matrix is the same as the column rank (which is the dimension of the span of the set of columns of the matrix).

Definition 40 *If V and W are both vector spaces over the field F, then a mapping $T : V \to W$ is a **vector space isomorphism** over F iff T is 1-1, onto, $T(v+w) = T(v) + T(w)$, and $T(\alpha v) = \alpha T(w)$, for all $v, w \in V$ and all $\alpha \in F$. Then we write $V \cong W$ and say V is **isomorphic** to W. If we drop the 1-1 and onto requirement, we could call T a vector space homomorphism, but instead we will call such a function a **linear mapping** (or linear transformation, or linear operator).*

Example 1. If $B = \{b_1, ..., b_m\}$ is a basis of the vector space V over the field F, then $V \cong F^m$. The mapping M_B is defined by writing $v \in V$ in the form $v = \sum_{i=1}^{m} \alpha_i b_i$ and then setting

$$M_B(v) = (\alpha_1, ..., \alpha_m). \tag{5}$$

We leave it as an **exercise** to check that M_B is indeed a vector space isomorphism.
Example 2. Consider the linear mapping $T : \mathbb{Z}_2^4 \to \mathbb{Z}_2^7$ defined by mapping a row vector $v \in \mathbb{Z}_2^4$ to the row vector $T(v) = vG$, where

$$G = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

Show that the image space $T\left(\mathbb{Z}_2^4\right)$ is a vector subspace of $\mathbb{Z}_2^7$. Find a basis for $T\left(\mathbb{Z}_2^4\right)$. What is the dimension of the image space?

2 Methods to find a basis of a finite dimensional vector space *V* over a field *F*?

Method 1. If you have a spanning set S of vectors in V you keep deleting any vectors that can be written as a (finite) linear combination of other vectors from S.
Method 2. Start with one non-0 vector or any non-empty set of linearly independent vectors in V and keep adding vectors from V that are not in the span of the vectors you already have.

Exercise. Prove that the 2 methods to find a basis of V actually work.
Exercise. Prove that if V is a vector space over the field F, then a linear mapping $T : V \to V$ is 1-1 iff it is onto.

Given that V and W are both (finite dimensional) vector spaces over the field F, then the **matrix of a linear mapping** $T : V \to W$ with respect to the (ordered) bases $B = \{b_1, ..., b_m\}$ of V and $C = \{c_1, ..., c_n\}$ of W is defined to be the $n \times m$ array of scalars $\mu_{ij} \in F$ defined by:

$$Mat_{B,C}(T) = (\mu_{ij})_{\substack{1 \le i \le n \\ 1 \le j \le m}}, \quad \text{where } T(b_j) = \sum_{i=1}^{n} \mu_{ij} c_i, \quad \text{for } j = 1, ..., m.$$

We have defined the matrix of a linear transformation this way in order that the composition of linear transformations corresponds to product of matrices.

Exercise. Suppose V, W are vector spaces over the field F and $T : V \to W$, is a linear map. If B, C are (ordered) bases of V, W, show that, using the definition of M_B in formula (5) from Example 1 above, we have

$$^{T}M_C\left(T(v)\right) = Mat(T)_{B,C} \ \ ^{T}M_B(v),$$

where if $A = (\alpha_{ij})_{\substack{1\leq i\leq n \\ 1\leq j\leq m}}$, the **transpose** of A denoted ^{T}A is the matrix $(\alpha_{ji})_{\substack{1\leq j\leq m \\ 1\leq i\leq n}}$, where the rows and columns are interchanged.

Hint. By the linearity of T, we have

$$T\left(\sum_{j=1}^{m}\alpha_j b_j\right) = \sum_{j=1}^{m}\alpha_j T\left(b_j\right) = \sum_{j=1}^{m}\alpha_j \sum_{i=1}^{n}\mu_{ij}c_i.$$

Interchange the sums over i and j are you will have done the Exercise.

Exercise. Suppose V, W, U are vector spaces over the field F and $T : V \to W$, $S : W \to U$ are both linear maps. Show that the composition $S \circ T : V \to U$ is also a linear map. Then show that if B, C, D are (ordered) bases of V, W, U, respectively, then

$$Mat_{B,D}(S \circ T) = Mat_{C,D}\left(S\right) Mat_{B,C}\left(T\right), \tag{6}$$

where the product on the right is the usual matrix multiplication.

Given that V and W are both vector spaces over the field F, the **rank of a linear map** $T : V \to W$ is the dimension of the image $L(V) = \{Lv | v \in V\}$.

Exercise. a) Show that, assuming V and W are both vector spaces over the field F and $T : V \to W$ is linear, then the image $L(V)$ is indeed a vector subspace of W.

b) Show that the rank of a linear transformation L is the same as the rank of a matrix of L using bases B of V and C of W.

Exercise. Consider the field $\mathbb{Z}_3[i]$, where $i^2 + 1 = 0$.

a) Show that $\mathbb{Z}_3[i]$ is a vector space over the field $\mathbb{Z}_3$.

b) Define the map $F : \mathbb{Z}_3[i] \to \mathbb{Z}_3[i]$ by $F(z) = z^3$. Show that F is linear. Then find a matrix of F using the basis $\{1, i\}$ for $\mathbb{Z}_3[i]$.

Theorem 41 *Given that V and W are both (finite dimensional) vector spaces over the field F, and $T : V \to W$ is a linear map,*

$$\dim \ker T + \dim T(V) = \dim V.$$

Proof. Take a basis $B = \{b_1, ..., b_m\}$ of $\ker T$ and extend it to a basis $C = \{b_1, ..., b_m, b_{m+1}, ..., b_n\}$. Then we claim that $\{T\left(b_{m+1}\right), ..., T\left(b_n\right)\}$ is a basis for $T(V)$. ■

Exercise. Fill in the details in the preceding proof.

Exercise. Suppose that $B = \{b_1, ..., b_n\}$ and $C = \{c_1, ..., c_n\}$ are 2 (ordered) bases of the vector space V over the field F. Let $1_V(x) = x$, $\forall x \in V$ be the identity map (which is certainly linear).

a) Show that if $M = Mat_{B,C}(1_V)$, then M is invertible.

b) Show that for any linear map $L : V \to V$, we have

$$Mat_{B,C}(1_V)Mat_{B,B}(L) = Mat_{C,C}(L)Mat_{B,C}(1_V).$$

Hint on b). Use formula (6) to see that both sides are $Mat_{B,C}(T)$.

Definition 42 *If A and B are $n \times n$ matrices with entries in the field F, we say that A and B are **similar** iff there is an invertible $n \times n$ matrix U such that $B = U^{-1}AU$.*

Exercise. a) Show that similarity is an equivalence relation on $F^{n\times n}$.

b) Show that two different $n \times n$ matrices A and B over the field F are similar iff they are the matrices of the same linear transformation $T : F^n \to F^n$ with respect to two different bases.

The Smith normal form allows one to obtain canonical forms of matrices (such as the Jordan form) so that any matrix will be similar to only one matrix of a given canonical form. See Dornhoff and Hohn, *Applied Modern Algebra*, for more details. This can be useful despite the reluctance of some applied books to consider the Jordan form of a matrix.

For example if one needs to do Fourier analysis on the general linear group $G = GL(n, \mathbb{Z}_p)$ of invertible 2×2 matrices over the field $\mathbb{Z}_p$, p=prime, one must know the conjugacy classes $\{g\} = \left\{x^{-1}gx \mid x \in G\right\}$. That is, one needs to know the similarity classes.

These **similarity (conjugacy) classes in** $GL(2, \mathbb{Z}_p)$ are

$$\begin{array}{ll} \text{central} & \left\{\begin{pmatrix} r & 0 \\ 0 & r \end{pmatrix}\right\} \\ \text{parabolic} & \left\{\begin{pmatrix} r & 1 \\ 0 & r \end{pmatrix}\right\} \\ \text{hyperbolic} & \left\{\begin{pmatrix} r & 0 \\ 0 & s \end{pmatrix}\right\}, \text{ where } rs \neq 0 \\ \text{elliptic} & \left\{\begin{pmatrix} r & s\delta \\ s & r \end{pmatrix}\right\}, \text{ where } \delta \text{ is not a square in } \mathbb{Z}_p. \end{array}$$

See Terras, *Fourier Analysis on Finite Groups and Applications, p. 366,* for more information.

Recall that the characteristic p of a finite field F is a prime number p which is the order of 1 in the additive group of F. It follows that if F is a finite field of characteristic p, then F contains $\mathbb{Z}_p$ as a subfield, as we prove next.

Proposition 43 *A finite field F of characteristic p (necessarily a prime) is a vector space over $\mathbb{Z}_p$.*

Proof. Look at the additive subgroup H of F which is generated by 1. Then $T : \mathbb{Z} \to H \subset F$ defined by $T(n) = n \cdot 1$ is a ring homomorphism mapping $\mathbb{Z}$ onto H. By the definition of characteristic, we know that $\ker T = p\mathbb{Z}$. By the 1st isomorphism theorem, we know that H is isomorphic to $\mathbb{Z}/p\mathbb{Z} = \mathbb{Z}_p$. This finishes the proof. ■

Corollary 44 *A finite field F of characteristic p is a vector space over $\mathbb{Z}_p$ which is necessarily finite dimensional. If the dimension of F over $\mathbb{Z}_p$ is n, then $F \cong \mathbb{Z}_p^n$. This implies that F has p^n elements.*

Exercise. Show that there is no integral domain with exactly 6 elements.

Notation: We write $\mathbb{F}_{p^n}$ for the field with p^n elements since we will be able to show that there is only one such field up to isomorphism.

Example. The Quaternions. Consider a 4-dimensional vector space $\mathbb{H}$ over $\mathbb{R}$ with basis $1, i, j, k$. We define multiplication by first defining how to multiply the basis vectors as in the multiplication table for the quaternion group in Part I. That is, $i^2 = j^2 = k^2 = ijk = -1$. Then assume that the multiplication satisfies the usual associative and distributive laws plus $(\alpha v)\cdot w = \alpha(v \cdot w) = v \cdot (\alpha w)$, for all $\alpha \in \mathbb{R}$ and $v, w \in \mathbb{H}$. This gives a non-commutative ring. It turns out that you can divide by non-0 elements. That is because we have an analog of complex conjugate: $\overline{\alpha_1 + \alpha_2 i + \alpha_3 j + \alpha_4 k} = \alpha_1 - \alpha_2 i - \alpha_3 j - \alpha_4 k$. Then if $v = \alpha_1 + \alpha_2 i + \alpha_3 j + \alpha_4 k$, for $\alpha_n \in \mathbb{R}$, we have $v\overline{v} = \alpha_1^2 + \alpha_2^2 + \alpha_3^2 + \alpha_4^2$. This means that when $v \neq 0$, we have $v^{-1} = \frac{\overline{v}}{v\overline{v}} \in \mathbb{H}$.

We told some of the story of Hamilton's discovery of the quaternions in Section 19 of Part I where we introduced the quaternion group. The quaternions have proved useful in physics and number theory. The construction has been generalized, replacing $\mathbb{R}$ by other fields. Finite quaternions turn out not to be so interesting as they are full matrix algebras.

Exercise. Show that in the quaternions $\mathbb{H}$, we have $\overline{x \cdot y} = \overline{y} \cdot \overline{x}$

Exercise. Suppose we multiply two quaternions

$$(\alpha_1 + \alpha_2 i + \alpha_3 j + \alpha_4 k)(\beta_1 + \beta_2 i + \beta_3 j + \beta_4 k) = \gamma_1 + \gamma_2 i + \gamma_3 j + \gamma_4 k. \tag{7}$$

Show that $\gamma_1 = \alpha_1\beta_1 - \alpha_2\beta_2 - \alpha_3\beta_3 - \alpha_4\beta_4$. Obtain similar formulas for the rest of the γ_r, $r = 1, 2, 3$.

Exercise. Use what we know about quaternions to prove Lagrange's identity which says that if the relationship of the γs to the αs and βs is as in formula (7), then $\left(\alpha_1^2 + \alpha_2^2 + \alpha_3^2 + \alpha_4^2\right)\left(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2\right) = \gamma_1^2 + \gamma_2^2 + \gamma_3^2 + \gamma_4^2$.

Determinants

Hopefully everyone knows that $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$ and the analogous formula in 3 dimensions which has 6 terms. What happens in n dimensions? There are $n!$ terms - one term for every element of the symmetric group. In short, the formula is:

$$\det \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = \sum_{\sigma \in S_n} sgn(\sigma) a_{1\sigma(1)} a_{2\sigma(2)} \cdots a_{n\sigma(n)}. \tag{8}$$

This formula is not so good for evaluating determinants. For that one needs to know the properties. Properties 2) and 3) combine to say that the determinant is a linear function of each row holding the rest of the rows fixed. Property 4) says the determinant is an alternating multilinear function of its rows.

Proposition 45 ***(Properties of Determinants).***

1) The determinant of an upper (or lower) triangular matrix is the product of the entries on the diagonal.

2) If all entries in 1 row of a matrix are multiplied by the scalar c then the determinant is multiplied by c.

3) If we write the ith row of a matrix A as $v + w$ then the determinant of A is the sum of the determinant of matrix C and matrix D, where matrix C is the same as A except that the ith row is vector v and matrix D is the same as A except that the ith row is vector w.

4) If 2 rows of a matrix are interchanged the determinant is multiplied by -1.

5) $\det(A) = \det({}^T A)$.

6) If 2 rows of a matrix are equal, then the determinant is 0.

7) A matrix is invertible (or nonsingular) iff its determinant is not 0.

8) $\det(AB) = \det(A)\det(B)$.

For the proofs, see Herstein, *Topic in Algebra*, Chapter 6. You could really do them as exercises. The modern way of doing these things is called "Exterior Algebra" or alternating multilinear algebra. We do not have time to cover this subject but if you dislike messy formulas with subscripts like formula (8), then this is for you. It is important in several variables integral calculus for helping to understand why the Jacobean determinant appears in the change of variables formula for multiple integrals and in helping to understand Stokes' theorem. See the end of Serge Lang, *Undergraduate Analysis,* or Courant and John, *Calculus*, Vol. II, for an introduction.

11 Subfields and Field Extensions of Finite Fields

Next suppose that E is a finite field of characteristic p with subfield F. Then we say that E is a **field extension** of F. Both E and F are extensions of $\mathbb{Z}_p$. We can view E as a vector space over F and then $E \cong F^r$, where r is the vector space dimension of E over F. If the dimension of F over $\mathbb{Z}_p$ is n, then $|F| = p^n$ and $|E| = p^{nr}$.

Definition 46 *The* ***degree*** *d of an extension $F \subset E$ of finite fields is the dimension of E as a vector space over F. The notation is $d = [E : F]$.*

Exercise. Show that if $E = F[x]/\langle f(x)\rangle \cong F[\theta]$, where $f(\theta) = 0$ and f is an irreducible polynomial in $F[x]$, then $[E : F] = \deg f$.

Proposition 47 *We have $\mathbb{F}_{p^k} \subset \mathbb{F}_{p^n} \iff k$ divides n.*

Proof. $\Longrightarrow$ $\mathbb{F}_{p^k} \subset \mathbb{F}_{p^n}$ says $\left(p^k\right)^r = p^n$ where r is the dimension of $\mathbb{F}_{p^n}$ as a vector space over $\mathbb{F}_{p^k}$. If follows that $n = kr$ and thus k divides n.

$\Longleftarrow$ We postpone this proof until we have proved that $\mathbb{F}_{p^n}$ is the splitting field of $x^{p^n} - x$, meaning the field where this polynomial factors completely into degree 1 factors. ■

The preceding proposition quickly gives the following corollary.

Corollary 48 *Subfields F of $\mathbb{F}_{p^n}$ are the fields $\mathbb{F}_{p^k}$ such that k divides n.*

Example. We compute $[\mathbb{F}_{5^{21}} : \mathbb{F}_{5^3}] = 7$ since $21 = 3 \cdot 7$.

In Figure 7 we draw the poset diagram for the subfields of $\mathbb{F}_{2^{24}}$. It is the same as the poset diagram for the divisors of 24.

Definition 49 *The* ***splitting field of a polynomial*** *$f(x) \in F[x]$ over the field F is the smallest extension field E of F such that f factors completely into linear factors from $F[x]$; i.e.,*

$$f(x) = c\prod_{i=1}^{n}(x - a_i), \quad \text{for} \quad a_i, c \in E.$$

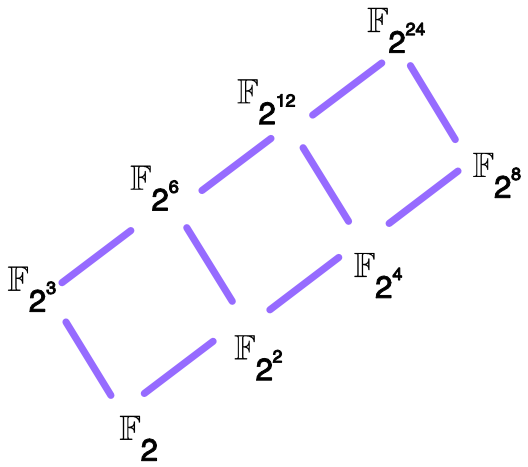

Figure 7: The poset of subfields of $\mathbb{F}_{2^{24}}$

Example 1. **The splitting field of x^2+1 over $\mathbb{R}$ is $\mathbb{C}$, the complex numbers.**
Example 2. **The splitting field of x^2-2** over $\mathbb{F}_5$ is

$$\mathbb{F}_5\left[\sqrt{2}\right] = \left\{ a + b\sqrt{2} \middle| a, b \in \mathbb{F}_5 \right\} \cong \mathbb{F}_5[x] / \langle x^2 - 2 \rangle .$$

Exercise. Find the splitting field E of the polynomial $x^3 + x + 1$ over $\mathbb{F}_2$. What is the degree $[E : \mathbb{F}_2]$?

Theorem 50 *Any polynomial $f(x) \in F[x]$, where F is a field, has a splitting field E and this splitting field is unique up to field isomorphism fixing elements of F.*

Proof. Existence of E. (Induction on $\deg f$). If $\deg f = 1$, f is already in the desired form $c(x - a)$. We know that we can construct a field E_1 containing F and a root θ of an irreducible factor $g(x)$ of f; namely, $E_1 = F[x]/\langle g(x) \rangle$. So we can factor $f(x) = (x - \theta)h(x)$ with $h \in E_1[x]$. By induction on $\deg f$ we may assume that $h(x)$ is completely factored into linear factors in an extension field E of E_1.

Uniqueness of E up to isomorphism. See Gallian, *Contemporary Abstract Algebra.* ■

Moral. Even though the splitting field of a polynomial $f(x)$ over F is only unique up to isomorphism, we will still say **"the"** splitting field.

If F is a field with p^n elements, then F must be the splitting field of $x^{p^n} - x$ over $\mathbb{F}_p = \mathbb{Z}_p$. Why? By Lagrange's theorem from Section 16 of Part I, any non-0 element of a field F with p^n elements is a root of the polynomial $x^{p^n-1} - 1$, since the order of the multiplicative group F^* is $p^n - 1$. So the elements of F are roots of $x\left(x^{p^n-1} - 1\right) = x^{p^n} - x$. Moreover the polynomial $x^{p^n} - x$ has at most p^n distinct roots in F, since by a corollary of the division algorithm it has p^n roots counting multiplicity. Therefore this polynomial has exactly p^n roots in F and F is the splitting field of $x^{p^n} - x$.

Next we want to know whether a general polynomial in $F[x]$ has multiple roots. For this, one needs to take derivatives. We don't want to talk about limits since we usually thinking that F is a finite field, not the real numbers. So we define the formal derivative of a polynomial by the formula that was proved from the limit definition in calculus.

Definition 51 *Suppose that F is any field Then the **formal derivative** of $f(x) = a_r x^r + a_{r-1}x^{r-1} + \cdots + a_1 x + a_0$, with $a_j \in F$ is defined by $f'(x) = r a_r x^{r-1} + (r-1)a_{r-1}x^{r-2} + \cdots + a_1$.*

Example. Show that the formal derivative has the following familiar properties of derivatives, for any $f, g \in F[x]$.
a) $(f + g)' = f' + g'$;
b) $(fg)' = f'g + fg'$;
c) $(f(x)^n)' = n\left(f(x)^{n-1}\right) f'(x)$.

