Fractional Linear Functions:
Some Thoughts on Permutations

by Justin Lanier

Suppose there are some objects on the table in front of you, arranged in an order.

A B C D E F

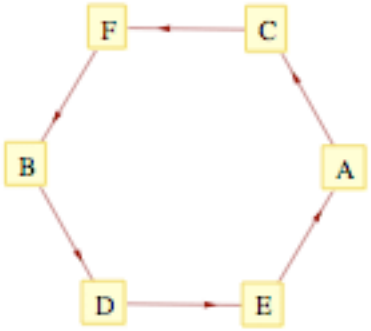
I want you to rearrange them in a certain way, and I want to communicate to you just how that is. How can I do this? The most straightforward thing for me to do is to tell you, for each object, where I want you to move it. For instance, I could write out for you a small table:

Move A to where C was.
Move B to where D was.
Move C to where F was.
Move D to where E was.
Move E to where A was.
Move F to where B was.

or, to abbreviate a bit, a chart:

This element	A	B	C	D	E	F
Goes to	C	D	F	E	A	B

or even draw you a picture:



All of these would tell you what I wanted you to do with the objects. Each one tells you, in a different format, six individual pieces of information—one for each object.

However, some rearrangements—or *permutations*—of these objects could be communicated to you more briefly. The simplest request I might make of you (as long as you aren't too antsy) would be:

`Leave all of the objects where they are.`

Here what I want you to do can be summed up in just one piece of information. I am able to say the same thing about every object—leave it alone. Instead of having to specify what I want you to do to each object individually, I'm able to deal with all of them in one fell swoop. We'll call this permutation the *identity permutation*, since the arrangement of objects stays the same.

Are there other permutations that can be explained so concisely? To do so, we must in some sense do the same thing to each element—to have a single rule to apply to each element that will tell us where each one will go. Of course, we need not be so concise; if we had two distinct rules—one for some elements, another for the rest—this would be an improvement over giving six very specific rules. For now, though, let's think about what we could accomplish with a single rule. One such rule is

`Reverse the order of the objects.`

which yields the arrangement

F E D C B A

Another is

`Move the last object to the first place.`
`Move each other object to the right one place.`

which yields

F A B C D E

This second example may at first seem like two rules, since the last object does something special; if it also moved to the right one place, there wouldn't be a rearrangement at all! Here two ways of thinking about this seemingly two-part rule. First, what if I instead wrote

`Move the last object to the front.`

What have we lost by eliminating that second part? Well, our objects no longer occupy the same locations that they did before. But if we only care about their ordering and not their location, this single rule serves just as well. Here, like so often in mathematical tasks, we have to choose and bear in mind what we care about and feel free to ignore the rest.

The second approach to thinking about this rule lies in a question: how can moving everything to the right act specially on one object, bringing it back to the beginning? If we thought of the objects as lying on a circle, this seeming exception disappears. We have a mathematical apparatus in hand to model this circle-like structure: modular arithmetic. For the sake of convenience, since our objects' only significant quality is just their order, let's exchange them for some different ones:

0 1 2 3 4 5

We can now write our permutation as

`Move x to x+1 (mod 6)`

or, formalizing a bit:

$$f(x) = x + 1 \pmod{6}.$$

We can obtain other shifting permutations by adding different constants; in fact we can summarize these permutations in a mod 6 addition table:

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

What other functions—which is just to say "rules couched in mathematical terms"—can we write that will be concise ways of stating permutations? One idea might come from our earlier order-reversing rule. This permutation has the feel of making each element into its "negative", but doing this in mod 6 holds 0 and 3 fixed, while switching out pairs of the other elements. Our feeling, however, leads us in the right direction and bears fruit if we relocate to a slightly different environs. If we number our elements

1 2 3 4 5 6

and consider them mod 7, the function $g(x) = -x \pmod{7}$ produces the desired result.

When we consider our objects mod 7, we cannot add a nonzero constant to permute them, since this would carry some object to 0, a place where no object was before. We can, however, multiply by a nonzero constant. Since there are six of these, and we don't want to recount the identity, we have found five new permutations.