Lemma 52 *A polynomial $f \in F[x]$ does not have a multiple root in an extension field E of F iff $\gcd(f, f') = 1$. Here f' is the formal derivative of f.*

Proof. $\Longleftarrow$ Suppose that $f(x) = (x-a)^2 g(x)$, with $g(x) \in E[x]$ and $a \in E$. Then by the usual properties of derivatives, we have $f'(x) = 2(x-a)g(x) + (x-a)^2 g'(x)$. It follows that $(x-a)$ divides $\gcd(f, f')$.

$\Longrightarrow$ Suppose $\deg \gcd(f, f') \geq 1$. Then $\gcd(f, f')$ is divisible by $x-a$ for some $a \in E$. This means that a is a root of f and f'. So $f(x) = (x-a)h(x)$ for some $h \in E[x]$. But then $f'(x) = h(x) + (x-a)h'(x)$ and $0 = f'(a) = h(a)$. This means $h(x) = (x-a)k(x)$ for some $k \in E[x]$ and thus $(x-a)^2$ divides $f(x)$ and a is a multiple root of f. ■

Theorem 53 *For every prime p and every $n = 1, 2, 3, ...,$ there is a finite field with p^n elements which is isomorphic to the splitting field of $x^{p^n} - x$ over $\mathbb{F}_p$.*

Proof. The splitting field F of $x^{p^n} - x$ over $\mathbb{F}_p$ has p^n distinct roots of $f(x) = x^{p^n} - x$, since $x^{p^n} - x$ cannot have multiple roots using the preceding Lemma and the fact that $\gcd(f, f') = \gcd(x^{p^n} - x, -1) = 1$. If we set $K = \left\{a \in F \mid a^{p^n} = a\right\}$, we can show that K is a subfield of F (**Exercise. Hint.** To see that K is closed under addition, note that $(x+y)^p = x^p + y^p$, for all x, y in a field of characteristic p). We know that $x^{p^n} - x$ splits in K and thus $F = K$. Moreover, then F is a finite field that has p^n elements.

By the moral above, any other field E with p^n elements must be a splitting field of $x^{p^n} - x$ over $\mathbb{F}_p$. This makes E isomorphic to F by an isomorphism which fixes $\mathbb{F}_p$. ■

Examples. We have looked at $\mathbb{F}_8$, $\mathbb{F}_9$, and $\mathbb{F}_{25}$ in Sections 5 and 6. Next let's consider $\mathbb{F}_{16} \cong \mathbb{F}_2[x]/\left\langle x^4 + x + 1\right\rangle \cong \left\{a\theta^3 + b\theta^2 + c\theta + d \;\middle|\; a, b, c, d \in \mathbb{F}_2\right\}$, where $\theta^4 + \theta + 1 = 0$. Here the degree of $\mathbb{F}_{16}$ over $\mathbb{F}_2$ is 4 and $\left\{\theta^3, \theta^2, \theta, 1\right\}$ is vector space basis of $\mathbb{F}_{16}$ as a vector space over $\mathbb{F}_2$.

Exercise. With the notation of the preceding example, show that

a) the polynomial $x^4 + x + 1$ is irreducible in $\mathbb{F}_2[x]$;

b) the polynomial $x^4 + x + 1$ is primitive; i.e., θ generates the multiplicative group $\mathbb{F}_{16}^*$.

Hint. For part b) you need to make a table of powers $\theta^j = a_3\theta^3 + a_2\theta^2 + a_1\theta + a_0$, $a_i \in \mathbb{F}_2$, using the feedback shift register idea from Section 5.

Now we can pursue

the completion of the proof of Proposition 47.

Proof. of the fact that m divides n implies $\mathbb{F}_{p^m}$ is a subfield of $\mathbb{F}_{p^n}$.

If $n = m \cdot r$, then $p^{m \cdot r} - 1 = (p^m - 1)\left(p^{m(r-1)} + p^{m(r-2)} + \cdots + p^m + 1\right)$. This says that $(p^m - 1)$ divides $(p^{m \cdot r} - 1) = (p^n - 1)$. It follows that $\left(x^{p^m - 1} - 1\right)$ divides $\left(x^{p^n - 1} - 1\right)$ in $\mathbb{F}_p[x]$. **Exercise**. Prove this last statement. **Hint.** Use the formula for a geometric progression in the form:

$$\frac{x^{m \cdot k} - 1}{x^m - 1} = (x^m)^{k-1} + (x^m)^{k-2} + \cdots + x^m + 1.$$

Since $\mathbb{F}_{p^r}$ is the splitting field of $x^{p^r} - x$ over $\mathbb{F}_p$, the proof is over. ■

There are still several topics to complete our theory of finite fields. The first is to prove something that we have found to be true in the examples we have considered.

Theorem 54 *The multiplicative group of a finite field is cyclic.*

Proof. Let F be a finite field with q elements. Suppose that a is an element of maximal order in F^* with $n = |a|$. If every element $b \in F$ has order dividing n, then $b^n = 1$ for all $b \in F^*$. We know by a corollary to the division algorithm for polynomials that $x^n - 1$ has at most n roots in F. So $n \geq |F^*| \geq n$ and $F^* = \langle a \rangle$.

So assume there is some $b \in F$ of order k such that k does not divide n. We will derive a contradiction. Then $k = p^t k_0$, $n = p^u n_0$, where p is a prime s.t. p does not divide $k_0 n_0$ and $t > u \geq 0$. Set $a_0 = a^{p^u}$ and $b_0 = b^{k_0}$. Then a_0 has order n_0 and b_0 has order p^t. Since $\gcd(n_0, p^t) = 1$, we know that

$$\langle a_0 \rangle \cap \langle b_0 \rangle = \{1\}.$$

Then $(a_0 b_0)^k = 1 \Longrightarrow a_0^{-k} = b_0^k \in \langle a_0 \rangle \cap \langle b_0 \rangle = \{1\}$, which implies $a_0^k = 1 = b_0^k$. Thus both n_0 and p^t must divide k. It follows that $a_0 b_0$ has order $\geq p^t n_0 > p^u n_0 = n =$order of a. This contradicts the maximality of the order of a which means no such b can exist. The theorem is proved. ■

Exercise. Find all the generators of the multiplicative group of $\mathbb{F}_9 \cong \mathbb{Z}_3[i]$, where $i^2+1=0$.

There are lists of primitive polynomials in the books on finite fields such as R. Lidl and H. Niederreiter, *Introduction to finite fields and their applications*. Here we give a short list containing one primitive polynomial over $\mathbb{F}_2$ for each small degree.

Some Primitive Polynomials over $\mathbb{F}_2$

$$x+1,\ \ x^2+x+1,\ \ x^3+x+1,\ \ x^4+x+1,\ \ x^5+x^2+1,$$
$$x^6+x+1,\ \ x^7+x^3+1,\ \ x^8+x^4+x^3+x^2+1,\ \ x^9+x^4+1,$$
$$x^{10}+x^4+1,\ \ x^{11}+x^2+1$$

Exercise. Show that the fields $\mathbb{F}_2[x]/\left\langle x^3+x+1\right\rangle$ and $\mathbb{F}_2[x]/\left\langle x^3+x^2+1\right\rangle$ and are isomorphic.
Exercise. Show that the mapping of $\mathbb{F}_{p^n}$ onto itself defined by $Tx=x^p$, is a field isomorphism (called the **Frobenius automorphism**) fixing elements of $\mathbb{F}_p$ (viewing $\mathbb{F}_p$ as a subfield of $\mathbb{F}_{p^n}$).
Hint. Use Lemma 17 in Section 3.

The **Galois group** $G(\mathbb{F}_{p^n}/\mathbb{F}_p)$ (which is read as the Galois group of $\mathbb{F}_{p^n}$ over $\mathbb{F}_p$) is defined to be the set of field automorphisms $T:\mathbb{F}_{p^n}\to\mathbb{F}_{p^n}$ such that $T(x)=x$, $\forall x\in\mathbb{F}_p$ (viewed as a subfield of $\mathbb{F}_{p^n}$). It turns out that $G(\mathbb{F}_{p^n}/\mathbb{F}_p)$ is a cyclic group generated by the Frobenius automorphism T of the last exercise. For a proof and a full discussion of Galois theory for finite fields, see L. Dornhoff and F. Hohn, *Applied Modern Algebra.* The fundamental theorem of Galois theory basically says that there is a 1-1 correspondence between intermediate fields E s.t. $\mathbb{F}_p\subset E\subset\mathbb{F}_{p^n}$ and subgroups H of $G=G(\mathbb{F}_{p^n}/\mathbb{F}_p)$. The subgroup H of G corresponding to E is $H=\{T\in G|Tx=x,\forall x\in E\}$. The intermediate field E corresponding to a subgroup H of G is $E=\{x\in\mathbb{F}_{p^n}|Tx=x,\forall T\in H\}$. The fundamental theorem of Galois theory says that the correspondence between intermediate fields and subgroups is 1-1, onto and inclusion reversing. Moreover the degree of the extension $n=[\mathbb{F}_{p^n}:\mathbb{F}_p]$ is equal to the order of the Galois group G.
Example. What does Galois theory say about the extension $\mathbb{F}_{2^8}/\mathbb{F}_2$? What are the intermediate fields between $\mathbb{F}_2$ and $\mathbb{F}_{2^8}$? They are $\mathbb{F}_{2^1},\mathbb{F}_{2^2},\mathbb{F}_{2^4},\mathbb{F}_{2^8}$. So the only non-trivial intermediate fields are $\mathbb{F}_4$ and $\mathbb{F}_{16}$. This implies that, if we believe the fundamental theorem of Galois theory as well as the fact that Galois groups like this are cyclic generated by the Frobenius automorphism F, the Galois group $G=G(\mathbb{F}_{2^8}/\mathbb{F}_2)=\langle F\rangle$ is a cyclic group of order 8 which has only 2 non-trivial proper subgroups $\left\langle F^2\right\rangle$ and $\left\langle F^4\right\rangle$. which correspond to the 2 non-trivial intermediate fields.
The poset diagram for subfields of $\mathbb{F}_{p^n}$ is the same as that for subgroups of $G=G(\mathbb{F}_{p^n}/\mathbb{F}_p)$, except that all inclusion lines are reversed.
R. Dedekind gave the first formal lectures on Galois theory in 1857. It is much easier for finite field extensions than for field extensions of other fields like $\mathbb{Q}$. There are analogs of Galois theory for coverings of Riemann surfaces, topological manifolds and graphs. See A. Terras, *Zeta Functions of Graphs*, for the graph theory version.

12 Random Number Generators

References for this section include:

D. E. Knuth, *The Art of Computer Programming*, Vol. II.
W. H. Press et al, *Numerical Recipes.*
D. Austin, "Random Numbers: Nothing Left to Chance," on the American Math. Society website under feature column.
G. Marsaglia, Random numbers fall mainly in the planes, *Proc. Natl. Acad. Sciences*, 61 (1968), 22-28.
R.P. Brent, Note on Marsaglia XOR Shift random number generators, *J. Stat. Software*, 11 (2004), 1-5.
H. Niederreiter, *Random Number Generation and the Quasi-Monte Carlo Method.*
P. Diaconis, *Group Representations in Probability and Statistics.*
B. Cipra, The best of the 20th Century. Editors name top 10 algorithms, *SIAM News*, Vol. 33, No. 4.
A. Terras, *Fourier Analysis on Finite Groups and Applications.*

There are many uses for sequences of random numbers; e.g., simulations of natural phenomena using "Monte Carlo" methods, systems analysis, software testing. For example, you can approximate $\frac{1}{V}\int\limits_D f(x)dx$, where V is the volume of the

domain D in $\mathbb{R}^n$, by the average value of f on a "random" finite set of points in D. Sadly, the first use of Monte Carlo methods seems to have been in work of Metropolis, Ulam and von Neumann that led to the atomic bomb.

Where do random numbers come from? In the old days there were tables (e.g., that of the RAND corporation from1950). One can also get random sequences of 0's and 1's from tossing a fair coin or from times between clicks of a geiger counter near some radioactive material.

Algebra gives us random numbers (technically, pseudo-random numbers) much more simply, as D.H. Lehmer found in 1949. Let's consider a simple example. Take a random walk through the multiplicative group $\mathbb{Z}_{17}^* = \langle 3(\mathrm{mod}\, 17)\rangle$. In Figure 8 we list the elements in order of the powers $3^j(\mathrm{mod}\, 17)$. The figure is the directed Cayley graph $X(\mathbb{Z}_{17}^*, \{3(\mathrm{mod}\, 17)\})$.This

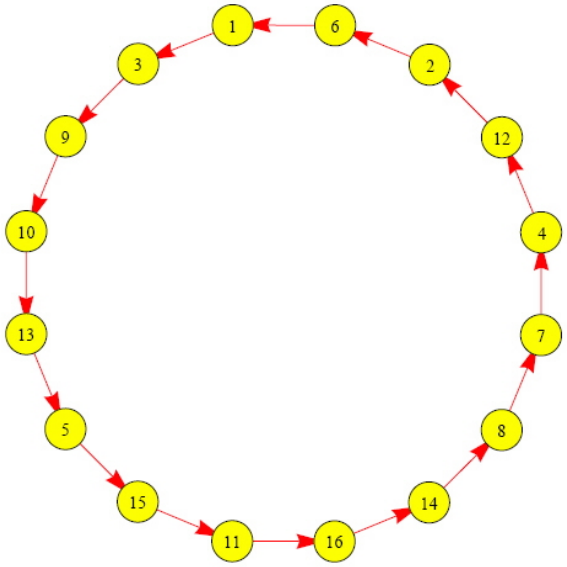

Figure 8: The Cayley graph $X(\mathbb{Z}_{17}^*, \{3(\mathrm{mod}\, 17)\})$.

is very non-random! If instead, we order the vertices according to the usual ordering of the elements of $\mathbb{Z}_{17}^*$, thinking of the group as a set of integers $\{1, 2, 3, ..., 16\}$, then we get the view of the same directed graph that is found in Figure 9.This second view of the same Cayley graph looks way more random. If you imagine doing the same thing with a truly large prime p instead of 17, you would certainly expect to get a random listing of integers by taking the random walk.

The method of generating random numbers created by D.H. Lehmer in 1949 is called the **linear congruential method**. To get a sequence $\{X_n\}_{n\geq 1}$ of (pseudo-) random integers, you fix 4 numbers: m, the modulus; a, the multiplier; c, the increment; and X_0, the starting point. Then you generate the sequence recursively with the formula

$$X_{n+1} = aX_n + c(\mathrm{mod}\, m).$$

If $m = 17$, $a = 3$, $c = 0$, we get the sequence envisaged in Figures 8 and 9. In the old days, popular choices were $m = 2^{31} - 1$ (a Mersenne prime) and $m = 2^{31}$. There was an infamous random number generator RANDU, which was built into the IBM mainframe computers of the 1960s. Randu took $m = 2^{31}, a = 65539, c = 0$. For many years Matlab took $m = 2^{31} - 1, a = 7^5, c = 0$.

What is the meaning of random? That is a question for statisticians who have devised tests to tell whether our lists are reasonably random. This is not a statistics course and thus we refer you to some of the references listed at the beginning of this section. Usually the applied mathematician wants a sequence of random real numbers in the interval $[0, 1]$. To get these from X_n, you just divide by the modulus m which you used to generate them.

In 1968 or so applied mathematicians using random numbers for Monte Carlo methods began to get angry. For it was discovered that if they used the linear congruential random integers X_n to produce vectors in $[0, 1]^n$, for $n > 1$, by writing $v = \frac{1}{m}(X_1, ..., X_n)$, they would have vectors lying in hyperplanes. Thus Marsaglia wrote his paper (see the references above) proving that such vectors will fall into $< (n!m)^{1/n}$ hyperplanes. For example if the modulus $m = 2^{32}$ and $n = 10$, we get < 41 hyperplanes. Of course no one had really tested the random vectors so produced for their uniform distribution in the hypercubes. So perhaps it should not have been a shock.

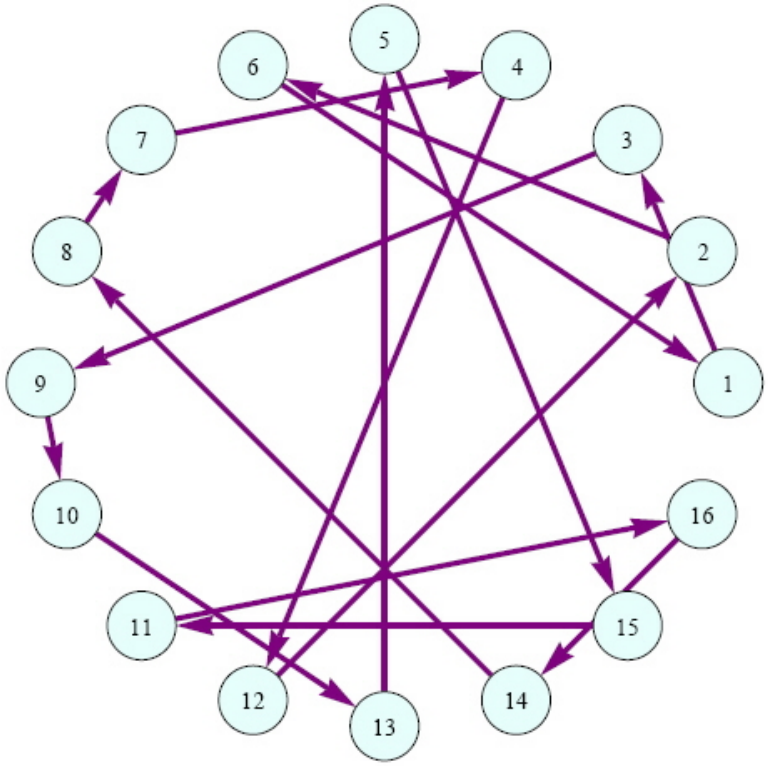

Figure 9: The same graph as in Figure 8 except that now the vertices are given the usual ordering 1,2,3,4,..., 16.

Let's do a simple experiment along these lines. Again we take a fairly small prime, namely $p = 499$ and note that $\mathbb{Z}^*_{499} = \langle 7 (\mathrm{mod}\, 499) \rangle$. We compute a vector $v \in [0,1]^{498}$ whose jth component is the real number $\frac{1}{499}$ times $7^j (\mathrm{mod}\ 499)$, identifying $7^j (\mathrm{mod}\, 499)$ as an integer between 1 and 498. Here we use Mathematica's `PowerMod` as described in Part I, Section 21 on public-key cryptography. If we do Mathematica's `ListPlot[v]` for this vector of points in $[0,1]$, we will get a fairly random looking set of points in the plane. See Figure 10.

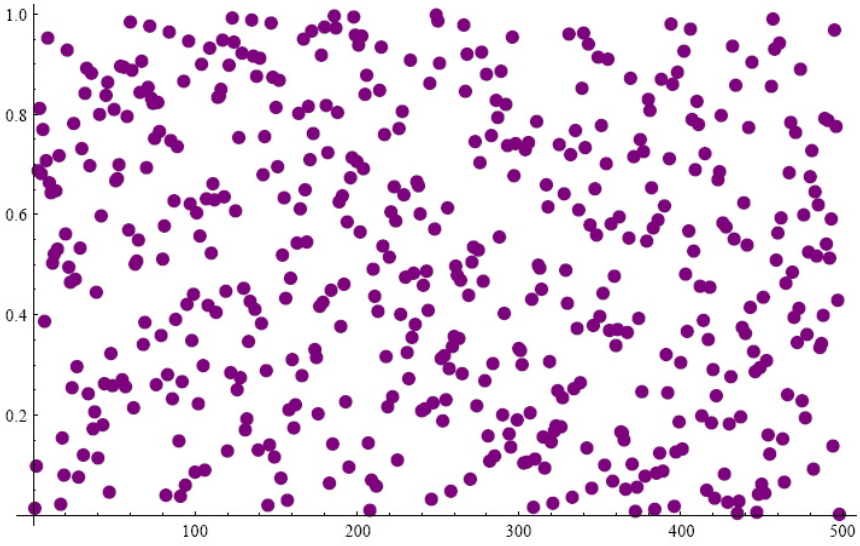

Figure 10: Plot of points $P_j = (v_j, j)$ whose 1st component is the real number $\frac{1}{499}$ times $7^j (\mathrm{mod}\, 499)$, identifying $7^j (\mathrm{mod}\, 499)$ as an integer between 1 and 498.

However, there is a pitfall in the method. We are really only allowed to think of this as a one dimensional thing. For if we try to plot points $(v_j, v_{j+1}) \in [0,1]^2$, we get Figure 11

The same sort of thing happens in 3D, taking points $(v_j, v_{j+1}, v_{j+2}) \in [0,1]^3$ to give Figure 12.

Suppose now we compute another random vector for a different prime modulus. We form $w \in [0,1]$, with w_j being the real number $\frac{1}{503}$ times $5^j (\mathrm{mod}\ 503)$, identifying $5^j (\mathrm{mod}\, 503)$ as an integer between 1 and 502. Then we plot points $P_j = (v_i, w_i) \in [0,1]$ in Figure 13

So the points formed using 2 random number generators look more random although connecting some dots might create some creatures. We can do a 3D plot of points formed using a 3rd random number generator. That is we create another vector $z \in [0,1]$, with z_j being the real number $\frac{1}{521}$ times $3^j (\mathrm{mod}\ 521)$, identifying $3^j (\mathrm{mod}\, 521)$ as an integer between 1 and 520. This gives us the Figure 14 plotting points $P_j = (v_i, w_i, z_i) \in [0,1]^3$.

Now we have no obvious hyperplanes.

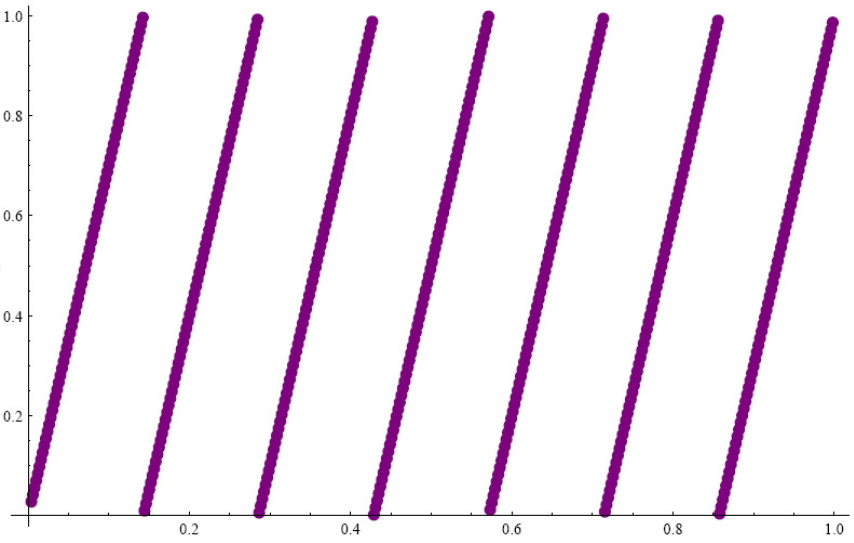

Figure 11: Plot of points $P_j = (v_j, v_{j+1})$ whose 1st component is the real number $\frac{1}{499}$ times $7^j \pmod{499}$, identifying $7^j \pmod{499}$ as an integer between 1 and 498.

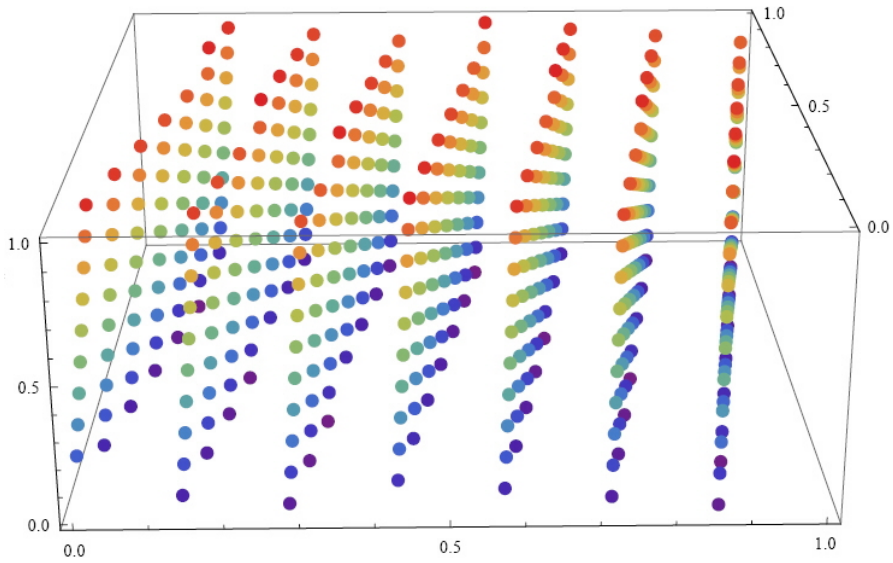

Figure 12: Plot of points $P_j = (v_j, v_{j+1}, v_{j+2})$ whose 1st component is the real number $\frac{1}{499}$ times $7^j \pmod{499}$, identifying $7^j \pmod{499}$ as an integer between 1 and 498.

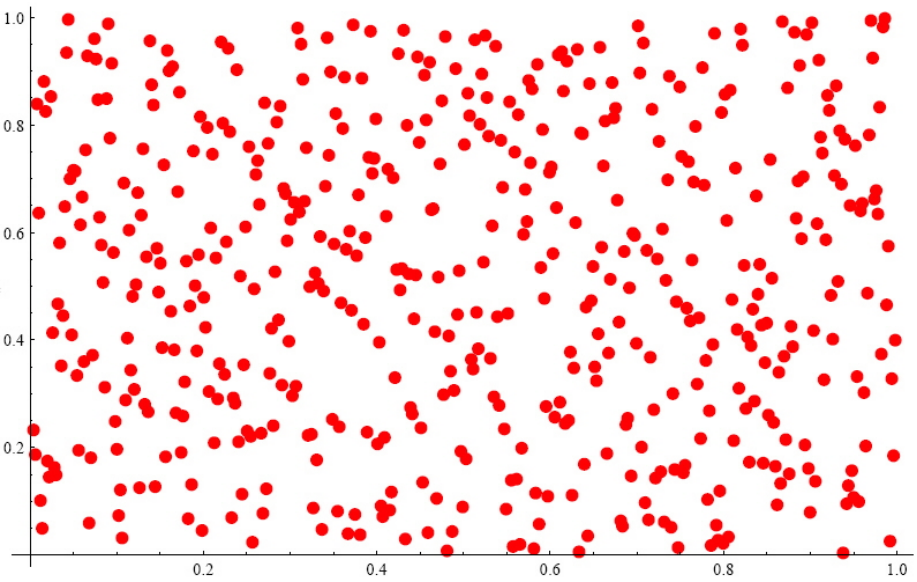

Figure 13: Plot of points $P_j = (v_j, w_j)$ whose 1st component is the real number $\frac{1}{499}$ times $7^j \pmod{499}$, identifying $7^j \pmod{499}$ as an integer between 1 and 498 and whose 2nd component is the analog with 499 replaced with 503.

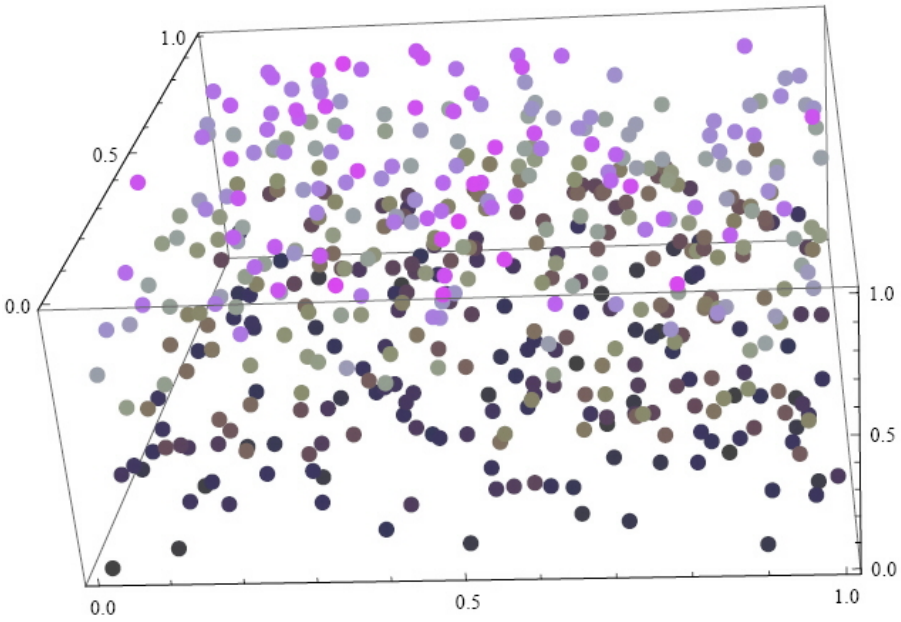

Figure 14: points (v_i, w_i, z_i) from 3 vectors v, w, z formed from powers of generators of $\mathbb{F}_p^*$ for $p = 499, 503$, and 521.

Applied mathematicians were in an uproar over the hyperplanes and did not want to use more than 1 generator, presumably worrying about slowing down the whole process. So new methods for random numbers arose. In 1995 Matlab switched to a Marsaglia generator. In the article listed at the beginning of the section, Brent noticed that the Marsaglia Xor Shift generator can be viewed as a linear feedback shift register. W. H. Press et al spend many pages bad mouthing the linear congruential generators and then list them as 2/3 of their methods. Mathematica gives 8 basic methods. Not surprisingly Wolfram's favorite - cellular automata - appear. You are also allowed to create your own generator.

Some of the generators come from generalizing the method we just illustrated in $\mathbb{F}_p$ to a method in $\mathbb{F}_{p^r}$ where you take a generator θ of the multiplicative group $\mathbb{F}_q^*$ and create the list $1, \theta, \theta^2, ...$ Writing $\theta^n = \sum_{j=0}^{n-1} c_j \theta^j$, where $c_j \in \mathbb{F}_p$. Here $\sum_{j=0}^{n} c_j \theta^j = 0$ with $c_n = -1$, is the irreducible primitive polynomial satisfied by θ. We can then look at the process via Feedback Shift Registers. Each successive multiplication of θ takes the registers $(a_0, a_1, ..., a_{n-1})$ to the registers $(a'_0, a'_1, ..., a'_{n-1}) = (a_{n-1}c_0, a_0 + a_{n-1}c_1, ..., a_{n-2} + a_{n-1}c_{n-1})$. This is the same as the matrix equation: $a' = Ma$, where our vectors are written as column vectors now and the matrix is

$$M = \begin{pmatrix} 0 & 0 & \cdots & 0 & c_0 \\ 1 & 0 & \cdots & 0 & c_1 \\ 0 & 1 & \cdots & 0 & c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & c_{n-1} \end{pmatrix}.$$

Now you can generate vectors in $\mathbb{F}_p^n$ and use them to produce vectors in $[0,1]^n$. In order to do this with big primes you would need a bigger list of primitive polynomials than that of the preceding section. Mathematica actually does calculations in finite fields. So perhaps it has a big list hidden somewhere.

Exercise. Do some experiments using finite fields to generate sequences of random vectors in $[0,1]^3$.

D. H. Lehmer (1905-1991) and his wife Emma were well known number theorists of the last century. Derrick's father was also a number theorist who had built a prime generating machine. Like his father, D.H. was on the U.C. Berkeley faculty - except for that short time in the 1950s when he was fired for refusing to sign the U.C. Regents' loyalty oath, which was ultimately declared unconstitutional. That was the era of Joe McCarthy's un-American activity committee and the blacklists. It was feared that communists were everywhere. During that period D.H. went to work for the National Bureau of Standards. Emma, being a woman, was never allowed to join the U.C.B. faculty. Yes, it was the bad old days.

I remember the Lehmers as 2 of the principal organizers of the West Coast Number Theory conference, a conference that I first attended as a young assistant professor in the 1970s. It is still meeting each year. Under their leadership this conference, was about the only democratically run math. conference that I ever attended (except the one I helped to start on automorphic forms). There was no elite bunch of organizers deciding who could speak and who couldn't. Instead, at the 1st night of the conference, anyone who wanted to give a talk would put their title on a piece of paper and the conference would be organized by putting talks on nearby subjects together. It was a great idea. Too bad more meetings are not more democratic and less elitist.

13 Error-Correcting Codes

References for this section include L. Dornhoff and F. Hohn, *Applied Modern Algebra,* J. Gallian, *Contemporary Abstract Algebra*; Vera Pless, *An Introduction to the Theory of Error-Correcting Codes;* Judy Walker, *Codes and Curves;* and A. Terras, *Fourier Analysis on Finite Groups and Applications.*

Suppose that I must send a message of $0's$ and $1's$ from my computer on earth to Mr. Spock's computer on Vulcan. No doubt errors will be introduced by transmission over such a long distance and some random 1 will turn into a 0. In order for Mr. Spock to figure out my message, there must be some redundancy built in. Error-correcting codes are created for that purpose. The original signal $s \in \mathbb{F}_2^n$ will be encoded as $x \in \mathbb{F}_2^{n+r}$. If errors are added in transmission of the encoded signal, Mr. Spock will use a decoder to find the most likely original signal $s' \in \mathbb{F}_2^n$ hoping that there is enough redundancy to do so. Such methods are used in compact discs as well as communications with spacecraft. The goal of error correction is really the opposite of the goal of cryptography. Here we want our message to be understood.

Figure 15: sending a message of 0's and 1's to Mr. Spock

R. W. Hamming (1915-1998) of Bell Telephone Labs published his codes in 1950. He had been working on a computer using punched cards. Whenever the machine detected an error, the computer would stop. Hamming got frustrated and began work on a way to correct errors. Hamming also introduced the Hamming distance defined below to get an estimate of the error in a signal. And he worked on the Manhattan project - doing simulations to model whether the atomic bomb would ignite the atmosphere. We cannot resist including a quote from Hamming's book *Digital Filters*: "... we will avoid becoming too involved with mathematical rigor, which all too often tends to become rigor mortis."

Definition 55 *A linear code C is a vector subspace C of $\mathbb{F}_q^n$.*

Here $\mathbb{F}_q$ denotes the field with q elements. If the dimension of C as a vector space over $\mathbb{F}_q$ is k, we call C an **$[n, k]$-code**. Since all codes we consider are linear, we will drop the word "linear" and just call them "codes". Here q will be 2 mostly. Such codes are called "**binary**." If $q = 3$, the code is "ternary."

Definition 56 *The **Hamming weight** of a codeword $x \in C$ is $|x|$= the number of components of x which are non-zero. The **distance** between $x, y \in C$ is defined to be $d(x, y) = |x - y|$.*

Exercise. For the vector space $V = \mathbb{F}_q^n$, show that the Hamming distance $d(x, y)$ has the following properties for all $x, y, u \in V$:

a) $d(x, y) = d(y, x)$;
b) $d(x, y) \geq 0$ and $d(x, y) = 0 \iff x = y$;
c) (triangle inequality) $d(x, y) \leq d(x, u) + d(u, y)$;
d) $d(x, y) = d(x + u, y + u)$.
e) $d(x, y) \in Z^+ \cup \{0\}$.

The first 3 properties of the Hamming distance in this exercise make it a **metric** on V. The 4th property makes it a translation invariant metric on V.

Exercise. a) For the prime p, show that when $p > n > 1$, the Hamming weight on $x \in \mathbb{F}_p^n$ satisfies

$$|x| \equiv x_1^{p-1} + \cdots + x_n^{p-1} (\mathrm{mod}\, p).$$

b) Consider $\mathbb{F}_p^n$ as a group under addition and form the Cayley graph $X(\mathbb{F}_p^n, S)$, where $S = \left\{ s \in \mathbb{F}_p^n \;\middle|\; |s| = 1 \right\}$. Draw some pictures for small values of p and n.

Definition 57 *If C is an $[n,k]$-code such that the minimum distance of a non-zero code word from 0 is d, we say that C is an* ***$[n, k, d]$-code****.*

The following theorem assumes you decode a received vector as the nearest codeword using the Hamming distance.

Theorem 58 *If $d = 2e + 1$, an $[n,k,d]$-code C corrects e or fewer errors.*

Proof. Suppose distinct $x, y \in C$ are such that $d(x,y) \geq 2e+1$. If the received word r has at most e errors, it cannot be in the Hamming ball of radius e about both x and y, since that would imply $0 < |x-y| = d(x,y) \leq d(x,r)+d(r,y) \leq e+e = 2e$. So the code can correct e errors. ■

We need more definitions.

Since an $[n,k]$ binary code C is a k-dimensional vector space over $\mathbb{F}_q$, C has a k-element basis, we can form a matrix whose rows are the basis vectors. This is called a **generator matrix** G of the code C. A generator matrix of an $[n,k]$-code is a $k \times n$ matrix of rank k with elements in $\mathbb{F}_q$. The code C is the image of the map sending the row vector $v \in \mathbb{F}_q^k$ to vG. Since C has more than one basis, it also has many generator matrices. The **standard generator matrix** has the form $G = (I_k \; A)$, where the first k columns form the $k \times k$ identity matrix I_k. If the generating matrix is in standard form, with no errors, decoding is easy, just take the first k entries of the code word. We know that we can use elementary row operations over $\mathbb{F}_q$ to put any generator matrix into row echelon form and that this must be the standard form $(I_k \; A)$ since this matrix must have rank k.

To describe the encoding envisioned here, we take the original message viewed as a row vector $s \in \mathbb{F}_q^k$ and then we encode the message as sG, adding redundancy to be able to correct errors.

A **parity check matrix** H of a $[n,k]$-code C is a matrix with n columns and rank $n-k$ such that $x \in C$ if and only if $xH = 0$. (Most texts take transpose H instead). If $G = (I_k \; A)$, then $H = \begin{pmatrix} -A \\ I_{n-k} \end{pmatrix}$.

Exercise. For code C with generator matrix $G = (I_k \; A)$ show that $x \in C$ iff $xH = 0$.

Hint. It is easy to see that C lies in the kernel of the linear transformation L sending $x \in \mathbb{F}_q^n$ to xH. Since

$$\dim \ker L + \dim L\left(\mathbb{F}_q^n\right) = n,$$

we find that the dimension of $\ker L$ is k and thus obtain the equality of the kernel of L and the code C.

The parity check matrix is quite useful for decoding. See Gallian, *Contemporary Abstract Algebra* for more information.

Our next question is: Where does all our theory of finite fields come in?

Definition 59 *A linear* ***cyclic code*** *is a linear code C with the property that if $c = (c_0, c_1, ..., c_{n-2}, c_{n-1})$ is a code word then so is $(c_{n-1}, c_0, ..., c_{n-3}, c_{n-2})$.*

Let R denote the factor ring $R = \mathbb{F}_q[x]/ < x^n - 1 >$. Represent elements of R by polynomials with coefficients in $\mathbb{F}_q$ of degree $< n$. Identify codeword $c = (c_0, c_1, ..., c_{n-2}, c_{n-1})$ with (the coset of) the polynomial $c_0 + c_1 x + \cdots + c_{n-1}x^{n-1}$.

Theorem 60 *A linear code C in R is cyclic if and only if it is an ideal in the ring $R = \mathbb{F}_q[x]/ < x^n - 1 >$.*

Proof. First note that a subspace W of R is an ideal if $xW \subset W$, because this implies $x^j W \subset W$, for all $j = 2, 3, ...$ Thus $RW \subset W$.

Now suppose that C is an ideal and $c_0 + c_1 x + \cdots + c_{n-1}x^{n-1} \in C$. Then C contains $x\left(c_0 + c_1 x + \cdots + c_{n-1}x^{n-1}\right) = c_0 x + c_1 x^2 + \cdots + c_{n-1}x^n = c_{n-1} + c_0 x + c_1 x^2 + \cdots + c_{n-2}x^{n-1} \left(\mathrm{mod}(x^n - 1)\right)$. The last happens because x^n is congruent to 1 modulo $\langle x^n - 1 \rangle$. So C is cyclic. ■

Question. What are the ideals A in the ring $R = \mathbb{F}_q[x]/ < x^n - 1 >$?

Answer. Just as we found in Section 4 for ideals in $\mathbb{Z}_{12}$, they are principal ideals $\langle g(x) \rangle$, where $g(x)$ divides $x^n - 1$. We call $g(x)$ the generator of A. If $g(x) = c_0 + c_1 x + \cdots + c_r x^r$ has degree r, then the corresponding code is an $[n, n-r]$-code and a generator matrix for the code (as defined above) is the $(n-r) \times n$ matrix:

$$\begin{pmatrix} c_0 & c_1 & c_2 & \cdots & c_r & 0 & 0 & \cdots & 0 \\ 0 & c_0 & c_1 & \cdots & c_{r-1} & c_r & 0 & \cdots & 0 \\ 0 & 0 & c_0 & \cdots & c_{r-2} & c_{r-1} & c_r & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & c_0 & c_1 & c_2 & \cdots & c_r \end{pmatrix}. \tag{9}$$

Exercise. Show that the code described above has dimension $n - r$.