Where else can we turn? What other frameworks can we set our elements in, or in what new ways can we utilize the frameworks we've already established? Using larger moduli seems problematic, since they will contain places "off the table" that our functions will all-too-likely send our objects. Functions of a higher power will not give us permutations the vast majority of the time, and we would have to be very selective to avoid sending multiple objects to the same place.[1] Neither of these expansions seems particularly promising as a source for permutations of an arbitrary number of objects, the way our shifts and multiples are.

Looking to our experience of modular arithmetic, we find another kind of operation available to us: multiplicative inverses. If we consider our elements again in mod 7 and enumerated as

$$1\ 2\ 3\ 4\ 5\ 6$$

then by taking the inverse of each element, we get back

$$1\ 4\ 5\ 2\ 3\ 6$$

A very nice permutation—but I'd note that this is the first rule we've encountered for which I can't manage to envision as a straightforward spatial process. The sitting, shifting, stretching, and flipping of addition and multiplication can be seen; with multiplicative inverses, my spatial intuition yields to calculation and understanding. Something is lost when this happens, but something is also gained; letting the mathematical structure take us where it will means we can say more, but we must leave the seeing to the machine.

To build upon the permuting that we did by inverting each element, we could also multiply them all by a constant as we did above to further shuffle them. We can't, however, add a constant, since this would take an element "off the table." Is there a way that we could have the permuting tools of addition, multiplication, and inverses all together in the framework of one modulus? Mod 7 leaves out addition, since it's too big to allow for shifts—as would any larger modulus. Mod 6 is limited both with respect to multiplication and to inverses, since it contains zero divisors. And mod 5—while it has no problem with shifts or zero divisors—seems like an unlikely candidate; it doesn't even have enough room in it for all of our objects! Besides, if we numbered our objects in mod 5,

$$0\ 1\ 2\ 3\ 4\ ?$$

we would have an uninvertible element: 0. Well, if mod 5 is our best chance to have all the permuting functions we've thought of so far at our disposal, perhaps we can fudge a little bit, and even kill two birds with one stone. We need both another representative for

[1] Even the simplest quadratic function $f(x) = x^2$ would send pairs of objects to the same location when applied to all but the smallest sets.

our last object and an inverse for zero. Hmm… An inverse for zero… What should $\frac{1}{0}$ be? How about infinity?

$$0\ 1\ 2\ 3\ 4\ \infty$$

By adding this "point at infinity," our little number system is now closed under addition, subtraction, multiplication, and division—all of the tools we pointed out for permuting elements. We've gone from $\mathbf{Z}_5$ to $\mathbf{P}_5$—"P" for projective, or for point at infinity—and within this number system, we can apply a very general kind of function that will give us many different permutations:

$$f(x) = \frac{ax+b}{cx+d}\ ,\quad a,b,c,d \in Z_p,\ \ ad-bc \neq 0$$

This is the general form of a *fractional linear function*, so called because it is the division of one linear polynomial by another.[2] The structure of this kind of function and the permutations it yields on $\mathbf{P}_p$—$\mathbf{Z}_p$ plus a point at infinity, p a prime—will be our focus from here on out. $ad-bc$ must not equal zero, for if it did,

$$ad = bc$$

$$\frac{a}{c} = \frac{b}{d}$$

To simplify things some, let's establish a convention. Observe that if *c* is nonzero, we can divide both the numerator and denominator by *c* to change *c* to 1. If *c* is zero, then the numerator is being divided by a constant *d*, and can thus be rewritten as a simple linear function.[3] Thus, we really only need to consider fractional linear functions of the forms

$$f(x) = \frac{ax+b}{x+d} \quad \text{and} \quad f(x) = ax+b$$

To get a feel for how a fractional linear function operates, let's apply to $\mathbf{P}_5$ the function

$$f(x) = \frac{3x+1}{x+4}.$$

[2] I'll take up some of the more technical points about fractional linear functions in Appendix A: that they form a group under composition, that they are equivalent to GL(2, $\mathbf{Z}_p$), and that they do in fact yield permutations. For the body of the paper, take these for granted.

[3] For a very long time, it was my convention to let a be 1 or 0. I did this, I suppose, just because it was the first coefficient. Below we'll see that doing this to c instead is suggested by the geometry. Note: always look for a natural choice to replace an arbitrary one.