Hint. The cosets of the vectors $g(x)x^j$, $j = 0, ..., n - r - 1$, are linearly independent in the ring R. These vectors span the ideal $A = \langle g(x) \rangle$ since elements of A have the form $f(x)g(x)$, for some polynomial $f(x)$ of degree less than or equal to $n - r - 1$.

Example 1. The Hamming [7, 4, 3]-code.

Note that the polynomial $x^7 - 1$ can be completely factored into irreducibles over $\mathbb{F}_2$ as follows:

$$x^7 - 1 = (x - 1)(x^3 + x + 1)(x^3 + x^2 + 1) \text{ in } \mathbb{F}_2[x].$$

Take $g(x) = x^3 + x + 1$ to generate our ideal I in R corresponding to the code. The codewords in C in are listed below.

0	0	0	0	0	0	0		1	1	1	1	1	1	1
1	1	0	1	0	0	0		0	0	1	0	1	1	1
0	1	1	0	1	0	0		1	0	0	1	0	1	1
0	0	1	1	0	1	0		1	1	0	0	1	0	1
0	0	0	1	1	0	1		1	1	1	0	0	1	0
1	0	0	0	1	1	0		0	1	1	1	0	0	1
0	1	0	0	0	1	1		1	0	1	1	1	0	0
1	0	1	0	0	0	1		0	1	0	1	1	1	0

A generator matrix corresponding to $g(x) = x^3 + x + 1$ as in formula (9) is

$$G = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

Exercise. a) Use elementary row operations to put the generator matrix G above into row-reduced echelon (standard) form.

b) Explain why the code words listed above are correct by making a table listing the 16 elements of $x \in \mathbb{F}_2^4$ (the possible messages) in the 1st column and then listing the corresponding xG (the encodings of the messages) in the 2nd column.

Exercise. Imitate the preceding example except use the polynomial $x^3 + x^2 + 1$ to build the code instead of $x^3 + x + 1$.

Suppose $g(x)h(x) = x^n - 1$, in $\mathbb{F}_2[x]$, with $g(x)$ of degree r, the generator polynomial of a code C and $h(x)$ of degree $k = n - r$. Then we get a parity check matrix for the code from the matrix of the polynomial $h(x) = h_0 + h_1 x + \cdots + h_k x^k$ as follows:

$$\begin{pmatrix} h_k & 0 & \cdots & 0 \\ h_{k-1} & h_k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ h_0 & h_1 & \cdots & h_k \\ 0 & h_0 & \cdots & h_{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_0 \end{pmatrix}.$$

Exercise. a) Show that the preceding matrix is indeed a parity - check matrix for our code with generator polynomial $g(x)$ as described above.

b) Find the parity check matrix for the Hamming $[7, 4, 3]$-code above.

There is a method for constructing codes that correct lots of errors called BCH codes. See Dornhoff and Hohn, *Applied Modern Algebra*, page 442, for the mathematical details. Let us sketch a bit of the theory.

Definition 61 *Suppose that $\gamma \in E$ and $E \supset F$ are finite fields. Define the* ***minimal polynomial*** *of γ over F to be the polynomial $f \in F[x]$ of least degree such that $f(\gamma) = 0$.*

Recall that we obtained the Hamming $[7, 4, 3]$-code by looking at the generator polynomial $g(x) = x^3 + x + 1$. This is the minimal polynomial of an element γ of $\mathbb{F}_8$ whose other roots are γ^2 and γ^4. So we could say that any polynomial $f(x)$ is in our code C iff $f(\gamma^j) = 0$, $j = 1, 2, 4$. For any polynomial whose roots include the roots of $g(x)$ must be divisible by $g(x)$.

Definition 62 *A* ***primitive nth root of 1*** *in a field K is a solution γ to $\gamma^n = 1$ such that $\gamma^m \neq 1$, for $1 \leq m < n$.*

Theorem 63 *(Bose-Chaudhuri and Hoquenghem,1960) Suppose* $\gcd(n, q) = 1$. *Let γ be a primitive nth root of* 1 *in an extension field of $\mathbb{F}_q$. Suppose the generator polynomial $g(x)$ of a cyclic code of length n over $\mathbb{F}_q$ has $\gamma, \gamma^2, ..., \gamma^{d-1}$ among its roots. Then the minimum distance of a code element from* 0 *is at least d.*

For a proof see Dornhoff and Hohn, pages 442-3.

A Reed-Solomon code is a BCH code with $n = q - 1$. These codes are used by the makers of CD players, NASA, ... These can be used to correct amazing numbers of errors. If you suppose $q = 2^8$ so that $n = 255$, a 5-error-correcting code has $g(x) = (x - \gamma)(x - \gamma^2) \cdots (x - \gamma^{10})$ of degree 10. Elements of $\mathbb{F}_{2^8}$ are 8-dimensional vectors over $\mathbb{F}_2$. This code can be used as a code of length $8 * 255 = 2040$ over $\mathbb{F}_2$, which can correct any "burst" of 33 consecutive errors. See Dornhoff and Hohn, p. 444. For any 33 consecutive errors over $\mathbb{F}_2$ will affect at most 5 of the elements of $\mathbb{F}_8$.

Feedback shift registers are of use in encoding and decoding cyclic codes. See Dornhoff and Hohn, pp. 449 ff.

Example. Codes from the Hadamard Matrix.

The code used in the 1969 NASA Mariner 9 spacecraft which orbited Mars comes from the Hamming matrix $H_{2^5} = \left((-1)^{u \cdot v}\right)_{u,v \in \mathbb{F}_2^5}$, with u, v ordered as for the corresponding numbers in binary and $u \cdot v = \sum_{i=1}^{5} u_i v_i$. This matrix is pictured in Figure 16.

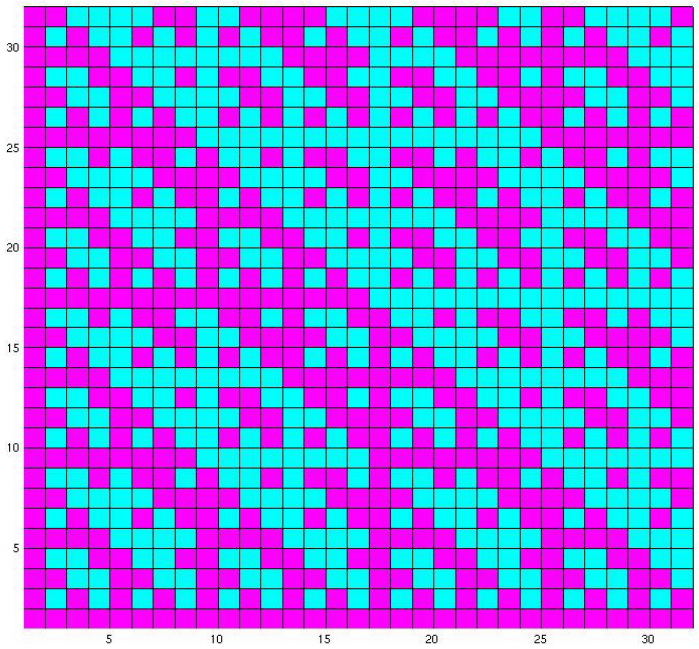

Figure 16: The matrix H_{32} where the $1's$ and -$1's$ have become purple and turquoise.

The code is found by forming the new matrix $G = \Phi \begin{pmatrix} H_{2^5} \\ -H_{2^5} \end{pmatrix}$, where Φ replaces 1s with 0s and -1s with 1s. The rows of G are the codewords of the $[32, 6, 16]$ biorthogonal Reed-Muller code used in the Mariner Mars probe.

Exercise. Why is the $[32, 6, 16]$ biorthogonal Reed-Muller code described above actually 6 dimensional with minimum weight 16 and how many errors can it correct?

The general Hadamard matrix $H_{2^n} = \left((-1)^{u \cdot v}\right)_{u,v \in \mathbb{F}_2^n}$ has the recursive definition $H_{2^n} = \begin{pmatrix} H_{2^n} & H_{2^n} \\ H_{2^n} & -H_{2^n} \end{pmatrix}$ with $H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$. The matrix H_m is defined to be a matrix of 1s and -1s such that $H_m \, {}^t H_m = mI$, where I is the identity matrix. For H_m to exist with $m > 2$, it is necessary that 4 divide m. The smallest values of m without a construction of H_m are $m = 428$ and 668, according to K. H. Rosen and J. G. Michaels, *Handbook of Discrete and Combinatorial Mathematics.* (1999).

Why did Hadamard study these matrices? He wanted a matrix such that H_m with entries h_{ij} such that $|h_{ij}| \leq 1$ and $|\det(H_m)|$ is maximal (i.e., $m^{m/2}$). In Terras, *Fourier Analysis on Finite Groups and Applications*, p. 172, we note that the Hadamard matrix H_{2^n} is the matrix of the Fourier Transform (or DFT) on the group $\mathbb{F}_2^n$.

H. B. Mann (Ed.), *Error Correcting Codes*, gives more information on the code used in the Mariner Mars probe, as well as on the history of error-correcting codes. In this book one finds a limerick inspired by the coding theorist Jessie MacWilliams:

"Delight in your algebra dressy
But take heed from a lady named Jessie
Who spoke to us here of her primitive fear
That good codes just might be messy."

W. W. Rouse Ball and H. S. M. Coxeter, *Mathematical Recreations and Essays*, give more recreational aspects of Hadamard matrices. See also F. Jessie Mac Williams and N. J. A. Sloane, *The Theory of Error-Correcting Codes.*

Exercise. Consider the $[12, 6]$ extended ternary Golay code with generator matrix $(I_6 \ A)$, where

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 2 & 2 & 1 \\ 1 & 1 & 0 & 1 & 2 & 2 \\ 1 & 2 & 1 & 0 & 1 & 2 \\ 1 & 2 & 2 & 1 & 0 & 1 \\ 1 & 1 & 2 & 2 & 1 & 0 \end{pmatrix}.$$

Show that the minimum Hamming weight of a codeword is 6.

This code was found by M. J. Golay in 1949. The ternary $[11, 6]$ cyclic code can be found by factoring

$$x^{11} - 1 = (x-1)\left(x^5 - x^3 + x^2 - x - 1\right)\left(-x^5 - x^4 + x^3 - x^2 + 1\right).$$

This is also a quadratic residue code. To obtain a ternary quadratic residue code we proceed as follows. Let p be a prime such that 3 is a square mod p. Here $p = 11$. Suppose ζ is a primitive pth root of unity in some field containing $\mathbb{F}_3$. Then let $\square$ denote the set of squares in $\mathbb{F}_p^*$ and let $\square'$ be the set of non-squares in $\mathbb{F}_p^*$. Define the polynomials

$$q(x) = \prod_{j \in \square} \left(x - \zeta^j\right) \text{ and } n(x) = \prod_{j \in \square'} \left(x - \zeta^j\right).$$

One can show that the polynomials $q(x)$ and $n(x)$ have coefficients in $\mathbb{F}_3$ and that

$$x^p - 1 = (x-1)q(x)n(x).$$

14 Finite Upper Half Planes and Ramanujan Graphs

We imitate the real Poincaré upper half plane H consisting of points $z = x + iy$, with $x, y \in \mathbb{R}$ and $y > 0$. It has a distance $ds^2 = \frac{dx^2+dy^2}{y^2}$ which is invariant under fractional linear transformation $z \to \frac{az+b}{cz+d}$, with $ad - bc > 0$. Moreover, the distance minimizing curves or geodesics are half lines and half circles perpendicular to the real axis. Viewing these as the straight lines of our geometry makes Euclid's 5th postulate false. Thus we get Poincaré's model of non-Euclidean geometry. See Terras, *Fourier Analysis on Symmetric Spaces*, I, for more information. Number theorists are enamoured of functions on H which have invariance properties under action of the modular group of fractional linear transformations with integer a, b, c, d and $ad - bc = 1$.

Here we want to consider a finite analog of the Poincaré upper half plane. Suppose that $\mathbb{F}_q$ is a finite field of odd characteristic p. This implies that $q = p^r$. Suppose δ is a fixed non-square in $\mathbb{F}_q$. The **finite upper half plane** over $\mathbb{F}_q$ is defined to be

$$H_q = \left\{ z = x + y\sqrt{\delta} | x, y \in \mathbb{F}_q, y \neq 0 \right\}.$$

We will write for $z = x + y\sqrt{\delta} \in x, y \in \mathbb{F}_q\left(\sqrt{\delta}\right)$, with $x, y \in \mathbb{F}_q$, **real part** of $z = \mathrm{Re}(z) = x$, **imaginary part** of $z = \mathrm{Im}(z) = y$, **Conjugate** of $z = \overline{z} = z - y\sqrt{\delta} = z^q$. **Norm** of $z = Nz = z\overline{z}$.

Perhaps you will object to the use of the word "upper." Since we have no good notion of $>$ for finite fields, we use the word "upper" thinking, for example, if $q = p$, the y-coordinate of a point is in the set $\{1, 2, .., p-1\}$ of "positive" numbers. That is perhaps a cheat and we should really view H_q as a union of an upper and a lower half plane, with the $y-$ coordinate of a point in the set $\left\{-\frac{p-1}{2}, ..., -1, 1, ..., \frac{p-1}{2}\right\}$.

The **general linear group** $GL(2, \mathbb{F}_q)$ of matrices $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $ad - bc \neq 0$ acts on $z \in H_q$ by fractional linear transformation: $gz = \frac{az+b}{cz+d}$.

Exercise. Show that $\mathrm{Im}(gz) \neq 0$, for any $g \in GL(2, \mathbb{F}_q)$ and any $z \in H_q$.

Exercise. a) Show that

$$K = \left\{ g \in GL(2, \mathbb{F}_q) \mid g\sqrt{\delta} = \sqrt{\delta} \right\} = \left\{ \begin{pmatrix} a & b\delta \\ b & a \end{pmatrix} \;\middle|\; a, b \in \mathbb{F}_q \text{ with } a^2 - \delta b^2 \neq 0 \right\}.$$

b) Show that K is a subgroup of $G = GL(2, \mathbb{F}_q)$ which is isomorphic to the multiplicative group $\mathbb{F}_q\left(\sqrt{\delta}\right)^*$.

Hint. The isomorphism is give by $\begin{pmatrix} a & b\delta \\ b & a \end{pmatrix} \to a + b\sqrt{\delta}$.

The subgroup K of $GL(2, \mathbb{F}_q)$ is analogous to the **orthogonal subgroup** of the general linear group $GL(2, \mathbb{R})$, namely, $O(2, \mathbb{R}) = \left\{ g \in GL(2, \mathbb{R}) \mid {}^T gg = I \right\}$ consisting of rotation matrices.

The **finite Poincaré distance** on H_q is defined to be $d(z, w) = \frac{N(z-w)}{\mathrm{Im}\, z\, \mathrm{Im}\, w}$. The distance has values in $\mathbb{F}_q$. Thus we are not talking about a metric here. There is no possibility of a triangle inequality.

Exercise. Let $z = x + y\sqrt{\delta}$ and $w = u + v\sqrt{\delta}$, with $x, y, u, v \in \mathbb{F}_q$ and $yv \neq 0$. Show that

$$d(z, w) = \frac{(x-u)^2 - \delta(y-v)^2}{yv}.$$

Exercise. Show that $d(gz, gw) = d(z, w)$ for all $g \in GL(2, \mathbb{F}_q)$ and all $z, w \in H_q$.

We can draw a contour map of the distance function by making a grid representing the finite upper half plane and coloring the point $x + y\sqrt{\delta}$ according to the value of $d(z, \sqrt{\delta}) = \frac{x^2 - \delta(y-1)^2}{y}$. When $q = 163$ we get Figure 17. This figure should be compared with the analogous figure obtained using an analog of the Euclidean distance on a finite plane given in Figure 1. I see monsters in Figure 17.

Exercise. Make a figure analogous to Figure 17 using the distance $d((x, y), (0, 0)) = x^4 + y^4$, for $(x, y) \in \mathbb{F}_p \times \mathbb{F}_p$, where p is your favorite prime.

Next we want to define some graphs attached to this stuff.

Definition 64 *Let $a \in \mathbb{F}_q$ and define the **finite upper half plane graph** $X_q(\delta, a)$ to have vertices the elements of H_q and then draw an edge between 2 vertices z, w iff $d(z, w) = a$.*

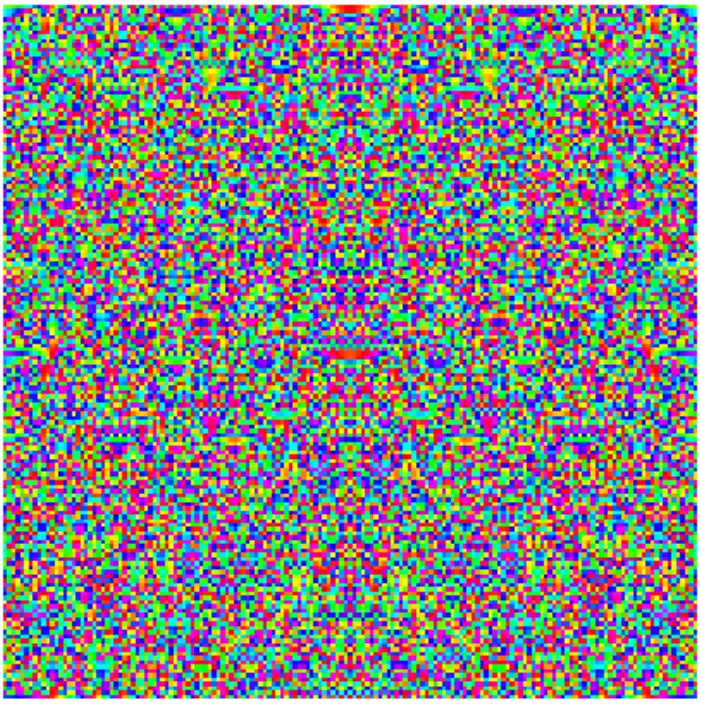

Figure 17: color at point $z = x + y\sqrt{\delta}$ in H_{163} is found by computing the Poincaré distance $d(z, \sqrt{\delta})$.

Example. The Octahedron. Let $q = 3$, $\delta = -1 \equiv 2 (\mathrm{mod}\, 3)$. We will write $i = \sqrt{-1}$. To draw the graph $X_3(-1, 1)$ we need to find the points adjacent to i for example. These are the points $z = x + iy$ such that $d(z, i) = \frac{N(z-i)}{y} = 1$. This is equivalent to solving $x^2 + (y-1)^2 = 1$. Solutions are the 4 points $1 + i, 1 - i, -1 + i, -1 - i$. To find the points adjacent to any point $a + bi \in H_3$, just apply the matrix $\begin{pmatrix} b & a \\ 0 & 1 \end{pmatrix}$ to the points $\pm 1 \pm i$ that we just found. The graph $X_3(-1, 1)$ is drawn on the left in Figure **??** It is an octahedron.

The adjacency matrix A of the octahedron graph is the 6×6 matrix below of $0s$ and $1s$ where the i, j entry is 1 iff vertex i is adjacent (i.e., joined by an edge) to vertex j.

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

The eigenvalues $\lambda \in \mathbb{C}$ of A are the solutions of $\det(A - \lambda I) = 0$. The set of eigenvalues is $\mathrm{Spectrum}(A) = \{4, -2, -2, 0, 0, 0\}$. Note that the $2nd$ largest eigenvalue in absolute value, which is $|\lambda| = 2$, satisfies $|\lambda| \leq 2\sqrt{3} \cong 3.46$. This means that the graph $X_3(-1, 1)$ is what is called a **Ramanujan graph**.

Figure **??** also shows $X_5(2, 1)$ on the right. The solid lines are the edges of the graph. The dotted lines are the edges of a dodecahedron. We can view the graph $X_5(2, 1)$ as that which you get by putting a 5-pointed star on each face of a dodecahedron.