When we plug 0 in for x, we get $\frac{2}{3}$, or twice the inverse of 3. Since the inverse of 3 is 2, $f(0) = 2 \times 2 = 4$. Continuing in like fashion for other finite values of x, we find that

$$f(1) = \frac{3 \times 1 + 1}{1 + 4} = \frac{4}{5} = \frac{4}{0} = 4 \times \infty = \infty$$

$$f(2) = \frac{3 \times 2 + 1}{2 + 4} = \frac{7}{6} = \frac{2}{1} = 2$$

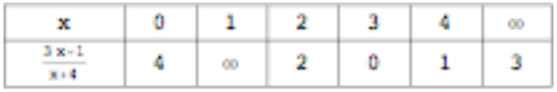
$$f(3) = \frac{3 \times 3 + 1}{3 + 4} = \frac{10}{7} = \frac{0}{2} = 0$$

$$f(4) = \frac{3 \times 4 + 1}{4 + 4} = \frac{13}{8} = \frac{3}{3} = 1$$

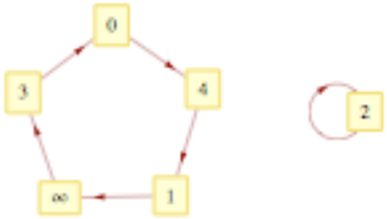
Finding $f(\infty)$ requires a little finesse. We are taught early on in our calculus classes to be wary of evaluating expressions that look like $\frac{\infty}{\infty}$, since such an expression could take on a whole host of values, depending on the relative speeds that the numerator and denominator are heading to infinity. In this case, however, we can be assured that the infinities are the "same size", since they are the same object from our set. Thus, when we encounter $\frac{\infty}{\infty}$, we can evaluate it as 1. Returning to our function, then, we have

$$f(\infty) = \frac{3 \cdot \infty + 1}{\infty + 4} = \frac{3 \cdot \infty}{\infty} = 3$$

In addition to conforming to our sense of these infinities being the "same size," the resulting evaluation fits with our expectations; 3 was the only value we had yet received from our function, and thus this interpretation of $f(\infty)$ completes our permutation. Now that we've evaluated our function for every element of **P**$_5$, we can display our results in either a table or a picture:



x	0	1	2	3	4	∞
$\frac{3x+1}{x+4}$	4	∞	2	0	1	3



So how many fractional linear functions are there, for a given p? Let's begin by considering how many possible denominators there are over $\mathbf{P}_p$. From our convention above, we have one possible denominator when c=0 (namely 1, which we don't even have to write). When c=1, d can range over the p finite elements of $\mathbf{P}_p$. This gives us a total of p+1 possible denominators. When we consider the a and b of the numerator, we find that they can each range over the p elements $\mathbf{Z}_p$, except that the numerator must never be a constant multiple of the numerator. If it were, then the function would map every element to the inverse of that constant—the reason why we forbade ad-bc to equal 0. Each denominator has p constant multiples of itself, and thus p forbidden numerators. Taking this figure from the p^2 total pairings of a and b leaves p^2-p numerators to be paired with each denominator. As there are p+1 different denominators, we have

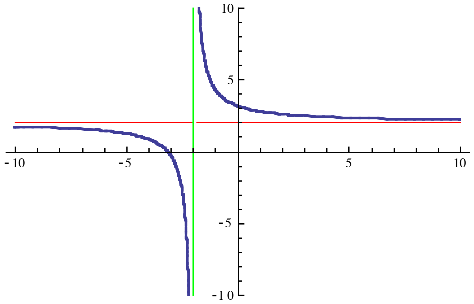
$$(p+1)\cdot(p^2-p) \;=\; (p+1)(p)(p-1)$$

fractional linear functions over $\mathbf{P}_p$, and so also that many different permutations. For $\mathbf{P}_5$, that's 120 permutations—many more than we were generating before!

How does this figure compare to the total number of permutations of the elements of of $\mathbf{P}_p$? Well, since of $\mathbf{P}_p$ contains p+1 elements, there are (p+1)! possible permutations of these elements; we get to choose where every element goes, that is, from among the places that have not already been occupied. For $\mathbf{P}_5$, that's 720 permutations, as we noted above. The similarity of the formula for the number of fractional linear functions to a factorial suggest that instead of getting to choose where every element goes—as we do with a general permutation—for a fractional linear function, we only get to choose where three elements go. Once these are chosen, the other mappings are frozen. Can we justify this suggestion?