A graph X is called $k - regular$ if there are k edges coming out of every vertex. We say that a k-regular graph is **Ramanujan** if for all eigenvalues λ of the adjacency matrix such that $|\lambda| \neq k$, we have $|\lambda| \leq 2\sqrt{k-1}$. This definition was made by Lubotzky, Phillips and Sarnak in a paper from 1988. It turns out that such graphs provide good communication networks as the random walk on them converges rapidly to uniform. In the 1980's Margulis and independently Lubotzky, Phillips and Sarnak found infinite families of Ramanujan graphs of fixed degree. See Guiliana Davidoff, Peter Sarnak, and Alain Valette, *Elementary number theory, group theory and Ramanujan graphs* or A. Terras, *Fourier Analysis on Finite*

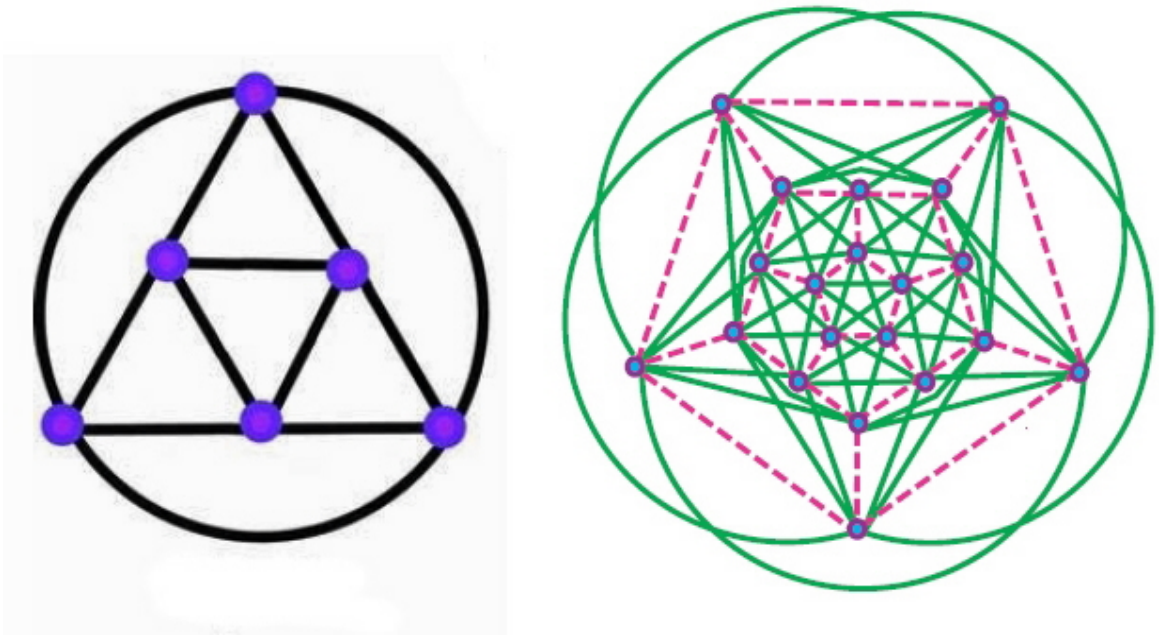

Figure 18: The graphs on the left is $X_3(-1,1)$, an octahedron, and that on the right is $X_5(2,1)$ with the edges in green. The red dashed lines on the right are the dodecahedron.

Groups and Applications. Denis Charles, Eyal Goren, and Kristin Lauter give applications of expanders to cryptography in "Cryptographic hash functions from expander graphs."

Of course, one really wants infinite families of Ramanujan graphs of fixed small degree. The finite upper half plane graphs $X_q(\delta, a)$ have degree $q+1$ provided that $a \neq 0$ or 4δ. These finite upper half plane graphs were proved to be Ramanujan by N. Katz using work of Soto-Andrade. Ramanujan graphs are also good expanders, meaning that if they form a gossip network the gossip gets out fast. Sarnak says in his article "What is an expander?" from the *Notices of the American Math. Society*, 51 (August 2004), pp. 762-3: "... it is in applications in theoretical computer science where expanders have had their major impact. Among their applications are the design of explicit superefficient communication networks, constructions of error-correcting codes with very efficient encoding and decoding algorithms, derandomization of random algorithms, and analysis of algorithms in computational group theory ...."

Now we can explain Figure 2 in Section 1. The picture is that of points (x, y), with $x, y \in \mathbb{F}_{121}$ and $y \neq 0$. Take $\delta \in \mathbb{F}_{121}$ to be a non-square. View a point (x, y) as $z = x + y\sqrt{\delta} \in H_{121} \subset \mathbb{F}_{121}[\sqrt{\delta}]$. Let 2×2 matrices $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL(2, \mathbb{F}_{11})$ act on $z = x + y\sqrt{\delta}$ by fractional linear transformation $gz = \frac{az+b}{cz+d}$. Color 2 points z and w the same color if there is a matrix $g \in GL(2, \mathbb{F}_{11})$ such that $w = gz$. This gives the picture in Figure 2. Figure 19 is another version of that figure. This figure is reminiscent of tessellations of the real Poincaré upper half plane H obtained by translating a fundamental domain $D \cong \Gamma\backslash H$ around using elements of the modular group $SL(2, \mathbb{Z})$. There are some beautiful tessellations on Helena Verrill's website: www.math.lsu/~verrill/.

Exercise. Apply Burnside's Lemma from Section 20 of Part I to $GL(2, \mathbb{F}_p)$ acting on H_{p^2} to find out how many colors need to be used in creating the analog of Figure for an arbitrary odd prime p.

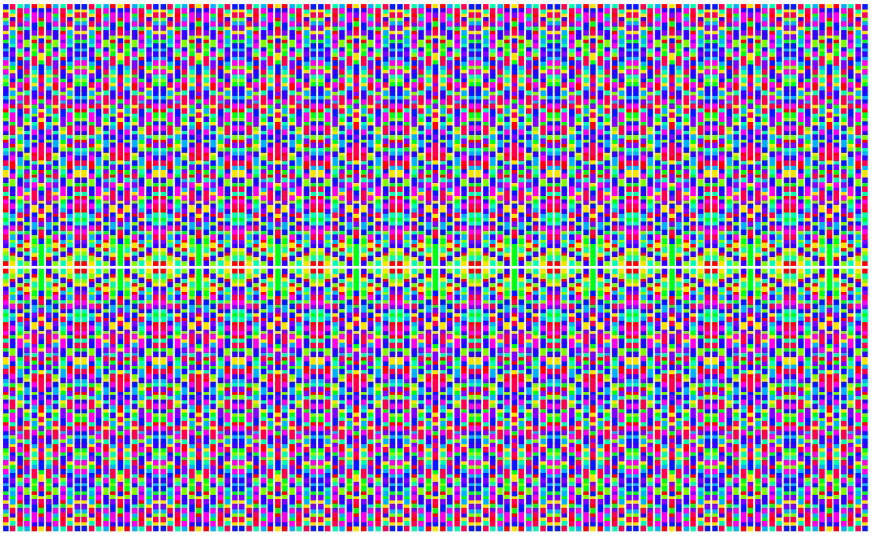

Figure 19: another version of Figure 2

15 Eigenvalues, Random Walks on Graphs, and Google

Notation. All the vectors in this section will be **column vectors** in $\mathbb{C}^n$. Thus our matrices $A \in \mathbb{C}^{n\times n}$ will act on the left taking $v \in \mathbb{C}^n$ to Av. For a matrix $M \in \mathbb{C}^{n\times n}$, the transpose of M is denoted ${}^TM$.

Given an $n \times n$ matrix A whose entries are complex numbers, we say that $\lambda \in \mathbb{C}$ is an **eigenvalue** of A iff $\det(A-\lambda I) = 0$, where I is the $n \times n$ identity matrix. This is the same thing as saying that the matrix $A - \lambda I$ is singular; or that $Ax = \lambda x$ for some non-0 column vector $x \in \mathbb{C}^n$. Then we say x is an **eigenvector** of A corresponding to the eigenvalue λ. The set of all the eigenvalues of the matrix A is called the **spectrum** of A. We will denote it $spec\,(A)$. The name eigenvalue comes from D. Hilbert in 1904. Many other words have been used. P. Halmos (in *Finite Dimensional Vector Spaces*, p. 102) said: "Almost every combination of the adjectives proper, latent, characteristic, eigen, and secular, with the nouns root, number, and value has been used in the literature ..."

Exercise. Find the eigenvalues of the following matrices:

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

If the following exercises are too terrible, you can find them in most linear algebra books.

Exercise. a) Show that for any matrix $A \in \mathbb{C}^{n\times n}$ there is a unitary matrix U (meaning that ${}^T\overline{U}\ U = I$) and an upper triangular matrix T, with $U, T \in \mathbb{C}^{n\times n}$, such that $A = {}^T\overline{U}\ TU$. This is called the **Schur decomposition** of A. Since ${}^T\overline{U} = U^{-1}$, this says that the matrix A is similar to T.

b) Then show that if $A = {}^T\overline{U}\ TU$ as in part a), the diagonal entries of T are the eigenvalues of A.

Hint on a). We know that $\det(A - \lambda I) = 0$ has a root λ_1. Therefore there is a corresponding eigenvector $v_1 \neq 0$ such that $Av_1 = \lambda_1 v_1$. Upon multiplying v_1 by a scalar, we may assume that $\|v_1\| = 1$. Complete v_1 to an orthonormal basis $\{v_1, v_2, \cdots, v_n\}$ of $\mathbb{C}^4$ using the Gram-Schmidt process. Then $U_1 = (v_1 v_2 \cdots v_n)$ is a unitary matrix. And $U_1^{-1}AU_1 = \begin{pmatrix} \lambda_1 & * \\ 0 & A_2 \end{pmatrix}$, where $A_2 \in \mathbb{C}^{(n-1)\times(n-1)}$. Use induction on n to complete the proof.

Exercise. a) Suppose that the matrix $A \in \mathbb{C}^{n\times n}$ is **Hermitian** meaning that ${}^T\overline{A} = A$. Show that then the upper triangular matrix T in the Schur decomposition of A can be taken to be diagonal. This is the **spectral theorem**.

b) Show that the eigenvalues of a Hermitian matrix are real numbers.

There are many applications of these concepts to engineering, physics, chemistry, statistics, economics, music, even the internet. Eigenvalues associated to structures can be used to analyze their stability under some kind of vibration such as that caused by an earthquake. The word "spectroscopy" means the use of spectral lines to analyze chemicals. We will investigate one such application in this section. References for this section include: Google's website, G. Strang, *Linear Algebra and its Applications*; C. D. Meyer, *Matrix Algebra and Applied Linear Algebra*; R. A. Horn and C. R. Johnson, *Matrix Analysis*; Amy N. Langville and Carl D. Meyer, *Google's Pagerank and Beyond: the Science of Search Engine Rankings*; D. Cvetković, M. Doob, and H. Sachs, *Spectra of Graphs*; A. Terras, *Fourier Analysis on Finite Groups and Applications*.

This section concerns real and complex linear algebra, the sort you learn as a beginning undergrad, for the most part, except for the Perron-Frobenius theorem. We will not be thinking about matrices with elements in finite fields in this section. Usually our matrices will have elements that are nonnegative real numbers. That's because our matrices will be Markov matrices from elementary probability theory. Markov invented this concept in 1907. Markov chains are random processes that retain no memory of where it was in the past. An example is a random walk on the pentagon graph below. References for the subject are J. C. Kemeny and J. L. Snell, *Finite Markov Chains* and J. R. Norris, *Markov Chains*.

A **Markov matrix** $M \in \mathbb{R}^{n\times n}$ means that the entries are in the interval $[0, 1]$ and the columns sum to 1. From such a matrix you get a Markov chain of probability vectors v, where **probability vector** means that the entries are in $[0, 1]$ and sum to 1. If we are given a **probability vector** $v_0 \in [0, 1]^n$ then we get a Markov chain $v_0, v_1 = Av_0, v_2 = Av_1, ..., v_{n+1} = Av_n$. All the vectors v_n are probability vectors. At time n, the vector v_n has jth component which should be interpreted as the probability that the system is in its jth state at time n. In the example which follows of a random walk on a pentagon graph, the jth component is the probability that the random walker is at vertex j of the graph. In the case of the Google Markov matrix, the jth component is the probability that a websurfer is at the jth website.

Exercise. Show that if M is a Markov matrix and v is a probability vector, then Mv is also a probability vector.

Because, in general, a Markov matrix need not be symmetric, its eigenvalues need not be real numbers. That makes the analysis of the behavior of the Markov chain a little more delicate. The spectral theorem of the Exercise above is not sufficient. One needs the Perron theorem (see Theorem 65 below) or more generally the Perron-Frobenius theorem.

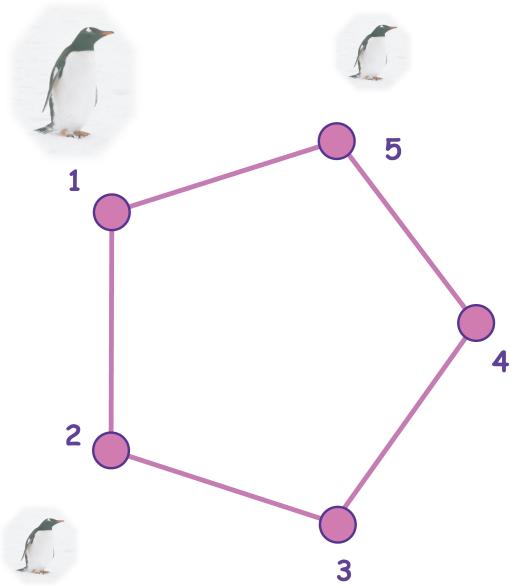

Figure 20: A random walk on a pentagon. At time $t = 0$, the big penguin is at vertex 1. At time $t = 1$ the penguin has probability $\frac{1}{2}$ of being at vertex 2 and probability $\frac{1}{2}$ of being at vertex 5. So the penguins at these vertices are half size.

Example. A Random Walk on a Pentagon.

Consider the pentagon graph. This is the Cayley graph $X(\mathbb{Z}_5, \{\pm 1 (\mathrm{mod}\, 5)\})$. It is undirected. The associated Markov matrix for the random walk in which a creature moves from vertex x to vertex $x + 1 (\mathrm{mod}\, 5)$ or $x - 1 (\mathrm{mod}\, 5)$ with equal probability is

$$M = \begin{pmatrix} 0 & .5 & 0 & 0 & .5 \\ .5 & 0 & .5 & 0 & 0 \\ 0 & .5 & 0 & .5 & 0 \\ 0 & 0 & .5 & 0 & .5 \\ .5 & 0 & 0 & .5 & 0 \end{pmatrix}.$$

See Figure 20. Note that $M = {}^T M$; i.e., M is real symmetric. Thus it has real eigenvalues which must include 1 since the vector ${}^t w = (1, 1, 1, 1, 1)$ is satisfies $Mv = v$. Scientific Workplace tells me that the other eigenvalues are approximately : -0.8090, 0.3090, each with multiplicity 2.

If we start our random walker at vertex 1, that corresponds to the probability vector ${}^T v_0 = (1, 0, 0, 0, 0)$. Then at time $t = 1$ the creature is either at vertex 2 or 5 with equal probability. That corresponds to the probability vector $v_1 = M v_0$, ${}^T v_1 = (0, 0.5, 0, 0, 0.5)$. Then at time $t = 2$ we have the probability vector $v_2 = M^2 v_0$, ${}^T v_2 = (0.5000, 0, 0.2500, 0.2500, 0)$. At time $t = 3$ we have $v_3 = M^3 v_0$, ${}^T v_3 = (0, 0.3750, 0.1250, 0.1250, 0.3750)$. Continue in this manner up to time $t = 10$ and you find that $v_{10} = M^{10} v_0$, ${}^T v_{10} = (0.2480, 0.1611, 0.2148, 0.2148, 0.1611)$. Already we see that we are approaching the eigenvector $\frac{1}{5} w = (.2, .2, .2, .2, .2)$ which is the probability that the poor creature is totally lost, also known as the uniform probability distribution. The speed of convergence to the uniform probability vector is governed by the second largest eigenvalue which is .8090 in this case. See my book, *Fourier Analysis on Finite Groups and Applications*, pp. 104-5, for a proof. You need to be a time t such that $.8090^t$ is negligible (depending on what metric you use on the space of vectors in $\mathbb{R}^5$). Anyway, for our example, at time $t = 30$, the probability vector is $v_{30} = M^{30} v_0$, ${}^T v_{30} = (0.2007, 0.1994, 0.2002, 0.2002, 0.1994)$ which is close enough to $u = (.2, .2, .2, .2, .2)$ not to be able to notice the difference on a picture. Note that $.8090^{30} \cong .00173$. The actual Euclidean distance between the 2 vectors is

$$\|v_{30} - u\|_2 = \sqrt{(.0007) + 2\,(.0006)^2 + 2\,(.0002)^2} \cong 0.03.$$

If our graph were the web, we'd be saying all websites have the same rank since all the coefficients of the steady-state vector u are equal.

Exercise. Prove that if M is a symmetric $n \times n$ Markov matrix and $v_1, ..., v_n$ is an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of M such that $Mv_1 = v_1$, then

$$\lim_{t\to\infty} M^t v = u, \quad \text{where} \quad u = {}^T\left(\frac{1}{n}, ..., \frac{1}{n}\right), \tag{10}$$

for any probability vector v.

Hint. Write $v = \sum_{j=1}^{n} \gamma_j v_j$, for $\gamma_j \in \mathbb{R}$. Apply M^t to both sides and take the limit as $t \to \infty$.

Exercise. a) What happens if you replace the pentagon in the preceding example with a square?

b) More generally consider the random walk on the Cayley graph $X(\mathbb{Z}_n, \{\pm 1 (\mathrm{mod}\, n)\})$ in which the random walker at vertex x has equal probability of moving to vertex $x+1 (\mathrm{mod}\, n)$ or to vertex $x-1 (\mathrm{mod}\, n)$. If you want to see convergence to the uniform probability vector $u = {}^T\left(\frac{1}{n}, ..., \frac{1}{n}\right)$, you will need to take n odd or change the random walk to allow the walker to have 3 choices, one being to stay at vertex x.

Now we want to apply similar reasoning to a random walk on an extremely large directed graph. If you websurf to www.google.com and type in some words such as "eigenvalues", you will get a long list of web sites, ordered according to importance. How does Google produce the ordering? Google has to take all over 1 trillion websites and rank them. It seemingly does this once a month. Google is a name close to googol which means 10^{100}. Google was invented by 2 computer science doctoral students (Brin and Page) at Stanford - in the mid-1990s. They only use ideas from a standard undergraduate linear algebra course, plus a bit of elementary probability. They view the web as a directed graph with a web surfer randomly hopping around. The main idea is that the more links a web site has to it, the more important it must be (these links are called "inlinks"). Figure 21 shows a tiny web with only 5 web sites. The sites are the vertices of a directed graph. An arrow from vertex x to vertex y means that vertex x contains a link to vertex y. So in the example of Figure 21 you might think vertex 5 is the most important, since it has the most arrows going to it. In short, if x_k is the number of links to site k, then $x = (1, 2, 1, 2, 3)$.

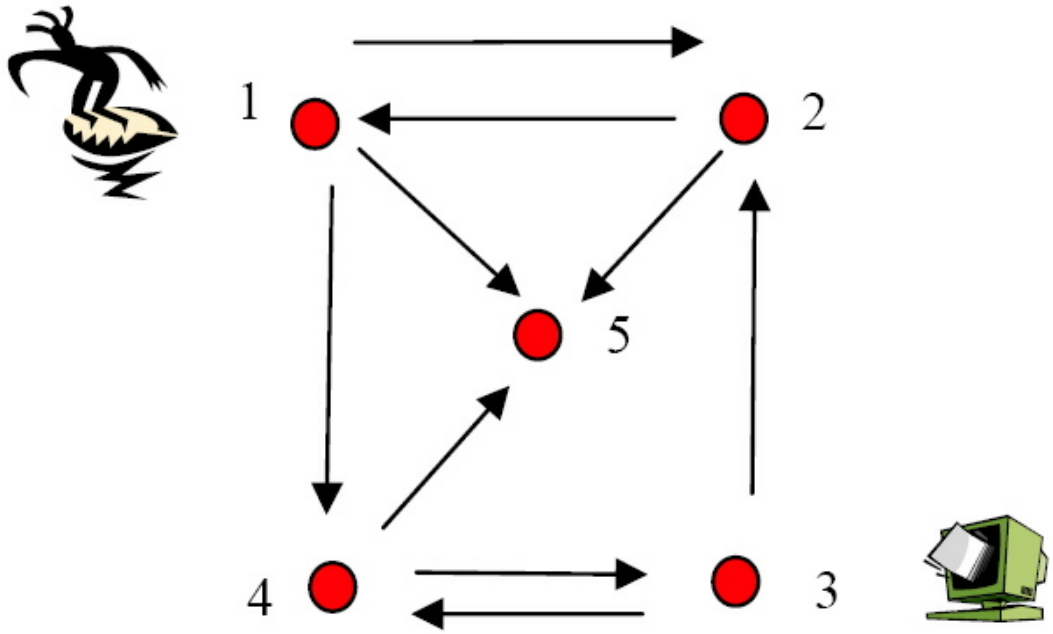

Figure 21: surfing a very small web

On the other hand, node 5 is what is called a "deadend". It has no links to any other site. If we imagine a web surfer bouncing around from web page to web page, that surfer will land at node 5 and have nowhere to go. Many web pages are like this; e.g., pdfs, gifs, jpgs.