We can, and in fact both geometrically and algebraically. If we graph a fractional linear function:

$$f(x)=\frac{ax+b}{x+d}$$



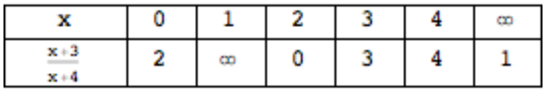
we obtain a pair of hyperbolae with asymptotes x=a and y=-d. This corresponds to what happens in the algebra when we allow x=∞ and x=-d:

$$f(\infty)=\frac{a\cdot\infty+b}{\infty+d} \;=\; \frac{a\cdot\infty}{\infty}=\;a \qquad \text{and} \qquad f(-d)=\frac{a\cdot(-d)+b}{-d+d} \;=\; \frac{-ad+b}{0}=\;\infty$$

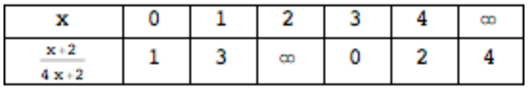
Our choices of *a* and *d*, then, can be seen as choices about where ∞ gets mapped and what gets mapped to ∞. Our remaining choice, that of *b*, scales the hyperbolae in and out while keeping the asymptotes fixed—picking one more input-output pair for our function. Any point we pick will have one and only one pair of hyperbolae that passes through it and has the asymptotes we've assigned. Now, we can construe the *meaning* of our selection of coefficients differently—as picking where three arbitrary elements map, rather than picking the special mappings that involve infinity, plus some other one. However, the fact that the form of a fractional linear function necessitates both a horizontal and vertical asymptote, this gives two constraints on the shape of the conic produced[4], and three more constraints—namely a, b, and d—will determine the hyperbola completely. That we have three choices corresponds exactly to the fact that there are $(p+1)(p)(p-1)$ fractional linear functions over $\mathbf{P}_p$.

The proportion of permutations of $\mathbf{P}_p$ that can be obtained through fractional linear functions gets small very quickly as *p* increases. Though we're getting many more than we were within our other frameworks, fractional linear functions still seem like a rather limited tool. We'll have a chance to reflect more on this fact and its repercussions for our larger project of giving concise descriptions of permutations after we have a better sense for what those limitations actually are.

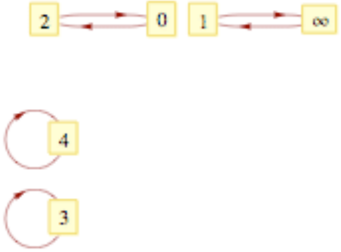
If we look at how a few more fractional linear functions act on $\mathbf{P}_5$—and believe me, I've looked at a whole bunch!—some patterns begin to emerge:

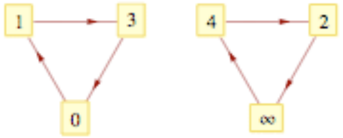


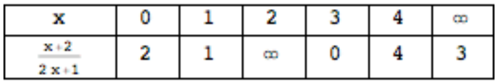
x	0	1	2	3	4	∞
$\frac{x+3}{x+4}$	2	∞	0	3	4	1



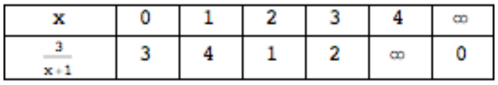
x	0	1	2	3	4	∞
$\frac{x+2}{4x+2}$	1	3	∞	0	2	4



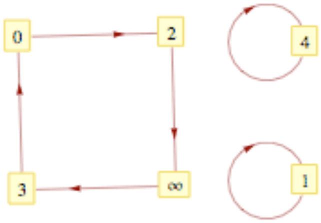


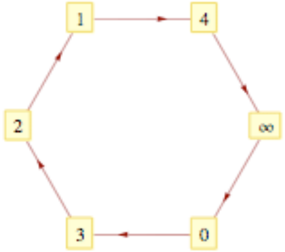


x	0	1	2	3	4	∞
$\frac{x+2}{2x+1}$	2	1	∞	0	4	3



x	0	1	2	3	4	∞
$\frac{3}{x+1}$	3	4	1	2	∞	0





[4] Since $y=\frac{ax+b}{x+d}$, $xy+dy=ax+b$, and since this is a quadratic in x and y, the curve is a conic.