We want to make a Markov matrix to give the transition matrix for a random web surfer. Let us first ignore the problem of node 5 and just look at the matrix H whose i, j entry is

$$h_{ij} = \begin{cases} \frac{1}{\#(\text{arrows going out from site } j)}, & \text{if there is an arrow from site } j \text{ to site } i \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

For the webgraph of Figure 21, we get

$$H = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}.$$

This is almost a **Markov matrix** except that the entries of the last column do not sum to 1. For the Google method to work, it would be nice to have an actual positive Markov matrix. For we want it to satisfy the hypotheses of Theorem 65 below that the largest eigenvalue is $\lambda_1 = 1$ and that λ_1 is an eigenvalue with a non-negative probability eigenvector corresponding to the steady state (i.e., limiting) behavior of the associated Markov chain. This theorem says that the other eigenvalues of M satisfy $|\lambda_j| < 1$.

You may want to use Matlab or Mathematica or Scientific Workplace (or whatever) to do the matrix computations in the following exercises.

Exercise. a) Find the largest eigenvalue and a corresponding positive eigenvector for the matrix H above. Since H is not a Markov matrix, don't expect $\lambda_1 = 1$.

b) What is the interpretation of this as far as ranking the websites? Which site is most important?

c) Follow a websurfer who starts at site 1 through 20 iterations. That is, compute $H^k v$, $v = {}^T(1, 0, 0, 0, 0)$, $k = 1, 2, 3 \ldots, 10$.

d) What seems to be happening to the vector $H^k v$ in the limit as $k \to \infty$?

e) Can you use the eigenvalues of H to explain what is happening in d) ?

Exercise. To produce a Markov matrix, one Google idea is to replace the last column in H by a column with 1/5 in each row. Call the new matrix S. Now it comes closer to satisfying the hypothesis of the Theorem 65 below.

a) Write down the matrix S. Does some S^k have all positive entries?

b) Compute a probability eigenvector of S corresponding to the eigenvalue 1. Which site does this eigenvector say is the most important?

c) Follow a websurfer who starts at site 1 through 20 iterations. What is the limit of $S^k v$, $v = {}^T (1, 0, 0, 0, 0)$, as $k \to \infty$? Compare answers in b) and c).

Google has one more trick. The matrix S obtained need not be such that all its entries are positive which is the hypothesis of Perron's Theorem 65 below (although the weaker hypothesis that S^k has all positive entries will also work but is harder to check on a matrix which is 1 trillion $\times$ 1 trillion). The new Google trick will also affect the second largest eigenvalue in absolute value. The new matrix is given, for $0 < \alpha < 1$, by setting

$$G = \alpha S + (1 - \alpha)\frac{1}{n} J, \tag{12}$$

where $n = 5$ and J is an $n \times n$ matrix all of whose entries are 1.

Exercise. a) Write down G in formula (12) for $\alpha = .9$ and then compute a probability eigenvector for G corresponding to the eigenvalue 1. Note that the entries of G are all positive. Which site does this eigenvector say is the most important?

b) Follow a websurfer who starts at site 1 through 20 iterations. What is the limit of $G^k v$, $v = {}^T(1, 0, 0, 0, 0)$, as $k \to \infty$? Compare answers in a) and b).

Notes. In the formula for G, Google chooses $\alpha = .85$. It could be any number between 0 and 1. If $\alpha = .85$, it means that 85% of the time the web surfer follows the hyperlink structure of the web and the other 15% of the time the web surfer jumps (teleports) to a random web page. Since 1/(1 trillion) is small, the alteration in the entries of the matrix H is not enormous. Of course the Google version of this matrix will be 1trillion x 1 trillion. How does Google find the dominant eigenvector of

G? It uses a very old method called the **power method** which works well for sparse matrices (meaning matrices most of whose entries are 0). That is, Google uses the fact that we noticed in the preceding problems. If you just keep multiplying some fixed probability vector v by G; in effect, computing $G^k v$, this should converge to the probability eigenvector for the eigenvalue 1.

In the case that the Markov matrix M is symmetric, the power method basically takes advantage of formula (10) which says that for arbitrary probability vectors v, the vectors $M^n v$ approach $u = {}^T(\frac{1}{n}, \cdots, \frac{1}{n})$, the steady-state of the Markov chain as $n \to \infty$. For a non-symmetric positive Markov matrix M, an analogous result comes from Theorem 65 below. But in the case of a non-symmetric Markov matrix the stationary state vector will not have all entries equal and that will of course give the web site rankings. The power method was published by R. von Mises in 1929.

Exercise. Suppose that the matrix $A \in \mathbb{R}^{n\times n}$ has n linearly independent eigenvectors $v_j \in \mathbb{R}^n$ with $Av_j = \lambda_j v_j$. Suppose that $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Show that for any vector $w \in \mathbb{R}^n$,

$$\lim_{k\to\infty} \frac{1}{|\lambda_1|^k} A^k w = \alpha v_1,$$

for some scalar $\alpha \in \mathbb{R}$.

Hint. Write $w = \sum_{j=1}^{n} \gamma_j v_j$, for $\gamma_j \in \mathbb{R}$. Apply A^k to both sides and take the limit.

However the power method is "notoriously slow" for non-sparse matrices like G. So why does it work for Google? The first part of the answer has to do with the fact that is proved in the following problem. Google only needs to compute the iterates of sparse matrices and not G itself.

The second part of the answer says that this method requires only about 50 iterations for the huge matrix Google is dealing with? Why should this be? This has to do with the size of the 2nd largest eigenvalue λ_2 of G in absolute value. It turns out that for the Google matrix, $|\lambda_2| \cong \alpha$.

Choosing $\alpha = .85$, one finds that $.85^{50} \cong .000296$. This means that Google can expect about 2-3 places of accuracy in the Page Rank vector after about 50 iterations of replacing v by Gv.

Exercise. Here we are trying to understand part of the second to last paragraph. Consider a web with n sites. Let H be the matrix whose i, j entry is defined by formula (11). Write e=the column vector of 1's. Let b be the column vector whose jth component is

$$b_j = \begin{cases} 1, & \text{if site } j \text{ has no arrows going out (i.e., it's a dead end),} \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

Using formulas (11) and (13), define

$$S = H + \frac{1}{n} e\ {}^T b, \quad \text{and} \quad G = \alpha S + (1-\alpha)\frac{1}{n} e\ {}^T e, \quad \text{for } 0 < \alpha < 1. \tag{14}$$

Show that if the Google matrix G is defined by formula (12) and if v is any probability (column) vector, meaning its entries are ≥ 0 and sum to 1 (which implies ${}^T e\ v = 1$), we have

$$Gv = \alpha H v + \frac{1}{n}(\alpha\ {}^T b\ v + 1 - \alpha)e. \tag{15}$$

Note that H is sparse (with on average only about 10 non-zero elements in a column) and the scalar ${}^T b\ v$ is easy to compute. It follows that iterating Gv will be quickly computed.

The next exercise is an attempt to explain "Google bombing." To do this, people are paid to set up link farms to fool Google into thinking a web page is more important than it otherwise would appear to be. Google attempts to find such occurrences and then give such pages lower ranks. It was sued for doing so in 2002. The lawsuit was dismissed in 2003. See the book of Langville and Meyer for more information. Now Google claims to be using many (200) factors to rank sites - not just the pagerank.

Exercise. In the example of the small web in Figure 21, suppose the site 1 people are angry to be rated below site 5. To increase the rating of site 1, they create a new site 6 with 3 links to site 1. Site 1 will also link to site 6. Does this help site 1's ranking?

a) Find the new H matrix from formula (11). Then form the S matrix in formula (14). Finally form the G matrix as in formula (15).

b) Then you need to find the probability eigenvector of the G matrix corresponding to the eigenvalue 1.

c) Would it help if site 1 created another new site with links to site 1?
d) What can Google do to minimize the effect of this sort of thing?

The Perron theorem was proved by Perron in 1907 and later generalized by Frobenius in 1912. The general version is called the Perron-Frobenius theorem. We given only a special case. To see the general version, look at Horn and Johnson, *Matrix Analysis.*

Theorem 65 ***(Perron).*** *Suppose that the $n \times n$ Markov matrix M has all positive entries. Then*
a) 1 is an eigenvalue of M and the corresponding vector space of eigenvectors is $1-$dimensional.
b) If the eigenvalues of M are listed as $\lambda_1 = 1, \lambda_2, ..., \lambda_n$, then $1 = |\lambda_1| > |\lambda_j|$, for all $j = 2, ..., n$.
c) There is an eigenvector v_1 corresponding to the eigenvalue 1 which is such that all its entries are > 0 and they sum to 1.
d) The vector v_1 is the steady state of the Markov chain with transition matrix M; i.e.,

$$\lim_{r \to \infty} M^r x = v_1, \quad \textit{for any probability vector} \quad x.$$

It is easy to see that if M is a Markov matrix, 1 is an eigenvalue of the transpose, ${}^TM$, with eigenvector ${}^T(1, 1, ..., 1)$. That's because the columns of M sum to 1. The eigenvalues of ${}^TM$ are the same as the eigenvalues of M, since the determinant of a matrix is the same as the determinant of its transpose. For the rest of the proof, see C. D. Meyer, *Matrix Algebra and Applied Linear Algebra, Chapter 8,* or the last chapter of R. A. Horn and C. R. Johnson, *Matrix Analysis.*
Exercise. Prove the Perron theorem in the case that the positive Markov matrix M is 2×2.
Exercise. a) If a real matrix A has non-negative entries, write $A \geq 0$. If A has positive entries write $A > 0$. Prove that if $A > 0$ and $x \geq 0, x \neq 0$, then $Ax > 0$.
b) If $A - B \geq 0$, write $A \geq B$. Show that if $N \geq 0$ and $u \geq v \geq 0$, then $Nu \geq Nv \geq 0$.
c) What are the properties of this inequality on matrices? Does it satisfy the properties listed in our list of facts about orders on $\mathbb{Z}$ in Section 3 of Part I of these lectures?

Define the **spectral radius** of $A \in \mathbb{C}^{n \times n}$ to be $\rho(A) = \max\{|\lambda| \mid \lambda \in spec(A)\}$. The following exercise is worked out in the book of Meyer, for example.
Exercise. a) Show that if $A \in \mathbb{R}^{n \times n}$ and $A > 0$, then the spectral radius $\rho(A)$ is positive.
b) Show that under the same hypotheses as in part a), $\rho(A) \in spec(A)$ and there is a positive eigenvector corresponding to the eigenvalue $\rho(A)$. In this case we call $\rho(A)$ the **Perron eigenvalue**.

Many proofs have been given of the Perron-Frobenius theorem. One uses the Brouwer fixed point theorem from analysis. Others come from H. Wielandt who shows that the Perron eigenvalue $\rho(A)$ of a matrix $A \in \mathbb{R}^{n \times n}$ such that $A > 0$ can be expressed as

$$\rho(A) = \max_{x \in \mathbb{R}^n, x \geq 0, x \neq 0} \left\{ \min_{\substack{1 \leq i \leq n \\ x_i \neq 0}} \frac{(Ax)_i}{x_i} \right\}.$$

You might still ask how Google finds the webpages with the words you typed. Google answers on its website that it has a large number of computers to "crawl" the web and "fetch" the pages and then form a humongous index of all the words it sees. So when we type in "eigenvalue" Google's computers search their index for that word and the pagerank of the websites containing that word among "200 factors."

16 Elliptic Curve Cryptography

See Ramanujachary Kumanduri and Cristina Romero, *Number Theory with Computer Applications*, Chapter 19, on elliptic curves. Another reference which also does cryptography is Neal Koblitz, *A Course in Number Theory and Cryptography.* Google.com gave us many hits when we typed in "elliptic curve cryptography."

What is an elliptic curve? First it is not an ellipse. According to the website of Jeff Miller which gives the origins of mathematical terms, the name "elliptic curve" comes from a poem to Isaac Newton by James Thompson. Here is a quote from the poem.

"He, first of Men, with awful Wing pursu'd The Comet thro' the long Elliptic Curve."

Let K be any field, for example $K = \mathbb{R}$, $\mathbb{C}$, or $\mathbb{Q}$ or $\mathbb{Z}/p\mathbb{Z}$, p=prime. We will mostly be interested in finite fields here.

Definition 66 *An **elliptic curve** $E = E(K)$ is the set of points (x, y) with x, y in K such that $y^2 = x^3 + ax^2 + bx + c$.*

We omit some technical conditions, which we will soon be forced to consider. You can also replace the y^2 on the left with some other quadratic function of y.

The real points on $E(\mathbb{R})$ are of interest. They will help us to visualize what we are doing over finite fields. So let's use draw some pictures of possible elliptic curves in the plane. Figure 22 shows the real points (x, y) on the elliptic curve $y^2 = x^3 + x^2$. Figure 23 shows the real points (x, y) on the elliptic curve $y^2 = x^3 + 2$. Figure 24 plots the real points (x, y) on the elliptic curve $y^2 = x^3 - x$.

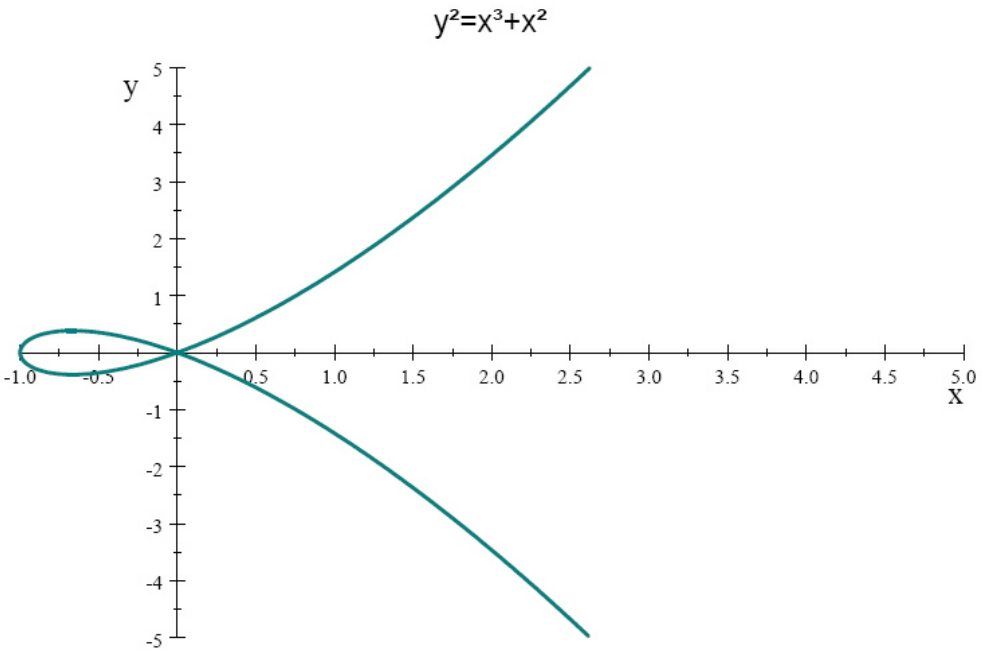

Figure 22: The real points (x, y) on the elliptic curve $y^2 = x^3 + x^2$.

Exercise. Plot the real points (x, y) on the elliptic curve $y^2 = x^3 + x$ and any other curves you find interesting.

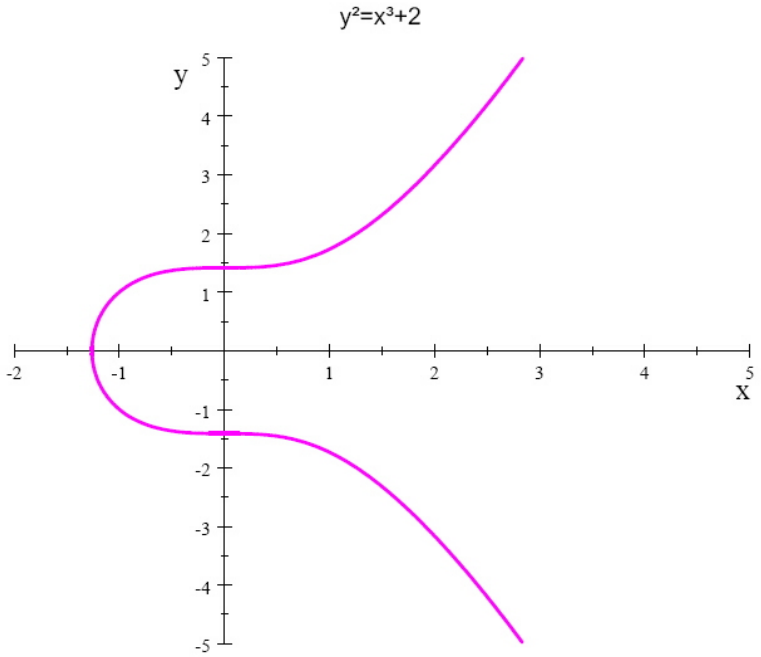

Figure 23: real points (x, y) on the elliptic curve $y^2 = x^3 + 2$.

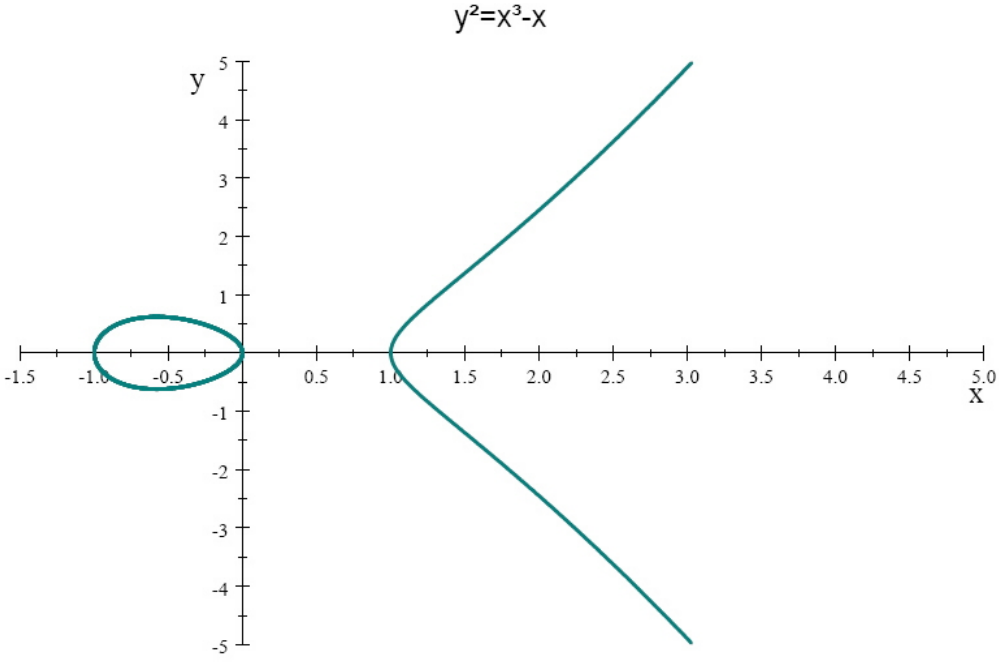

Figure 24: The elliptic curve $y^2 = x^3 - x$ over the reals.

It is useful to replace the plane with the projective plane. In general if K is some field, projective n-space is obtained by looking at points $x = (x_0, x_1, x_2, ..., x_n) \in K^{n+1}$ with $x \neq 0$, and setting up an equivalence relation $x \sim t$ iff $x = \alpha t$, for some $\alpha \in K$. Here we mean the usual multiplication of a vector by a scalar.