The simplest observation we might make is that fractional linear functions seem to put each element into a *cycle*. A goes to B, B goes to C, and so on, until eventually we arrive back at A. That this occurs is closely tied to the fact that a fractional linear function permutes the elements.

One notable feature of these representatives is that some of them have *fixed points*—values of x for which $f(x) = x$, objects that sit in place while others get shifted around. We see examples where there are no fixed points, one fixed point, and two fixed points. This raises the question: when will a fractional linear function have fixed points, and how many will it have?

Another feature of these representatives is the size of the cycles they contain. If we set aside any fixed points, it looks like the cycles that remain are all the same size for a given fractional linear function. Granted, there are only a few examples to draw upon above, but this pattern continues to hold throughout the rest of $\mathbf{P}_5$, as well as on $\mathbf{P}_p$ in general. Thus another question arises: what is at the root of these same-sized cycles, and how can we predict what sized cycles a given fractional linear function will produce?

It's not hard to demonstrate why a fractional linear function puts the elements of Pp into cycles. If we apply the function to an element, and then again to the resulting element, and so on, we'll eventually have to return to an element we've been to before, since the number of elements in $\mathbf{P}_p$ is finite. Either we return to the element we began with, in which case a cycle has been made, or we return to some other element that we didn't begin with. But then that element has two different elements that point to it—the one we came through originally, and the one from which we've returned to it. This, however, is impossible, since fractional linear functions are one-to-one.[5] Therefore we must return to where we began, and so we see that every element must be involved in a cycle.

Let's begin, then, with the question about fixed points. Given a fractional linear function, we can find its fixed points by examining the equation

$$\frac{ax+b}{cx+d} = x$$

Multiplying through by *cx*+*d*, bringing all terms to the right, and simplifying yields

$$ax+b = x(cx+d) \quad \rightarrow \quad ax+b = cx^2+dx \quad \rightarrow \quad 0 = cx^2+dx-ax-b \quad \rightarrow \quad 0 = cx^2+(d-a)x-b$$

Since this last equation is a quadratic over $\mathbf{P}_p$, it can have at most two roots. This entails that a fractional linear function can have at most two fixed points. To determine exactly how many and what they are, we can use the quadratic formula to obtain

[5] See Appendix A.

$$x = \frac{-(d-a) \pm \sqrt{(d-a)^2 + 4bc}}{2c}$$

Ordinarily, we could look at the discriminant to decide how many real roots our original equation would have: two if the discriminant is positive, a double root if it's zero, and no real roots if it's negative. Since we're working over $\mathbf{P}_p$, however, we have to rethink what the values of the discriminant entail. Finding $\sqrt{-1}$ in $\mathbf{P}_5$, for instance, is no trouble; since -1 is 4, 2 and -2 work just fine. We care, then, not whether the discriminant is positive or negative, but whether or not it's a quadratic residue in $\mathbf{P}_p$.

We can thus set up the following table:

$(d-a)^2 + 4bc$ is a QR (mod p) → $f(x) = \dfrac{ax+b}{cx+d}$ has **two** fixed points

$(d-a)^2 + 4bc = 0$ (mod p) → $f(x) = \dfrac{ax+b}{cx+d}$ has **one** fixed point

$(d-a)^2 + 4bc$ is a QNR (mod p) → $f(x) = \dfrac{ax+b}{cx+d}$ has **no** fixed points

That would seem to settle the question of fixed points entirely. But something struck me as I came upon this result: isn't there a fractional linear function that has *many* fixed points—namely, the identity? How does the theory account for this? Well, if we go ahead and plug in the coefficients of the identity function into our formula,

$$\begin{matrix} a=1 & b=0 \\ c=0 & d=1 \end{matrix} \qquad x = \frac{-(1-1) \pm \sqrt{(1-1)^2 + 4 \cdot 0 \cdot 0}}{2 \cdot 0} = \frac{0 \pm \sqrt{0+0}}{0} = \frac{0}{0}$$

we find yet another indeterminate form! What can this mean? Well, $\frac{0}{0}$ could be any number at all, so I took it that x could be any number, and thus every number is a fixed point—just what we'd expect from the identity. The result that every point is fixed by the identity is nothing to get worked up about, but the way in which the calculation showed that every element was *forced* to be a fixed point was to shape my thinking on other questions.

We've now given a complete account of fixed points of fractional linear functions over Pp. We now turn to questions concerning other sized cycles. First, when do fractional linear functions produce *n*-cycles on Pp—in particular, when do fractional linear functions force *n*-cycles on every Pp? Second, what sized cycles can fractional linear function produce on Pp for some particular p?[6] Now, asking when we get fixed points—which we've just provided an answer to—is the same as asking when we get 1-cycles. This connecting thought also brings up the possibility that the method we used for fixed points might also be useful for considering n-cycles. Let's begin small: when do we get a 2-cycle? Well, whenever A gets mapped to B and B gets mapped to A; that is, when the function is applied to an element twice, it gives back that element. More formally,

$$\text{for } f(x)=\frac{ax+b}{cx+d}, \quad \text{whenever } f(f(x))=x, \quad \text{x is involved in a 2-cycle,}$$

—unless, of course, it's a fixed point. After all, if $f(x)=x$, certainly $f(f(x))$ is as well. We'll have to be careful to weed out such "false m-cycles." To find out if any particular element is involved in a 2-cycle, we just have to calculate. For instance, in $\mathbf{P}_5$

$$\text{when } f(x)=\frac{2x+3}{x+3},$$

$$f(f(4))=f\left(\frac{2\cdot 4+3}{4+3}\right)=f\left(\frac{11}{7}\right)=f\left(\frac{1}{2}\right)=f(3)=\frac{2\cdot 3+3}{3+3}=\frac{9}{6}=\frac{3}{2}=4$$

Thus 4 and 3 are bound together in a 2-cycle. We could have associated our operations differently and found $f\circ f$ before $f(4)$. To do this, we could either have plugged f into itself and then done a lot of simplifying algebra, or we could have used the tool of matrix multiplication to arrive at the result more expediently.[7] More importantly, by carrying through this $f\circ f$ in general via either method, we can proceed to set up an equation to characterize what fractional linear functions force 2-cycles onto every $\mathbf{P}_p$ — what fractional linear functions yield the identity when "squared." In general,

$$f^2=\begin{pmatrix} a & b \\ c & d\end{pmatrix}\begin{pmatrix} a & b \\ c & d\end{pmatrix}=\begin{pmatrix} a^2+bc & ab+bd \\ ac+cd & bc+d^2\end{pmatrix}$$

The identity matrix is of the form

$$\begin{pmatrix} m & 0 \\ 0 & m\end{pmatrix}, \quad \text{since} \quad \frac{mx}{m}=x$$

Setting f^2 equal to the identity, or in matrix language,

[6] It is interesting to note the similarity between these questions and those we ask when we are investigating quadratic residues, namely: when is x a square mod p? and, what are the squares mod p? We look either at one object over all its environments, or at all of the objects in one environments.

[7] See Appendix A for details.

$$\begin{pmatrix} a^2 + bc & ab + bd \\ ac + cd & bc + d^2 \end{pmatrix} = \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix}$$

we get three equations in four unknowns:

$$a^2 + bc = bc + d^2 \qquad ab + bd = 0 \qquad ac + cd = 0$$

These equations give in full the restrictions on a, b, c, and d so that the fractional linear function will force 2-cycles on every non-fixed point in Pp. However, they aren't a very concise way of describing these constraints; while they give us the means to check whether a given fractional linear function forces 2-cycles, they don't let us see the answer at a glance. Our task, then, is to rewrite these restrictions in a different form more amenable to our psychology without changing their content. In other words, we're going to do some algebra.

$a^2 + bc = bc + d^2 \rightarrow a^2 = d^2 \rightarrow a = \pm d$

If $a = d$, $ab + bd = 0 \rightarrow ab + ba = 0 \rightarrow a(b + b) = 0 \rightarrow a = 0$ or $b = 0$

$a = 0 \rightarrow d = 0$, and so $f(x) = \dfrac{b}{cx}$ where b and c are nonzero.

$b = 0$ and a $\neq 0$: $ac + cd = 0 \rightarrow ac + ca = 0 \rightarrow a(c + c) = 0 \rightarrow c = 0$, and so $f(x) = I$.