Projective n-space over K is the set of equivalence classes of $K^{n+1} - 0$ under this equivalence relation; i.e., $\mathbb{P}_n(K) = (K^{n+1} - 0)/\sim$. This allows us to replace our elliptic curve $E(K)$ with a curve in the projective plane $\mathbb{P}_2(K)$:

$$y^2 = x^3 + ax^2 + bx + c \quad \text{becomes} \quad (y/z)^2 = (x/z)^3 + a(x/z)^2 + bx/z + c \quad \text{or} \quad y^2 z = x^3 + ax^2 z + bxz^2 + cz^3. \tag{16}$$

We will identify $(x, y, 1)$ in $\mathbb{P}_2(K)$ with (x, y) in K^2. So we view K^2 as a subspace of the projective plane called **affine space**. The **line at infinity** in $\mathbb{P}_2(K)$ consists of the equivalence classes of points $(x, y, 0)$ in $\mathbb{P}_2(K)$. The intersection of this line with the elliptic curve of formula (16) has $x = 0$. Then the equivalence class in $\mathbb{P}_2(K)$ containing the point $(0, 1, 0)$ is called the **point at infinity**. View it as a point on the intersection of the $y-$axis and the line at infinity in $\mathbb{P}_2(K)$.

Definition 67 *Suppose $a, b, c \in \mathbb{C}$. If $f(x) = x^3 + ax^2 + bx + c = (x - r_1)(x - r_2)(x - r_3)$, then the **discriminant** of f is $\Delta f = (r_1 - r_2)^2(r_2 - r_3)^2(r_1 - r_3)^2$. One can show that $\Delta f = a^2b^2 + 18abc - 27c^2 - 4a^3c - 4b^3$.*

An extra factors of 16 appears in the discriminant of an elliptic curve. See, for example N. Koblitz, *Introduction to Elliptic Curves and Modular Forms*, p. 26 or the extensive tables of J. E. Cremona, *Algorithms for Modular Elliptic Curves* (www.warwick.ac.uk/staff/J.E.Cremona/book/fulltext/index.html or the new website http://l-functions.org).

In order to create a group associated with our elliptic curve, we need to be able to draw tangents to our elliptic curves. The tangent lines to the curve $y = f(x)$ will be undefined at points r where both $f(r)$ and $f'(x)$ vanish. That is, when r is a double root of f and thus the discriminant vanishes. An elliptic curve for which the discriminant is non-zero is called **nonsingular**. For example, the curve $y^2 = x^3 + x^2$ in Figure 22 does not have a well defined tangent at the origin and its discriminant is 0.

What is the group of an elliptic curve? To associate an abelian group G to an elliptic curve $E(K)$ where K is any field, usually $\mathbb{Q}$ or $\mathbb{F}_q$, the simplest way is to say that 3 points p, q, r on $E(K)$ add to 0 iff they lie on a straight line. See Figure 25. We define the identity 0 to be the point at infinity on the curve. Then if $p = (x, y)$, we see that $-p = (x, -y)$. Think of 0 as a point infinitely far up any vertical line. If you need to compute $2p = p + p$, then define the intersection of the curve and its tangent at p to be $-2p$. Of course, this makes sense over $\mathbb{R}$. To figure out what is happening over a finite field, we just use the formulas derived from those over the real field. We will make this more precise in the examples below. The curve will have to have a well defined tangent at every point for our construction to work. Thus the curve in Figure 22 is a bad one, since there is no well defined tangent at the origin.

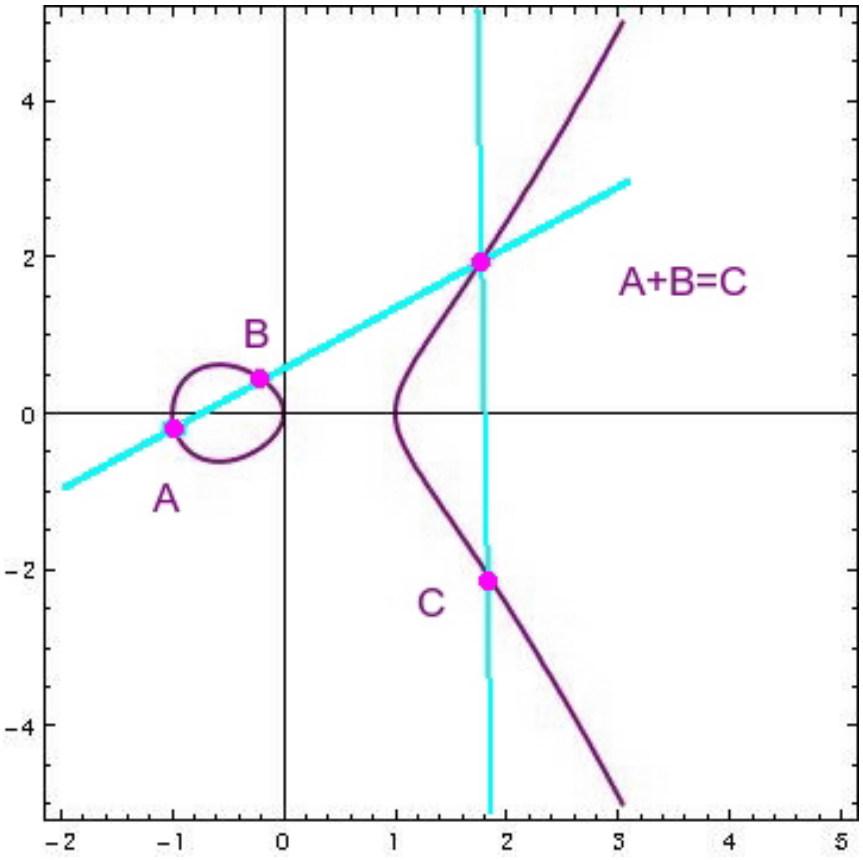

Figure 25: picture of the addition $A + B = C$ on the elliptic curve $y^2 = x^3 - x$ where the field is $\mathbb{R}$.

Theorem 68 *The preceding definition makes the nonsingular elliptic curve into an abelian group.*

A proof sketch is given in Kumanduri and Romero, *Number Theory with Computer Applications*, p. 496.

Example. Look at $y^2 + y = x^3 - x^2$ over the field $\mathbb{Q}$. This curve has 5 rational points $a = (0,0), b = (1,-1), c = (1,0), d = (0,-1)$ and 0. Can you prove it? The real points are shown in Figure 26.

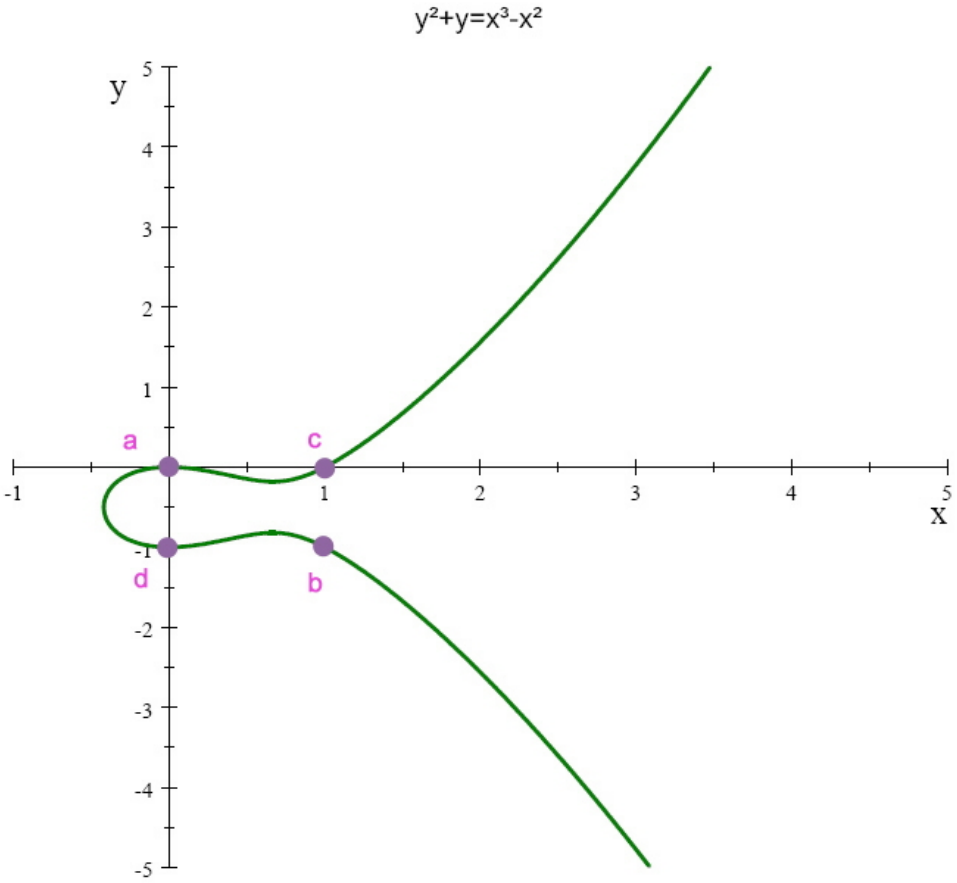

Figure 26: The rational points on the curve $y^2 + y = x^3 - x^2$ are a, b, c, d and the point at ∞.

The group G of this curve over $\mathbb{Q}$ turns out to be the cyclic group of order 5 generated by a. For you can see that $2a = b$. In this case you just need to see that the tangent to the curve at a (which is the x-axis) intersects the curve at c. So that means $a + a + c = 0$ and thus $a + a = -c = b$. And $b + c = 0 =$ the point at ∞, since the line through b and c is vertical and thus goes through the point at ∞.

Exercise. Compute the addition table for the group $G = \{0, a, b, c, d\}$ of the curve $y^2 + y = x^3 - x^2$. The group G is a cyclic group generated by a. Assume that we have found all the rational points on the curve.

Theorem 69 *(Mordell). The group G associated to an elliptic curve $E(\mathbb{Q})$ is finitely generated (not necessarily finite) abelian.*

Thus, by the fundamental theorem of abelian groups, the group G of $E(\mathbb{Q})$ is isomorphic to a direct sum $\mathbb{Z}_{a_1} \oplus \cdots \oplus \mathbb{Z}_{a_n} \oplus \mathbb{Z}^r$. Here r is called the **rank** of G. The finite part of G is called the **torsion subgroup**. Kumanduri and Romero, *Number Theory with Computer Applications,* finds the torsion subgroup in Section 19.5. The rank is harder.

The rank of an elliptic curve is connected to the congruent number problem which is still open. The congruent number problem asks which positive integers n are such that there a right triangle with rational sides whose area equals n. More precisely, the following 2 questions are equivalent:
Question A. For every $n \in \mathbb{Z}^+$ does there exist a right triangle with rational sides whose area equals n?
Question B. Is the rank of $y^2 = x^3 - n^2 x$ positive?

Remarks on Elliptic Curves over the Field $\mathbb{C}$.

To study elliptic curves over the complex numbers, one needs the theory of the Weierstrass $\wp$-function. The function $\wp(z)$ is a holomorphic function of z in the complex plane except for a double pole at each point of a lattice $L = w_1\mathbb{Z} + w_2\mathbb{Z}$. Moreover one has the differential equation:

$$\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3.$$

This implies that the point $(\wp(z), \wp'(z))$ lies on an elliptic curve $E(C)$. The "curve" is a subset of $\mathbb{C}^2 \cong \mathbb{R}^4$ which has 1 complex parameter and thus 2 real parameters. The graph would have to be drawn in 4 real dimensions. If you need to know, the numbers g_2, g_3 are given by (Eisenstein) series:

$$g_2 = 60 \sum_{\substack{m,n\in\mathbb{Z}\\(m,n)\neq(0,0)}} \frac{1}{(mw_1+nw_2)^4} \quad \text{and} \quad g_3 = 140 \sum_{\substack{m,n\in\mathbb{Z}\\(m,n)\neq(0,0)}} \frac{1}{(mw_1+nw_2)^6}.$$

Mathematica knows $\wp(z)$ as `WeierstrassP[z,{g1,g2}]`. This gives a one-to-one correspondence from z to $(\wp(z), \wp'(z))$ which takes the torus $\mathbb{C}/L$ (where we identify points in $\mathbb{C}$ which differ by a lattice point in $L = w_1\mathbb{Z}+w_2\mathbb{Z}$) to the elliptic curve $E(\mathbb{C})$. The mapping is an isomorphism of abelian groups. The Weierstrass $\wp$-function is not usually covered in undergraduate analysis. However you can find a discussion in Koblitz, *Introduction to Elliptic Curves and Modular Forms.*

Elliptic Curves over the Field $\mathbb{F}_q$

In order to do cryptography, we need to discuss elliptic curves over finite fields $\mathbb{F}_q$. Mostly let's consider the special case that $q = p$ =prime. How do we add points on a curve $E = E(\mathbb{Z}_p)$

$$y^2 = x^3 + ax^2 + bx + c (\mathrm{mod}\, p)? \tag{17}$$

Here we assume the prime $p > 3$, $a, b, c \in \mathbb{Z}$, and p does not divide the discriminant $a^2b^2 + 18abc - 27c^2 - 4a^3c - 4b^3$. We imitate the construction over $\mathbb{R}$. Now the "curve" is just a finite set of points.

See Figure 27 for an example mod 29. The purple points on the 29×29 grid correspond to (x, y) such that $y^2 = x^3 - x + 1 \pmod{29}$ and they look pretty random, though the pink line segments do look like lines.

In general, let $P = (x_1, y_1)$ and $Q = (x_2, y_2)$, $P + Q = (x_3, y_3)$, with $x_1 \neq x_2$. Let $y - y_1 = \mu(x - x_1)$ be the line L through P and Q. Points y on the "line" L must satisfy

$$y = \mu x + \beta \ (\mathrm{mod}\, p), \quad \text{where} \quad \mu = \frac{y_2 - y_1}{x_2 - x_1} (\mathrm{mod}\, p) \quad \text{and} \quad \beta = y_1 - \mu x_1 (\mathrm{mod}\, p). \tag{18}$$

In equation (18), you have to find the inverse $(\mathrm{mod}\, p)$ to find the "slope" μ of the line L.

To find the 3rd point on L and E, which is $-(P+Q)$, we plug equation (18) into formula (17) for the elliptic curve $E(\mathbb{Z}_p)$. You get a cubic equation for x:

$$(\mu x + \beta)^2 = x^3 + ax^2 + bx + c \ (\mathrm{mod}\, p).$$

So we can find our point $-(P + Q)$ by solving the cubic:

$$f(x) = x^3 + (a - \mu^2)x^2 + (b - 2\mu\beta)\, x + c - \beta^2 (\mathrm{mod}\, p). \tag{19}$$

We already have 2 roots of $f(x)$, namely x_1 and x_2. This means that

$$f(x) = (x - x_1)(x - x_2)(x - x_3) = x^3 - (x_1 + x_2 + x_3)x^2 + (x_1x_2 + x_1x_3 + x_2x_3)x - x_1x_2x_3 \pmod{p}. \tag{20}$$

Therefore if $x_1 \neq x_2$, $P = (x_1, y_1)$ and $Q = (x_2, y_2)$, we have $P + Q = (x_3, -y_3)$, with

$$x_3 \equiv \mu^2 - a - x_1 - x_2 \pmod{p}, \quad y_3 \equiv \mu(x_3 - x_1) + y_1 \pmod{p}, \quad \text{where} \quad \mu = \frac{y_2 - y_1}{x_2 - x_1} \pmod{p}. \tag{21}$$

Again in the formula for μ, you must divide mod p. When $x_1 = x_2$ but $P \neq Q$, the sum is 0, the point at infinity.

To figure out the rule when $P = Q$, look at the tangent to E at P. This is found by recalling that the "derivative" is the "slope" to the "tangent" and thus formally we have

$$2y\frac{dy}{dx} = 3x^2 + 2ax + b \pmod{p}.$$

It follows that the slope $\frac{dy}{dx}$ of the "tangent" to E at the point (x_1, y_1) is

$$\mu = \frac{3x_1^2 + 2ax_1 + b}{2y_1} \pmod{p}.$$

If $y_1 = 0$, the "tangent" is vertical and the third point on the curve is 0, the point at infinity. Once more, to find the sum $P + Q$, substitute the equation $y = \mu x + \beta$ into equation (17) for the elliptic curve. This time we have a double root at x_1. So instead of (20) we see that

$$f(x) = (x - x_1)^2(x - x_3) = x^3 - (2x_1 + x_3)x^2 + (x_1^2 + 2x_1x_3)x - x_1^2x_3 \pmod{p}. \tag{22}$$

This implies $2P = P + P = (x_3, -y_3)$ where

$$x_3 \equiv \mu^2 - a - 2x_1 \pmod{p}, \quad y_3 \equiv \mu(x_3 - x_1) + y_1 \pmod{p}, \quad \mu = \frac{3x_1^2 + 2ax_1 + b}{2y_1} \pmod{p}. \tag{23}$$

Note that equation (23) is just (21) with $x_1 = x_2$, except for finding μ with formal derivatives.

Example. Consider the elliptic curve in Figure 27. We plot the points as purple squares in a grid of $(x, y) \in \mathbb{F}_{29}^2$. Point A is $(6, 22)$. Point B is $(10, 24)$. The "line" through them is $y = \mu x + \beta$, where we find $\mu = 15 = 2^{-1} \pmod{29}$ and $\beta = 19 \pmod{29}$. Then we find $-C = (17, 28)$. It follows that $C = (17, 1) = A + B$. Of course the lines mod p are not always so easily seen. But mercifully all we really need is formula (18). Again, the 3rd point on the line through C and $-C$ is the point at ∞ which is viewed as infinitely far up the vertical line through C and $-C$.

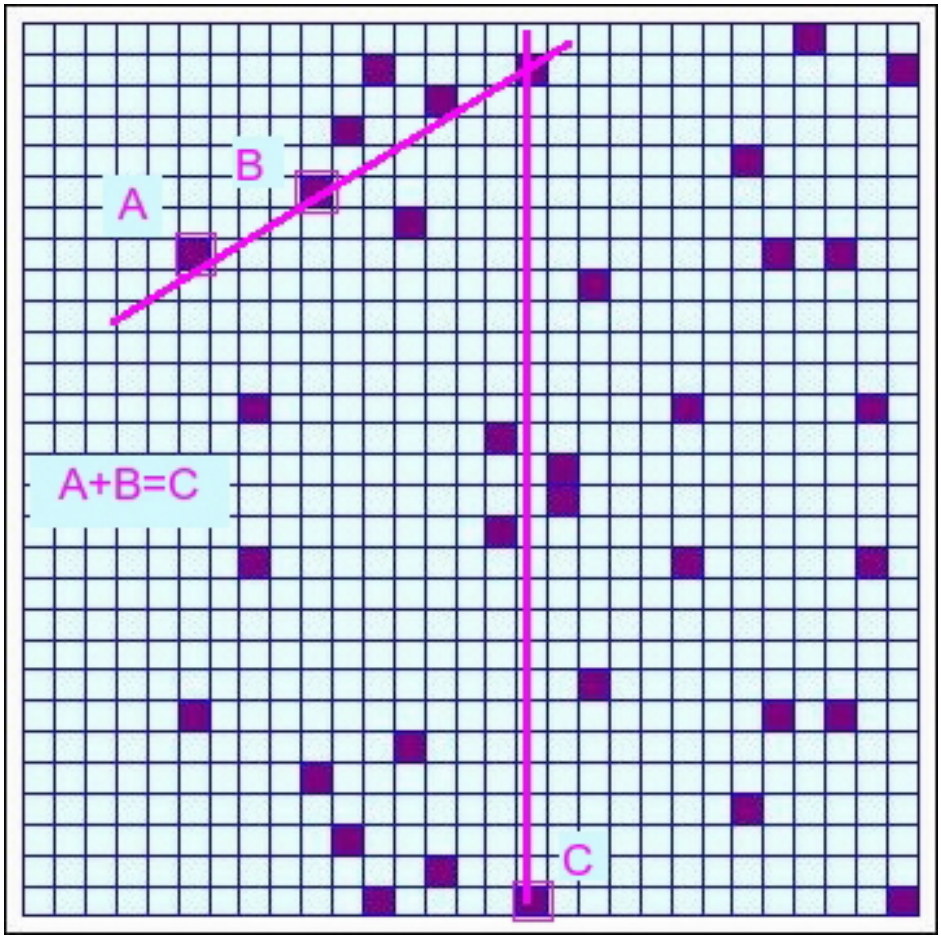

Figure 27: The purple squares indicate the points (x, y) on the elliptic curve $y^2 = x^3 - x + 1 \bmod 29$. The line through the points $A = (6, 22), B = (10, 24), -C = (17, 28)$ is indicated. Then $A + B = C = (17, 1)$.

Exercise. Compute the discriminant of the elliptic curve in Figure 27.
Exercise. Visually try to find 3 more points in a line in Figure 27. Then compute the sum of any 2 of the points using formula (18).
Example. Compute the group table for the elliptic curve E given by $y^2 = x^3 + 1 \pmod 5$.

It is now easy to find the points. Substitute $x = 0, 1, 2, 3, 4$ and solve for $y \pmod 5$. You find

$$a = (0,1), \quad b = (0,-1), \quad c = (2,3), \quad d = (2,-3), \quad e = (4,0) \text{ and } \quad 0 = \text{ point at } \infty .$$

We can use formulas (21) and (23) to compute the group table for the group G of points on E. First note that $a = (0,1)$ and $b = (0,-1)$. Thus $a + b = 0$=the point at infinity.

To find $a+c$, note that $a = (0,1)$ and $c = (2,3)$. So, in this case, the slope of the line L through a and c is $\mu \equiv \frac{3-1}{2-0} (\mod 5)$. So $\mu \equiv 1 \pmod 5$. Then, using (21), we see that

$$\begin{aligned} x_3 &\equiv \mu^2 - a - x_1 - x_2 \equiv 1 - 0 - 0 - 2 \equiv -1 \equiv 4 (\mod 5). \\ y_3 &\equiv \mu(x_3 - x_1) + y_1 \equiv 1(4-0) + 1 \equiv 5 \equiv 0 (\mod 5). \end{aligned}$$

Thus $a + c = (4,0) = e$.

To find $c + c$, use equation (23) and note that in this case $\mu = \frac{3*4}{2*3} \equiv 3 * 4 \equiv 2 (\mod 5)$.

$$\begin{aligned} x_3 &\equiv \mu^2 - a - 2x_1 \equiv 4 - 0 - 4 \equiv 0 (\mod 5). \\ y_3 &\equiv \mu(x_3 - x_1) + y_1 \equiv 2(0-2) + 3 \equiv -1 (\mod 5). \end{aligned}$$

It follows that $2c = a$.
Exercise. Compute the rest of the group table for the group G of the preceding example. Is G cyclic?

Remarks on the number of points on an Elliptic Curve Mod p.

Define the **Legendre symbol** by

$$\left(\frac{n}{p}\right) = \begin{cases} 0, & \text{if } p \text{ divides } n, \\ 1, & \text{if } p \text{ does not divide } m \text{ and } n \equiv x^2 (\mod p) \text{ has a solution } x. \end{cases}$$

Exercise. a) Let $f(x) = x^3 + ax^2 + bx + c$. Show that

$$1 + \left(\frac{f(x)}{p}\right) = \text{ the number of solutions } y \text{ to the congruence } \quad y^2 \equiv x^3 + ax^2 + bx + c (\mod p).$$

b) Then prove that the number of points on the elliptic curve $y^2 \equiv x^3 + ax^2 + bx + c (\mod p)$ is

$$N_p = p + 1 + \sum_{x=0}^{p-1} \left(\frac{f(x)}{p}\right).$$

c) For our example from the preceding problem, use this formula to see that there are 6 points on the curve.

Theorem 70 ***(H. Hasse, 1933).*** *If N_p is the number of points on an elliptic curve $\mod p$, set $a_p = p + 1 - N_p$. Then $|a_p| \leq 2\sqrt{p}$.*

To prove this theorem, one must bound the sum $\sum_{x=0}^{p-1} \left(\frac{f(x)}{p}\right)$. One expects the Legendre symbols to be randomly +1 or -1. That leads to the heuristic reason for the bound. If you want a real proof, see S. Lang, *Elliptic Curves: Diophantine Analysis.*

Much is known about elliptic curves. For example, it has also been proved that the group G of points on an elliptic curve $(\mod p)$ is a product of at most two cyclic groups. More references on the subject are: Jeff Hoffstein, Jill Pipher and J. H. Silverman, *Introduction to mathematical cryptography*; Kristin Lauter, The advantages of elliptic curve cryptography for wireless security, *IEEE Wireless Communications*, February 2004, 2-7; Karl Rubin and Alice Silverberg, Ranks of elliptic

curves, *Bull. Amer. Math. Soc.*, 39 (2002), 455-474; Alice Silverberg, Introduction to Elliptic Curves, in *IAS/Park City Math. Series*, to appear; J. Silverman, *The Arithmetic of Elliptic Curves*.
Exercise. Find the number of points on the curve $y^2 = x^3 - 1 (\mathrm{mod}\, p)$ for all odd primes $p \leq 30$.

Elliptic Curves over Finite Fields and Cryptography.

In Section 21 of Part I we saw how to get public key secret codes from the multiplicative group $(\mathbb{Z}/pq\mathbb{Z})^*$, when p and q are large primes. The usefulness of such codes derives from the difficulty of factoring pq when p and q are 2 large primes. Elliptic curve cryptography makes use of the group G of an elliptic curve $\mathrm{mod}\, p$. It seems that one can use smaller public keys and still have secure messages using elliptic curve cryptography.

Note that our analog of raising an element a in $(\mathbb{Z}/pq\mathbb{Z})^*$ to the kth power in G is multiplying an element a in G by k, or adding a to itself k times to get

$$\underset{k \text{ times}}{a + \cdots + a} = ka.$$

To do this fast, one can proceed in an analogous way to that with powers. For example,

$$100 = 2^6 + 2^5 + 2^2$$

The result is that in order to compute $100a$, we need 6 doublings and 2 additions.

We will want to encode our plaintext m as a point P_m on an elliptic curve E so that it will be easy to get m from P_m.
Remarks.

1) There does not exist an algorithm for writing down lots of points on $E(\mathbb{Z}/p\mathbb{Z})$ in $\log p$ time.

2) It is not sufficient to generate random points on $E(\mathbb{Z}/p\mathbb{Z})$ anyway.

A Probabilistic Method to Encode Plaintext m as P_m on an elliptic curve $E(\mathbb{Z}/p\mathbb{Z})$.

We will illustrate the method with an example from Koblitz, *A Course in Number Theory and Cryptography*. The curve is

$$y^2 + y \equiv x^3 - x (\mathrm{mod}\, 751).$$

This curve has 727 points.
Exercise. Check the last statement.

Take a number $\kappa = 20$ (or larger). The number κ is chosen so that a failure rate of $\frac{1}{2^\kappa}$ is OK when seeking our point. We will need to represent numbers m between 0 and 35 (meaning the usual alphabet plus the digits from 0 to 9):

$$0, 1, 2, ..., 9, A, B, C, D, ..., X, Y, Z.$$

So we want $p > M * \kappa = 700$. Our $p = 751$ so that is O.K.

Write x between 0 and 700 in the form $x = m * 20 + j$, where $1 \leq j \leq 20$. Then compute $y' = y + 376$ so that

$$2 * 376 \equiv 1 (\mathrm{mod}\, 751) \quad \text{and} \quad 376^2 \equiv 188 (\mathrm{mod}\, 751).$$

$$y'^2 \equiv (y + 376)^2 \equiv y^2 + y + 188 \equiv x^3 - x + 188 (\mathrm{mod}\, 751).$$

Thus we need a fast way to do square roots $\mathrm{mod}\, p$. Luckily Mathematica does square roots $\mathrm{mod}\, p$. So we do not need to program this algorithm (unless we hate Mathematica). Of course programs like SAGE actually know about elliptic curves.

If we can solve for y then set $P_x = (x, y)$. Otherwise replace j by $j + 1$ in the formula for x and try again. Since our curve has 727 points, probability says we shouldn't have to increment more than 20 times. Set $f(x) \equiv x^3 - x + 188 (\mathrm{mod}\, 751)$. There is a $(\frac{1}{2})^{20}$ chance that $f(m*20+j)$ will not be a square for any $j = 1, 2, ..., 20$; assuming that the events $f(m*20+j)$ =square and $f(m * 20 + j + 1)$ =square are independent.

Here's our alphabet table.

0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	G	I
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18
J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	
19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	

Let's find some points on the curve corresponding to our alphabet entries.

1) First $m = 0$ gives P_0. We look at $x = 0 * 20 + 1$ and plug that into

$$y'^2 \equiv (y + 376)^2 \equiv x^3 - x + 188 \equiv 1 - 1 + 188 \equiv 188 \equiv (376)^2 (\mathrm{mod}\, 751).$$

Clearly the solution is $y = 0$. So $P_0 = (1, 0)$.

2) Next we look at $m = 1$, and form $x = 1 * 20 + j$, with $j = 1, ..., 20$. For $j = 1$, we have $x = 21$ and solve

$$y'^2 \equiv (y + 376)^2 \equiv x^3 - x + 188 \equiv 21^3 - 21x + 188 \equiv 416 \equiv (618)^2 (\mathrm{mod}\, 751).$$

Then $y'^2 \equiv (618)^2 (\mathrm{mod}\, 751)$ has two solutions. We take $y' \equiv 618 (\mathrm{mod}\, 751)$. Then $y \equiv y' - 376 \equiv 212 (\mathrm{mod}\, 751)$. The other solution is $y \equiv 508 (\mathrm{mod}\, 751)$. We'll ignore it. So we get the point $P_1 = (21, 242)$. We could have equally well said $(21, 508)$.

3) Similarly we find $P_2 = (41, 101)$.

4) The next case is more interesting. If we set $m = 3$ and form $x = 3 * 20 + j$, with $j = 1$, we see that when $x = 61$, we cannot solve the congruence

$$y'^2 \equiv (y + 376)^2 \equiv x^3 - x + 188 \equiv 61^3 - 61x + 188 \equiv 306 (\mathrm{mod}\, 751).$$

So we must increment j to 2 and look at $x = 62$. Luckily this guy is a square mod 751 and we find that $P_2 = (62, 214)$.

5) Corresponding to the letter S is the number $m = 28$ (from our alphabet table). Then we find the point $P_{28} = (562, 576)$ on the curve E. It again takes 2 tries.

Of course we are using Mathematica to do this. Here is part of our Mathematica notebook.

In versions of Mathematica from 2001, we needed to include the package `<<NumberTheory‘NumberTheoryFunctions‘` in order to take square roots mod n. Now we just use `PowerMod[a,1/2,n]`.

For example, to do the Mathematica calculation for the 2nd point on the curve mod 751, we define the following Mathematica functions f,g,h,k and perform the calculations.

```
f[x_]:=f[x]=Mod[PowerMod[x,3,751] - x + 188,751]
g[x_]:=g[x]=PowerMod[x,1/2,751]
h[x_]:=h[x]=Mod[x-376,751]
k[x_]:=k[x]=Mod[-x-376,751]
f[21]=416
g[416] =618
h[618]= 242
k[618]=508
```

Exercise. a) Write the message

THEIR LANGUAGE IS THE LANGUAGE OF NUMBERS AND THEY HAVE NO NEED TO SMILE

as a sequence of points on the curve

$$y^2 + y \equiv x^3 - x (\mathrm{mod}\, 751).$$

using the method described above. This is a quote from the Dr. Who episode *Logopolis*.

b) Translate the following sequence of points on the curve in part a)

$$(421, 737)(361, 383)(621, 220)(283, 321)(421, 737)(484, 214)(461, 467)(324, 416)(201, 380)$$
$$(461, 467)(261, 663)(501, 220)(543, 436)(484, 214)(562, 576)(501, 220)(283, 321)(543, 436)$$

Hints. a) The answer to part a) is not unique.

b)We set up the mapping from m to P_m so that the inverse is easy to find.

A Few Remarks on Public Key Cryptography

Now we should discuss public key codes. Let's just say a little about the analog of the Diffie-Helman key exchange. Two other methods are also given in Koblitz, *A Course in Number Theory and Cryptography.*

Suppose that Delenn is on Minbari and John is on Babylon 5. They want to send messages to each other without having the shadows understand these messages. They publicly choose an elliptic curve $E(\mathbb{Z}/p\mathbb{Z})$. Their key will be built out of a point $z = (x, y)$ on $E(\mathbb{Z}/p\mathbb{Z})$. They must choose z such that all communication with each other is public and yet they are the only ones who know z.

1) Delenn and John choose g in $E(\mathbb{Z}/p\mathbb{Z})$ to be their base (not necessarily a generator of the group G of the elliptic curve, since G may not be cyclic). In fact, it may be hard to show g generates G even if it does, and it is even hard to find the order of G. But we do want the subgroup $< g >$ generated by g to be large (near the order of G). One also wants the order of G to be different from p.

2) To create a key, Delenn chooses an integer a near the order of G. She then computes ag in G which she makes public. John chooses an integer b near the order of G and makes bg in G public. The secret key they use is then $z = abg$ in G. Both users can compute z. For example, Delenn knows bg, which is public and her secret a. But a shadow creature knows only ag and bg. It should not be able to find abg, assuming it cannot find a and b.

3) One way to guarantee that g generates the group G is to make sure that the order of G is a prime. Any subgroup has order dividing the order of G and thus $< g >$ must equal G if g is not the identity in G.

4) What is the advantage of elliptic curve cryptography over RSA described in Chapter 21 of Part I of these Lectures? The public keys can be much smaller than those for RSA.

6) Actually one uses finite fields of prime power order as well.

To end this section, which is really the end of this set of lectures, we include Figures 28 and 29 which are 2 pictures of level "curves" of $y^2 - x^3 + x (\mathrm{mod}\, 29)$.

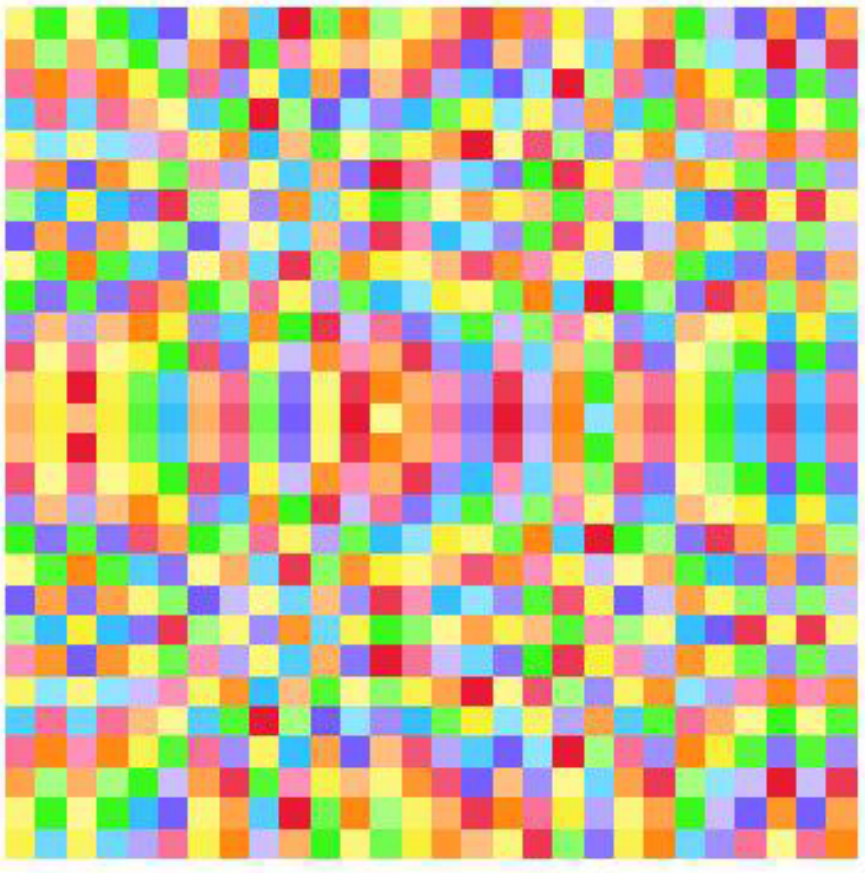

Figure 28: level "curves" of $y^2 - x^3 + x (\mathrm{mod}\, 29)$

Figure 30 is a rotated version of $x^4 + y^4 (\mathrm{mod}\, p)$.

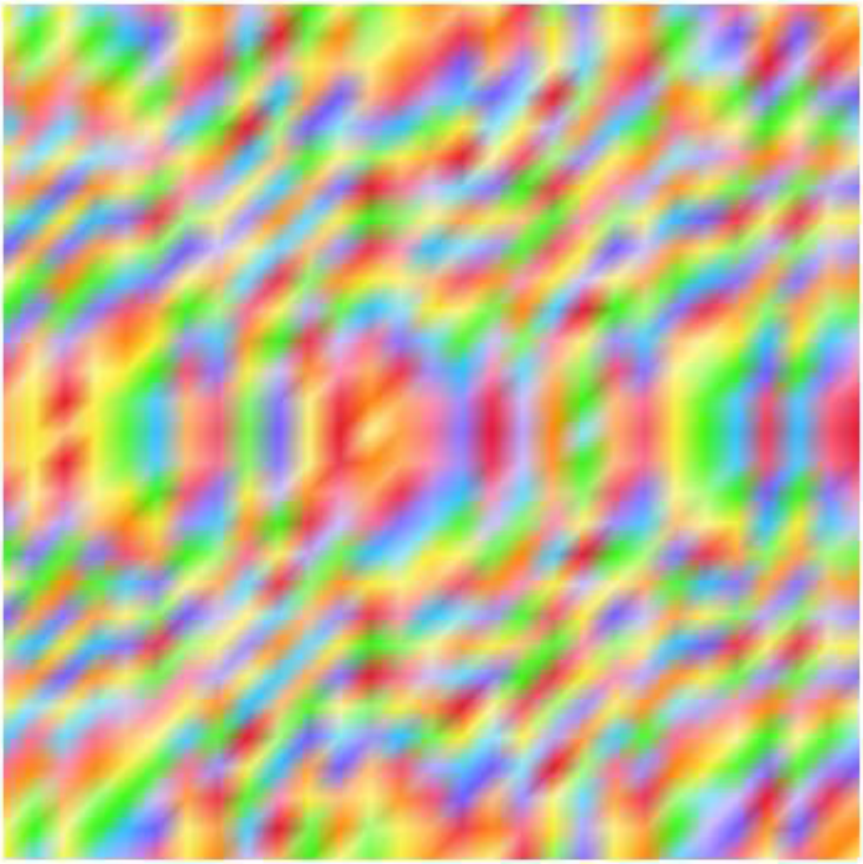

Figure 29: smoothed level "curves" of $y^2 - x^3 + x (\mathrm{mod}\, 29)$

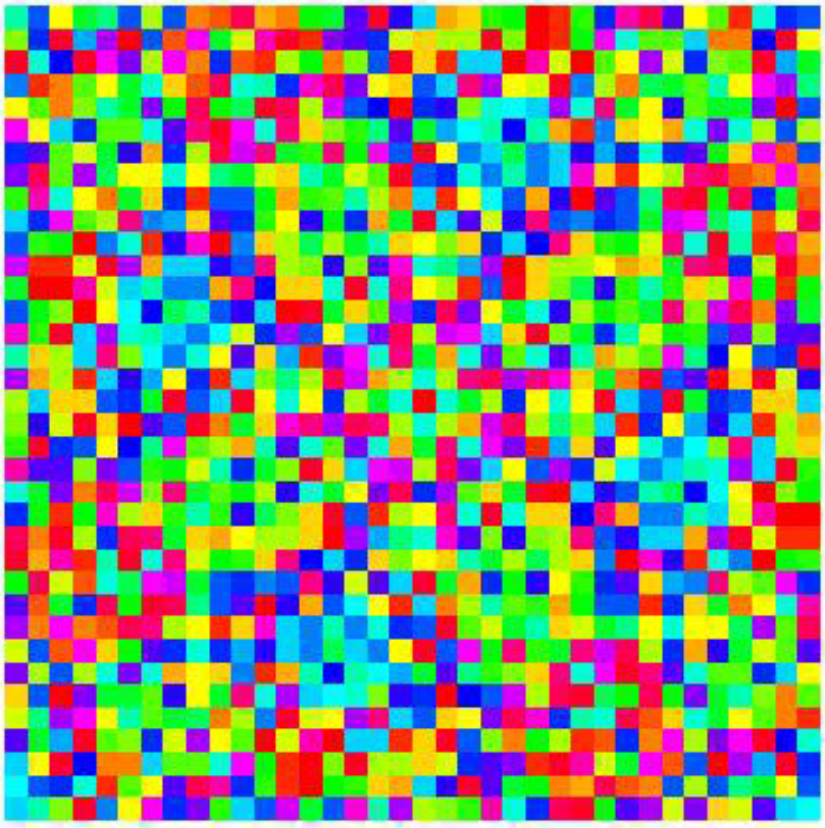

Figure 30: level "curves" of $(y+2x)^4+(x-2y)^4 (\mathrm{mod}\, 37)$.