If $a = -d$, $ab + bd = 0 \rightarrow ab + b(-a) = 0 \rightarrow b(a - a) = 0 \rightarrow b$ can be anything.

$ac + cd = 0 \rightarrow ac + c(-a) = 0 \rightarrow c(a - a) = 0 \rightarrow c$ can be anything.

So $f(x) = \dfrac{ax + b}{cx - a}$.

Disregarding the case of the identity, we find either

$$f(x) = \frac{b}{cx} \quad \text{or} \quad f(x) = \frac{ax + b}{cx - a},$$

where the former is a special case of the latter. In general, then, we can say that a fractional linear function will force 2-cycles whenever $a = -d$, or $a + d = 0$.

By much the same process, we can determine what fractional linear functions force other sized cycles on Pp. For $f(x)$ to have n-cycles, n must be the smallest power of f such that $f^n = I$. In order to find out what restrictions are placed on a, b, c, and d so that this will occur, we find

$$f^n = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^n = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \quad \text{and set it equal to the identity} \quad \begin{pmatrix} m & 0 \\ 0 & m \end{pmatrix}$$

and this yields the three equations:

$$A = D \qquad B = 0 \qquad C = 0$$

We can then shuffle these linear equations around into a form that's convenient. Now, actually carrying out this process for each n is a little tedious, and so I'll let the case for n = 2 that we looked at above serve as our example. The restrictions on a, b, c, and d that emerge from this process for various n, however, are extremely interesting. Here are the first several:

$$\text{the identity, or a single fixed point}^{8}\text{: } \frac{(a+d)^2}{4} = ad - bc$$
$$\text{2-cycles: } a + d = 0$$
$$\text{3-cycles: } (a+d)^2 = ad - bc$$
$$\text{4-cycles: } \frac{(a+d)^2}{2} = ad - bc$$
$$\text{6-cycles: } \frac{(a+d)^2}{3} = ad - bc$$

Two observations about these conditions: first, they seem to obey a regular law, relating the square of the trace (*a*+*d*) to the determinant (*ad*-*bc*) linearly. For now, this must remain a mere observation, as I haven't given enough thought as to why this should happen. Second, these results show that 2-, 3-, 4-, and 6-cycles can be forced on to Pp for any *p*. The conditions specify nothing about the structure or elements of Pp, and so a properly constructed linear fractional function will force these sized cycles on $\mathbf{P}_p$ for any *p*.

When we consider the criteria for forcing other sized cycles, the criteria cannot be met in Pp for every p. This first occurs in the criterion for 5-cycles:

$$\text{5-cycles: } \frac{(3 \pm \sqrt{5})(a+d)^2}{3} = ad - bc$$

5-cycles can thus only be forced when $\sqrt{5}$ is a quadratic residue mod *p*. That 2-, 3-, 4-, and 6- cycles can be forced on any Pp, while no other sized cycle can, fits with our earlier conjecture that (to be proven below) that all non-fixed points are involved in cycles of the same size. Since the number of non-fixed points in Pp is either p+1, p, or p-1, we would only expect cycle lengths in Pp that could divide p+1, p, or p-1. If p is 2 or 3, the set is so small that it's obvious that we can get any cycle length. For $p > 3$, p is odd, and so both p+1 and p-1 are even, and one is a multiple of four. Also, since p is not a multiple of 3, either p+1 or p-1 is. Finally, whichever of these is a multiple of three, it is also a multiple of two, and thus is a multiple of six. These are the only numbers that we can guarantee

[8] Above I wrote $(d - a)^2 + 4bc = 0$ for a single fixed point, but this formulation is equivalent and harmonizes more nicely with the other criteria.

divides at least one of p+1, p, or p-1 for every p. They are also the only sizes of cycles that can be force on any Pp.

5 only divides p+1, p, or p-1 when $p = 5$ or $p \equiv \pm 1 \pmod{10}$. This second option is exactly when $\sqrt{5}$ is a quadratic residue mod p. Thus it appears that here, too, we get 5-cycles every time they are possible. Now, the computation for deriving the conditions for forcing n-cycles becomes increasingly messier as n gets large. Still the same sort of phenomenon seems to hold for larger n—that the constant that shows up under the radical in the criterion for forced n-cycles is a quadratic residue mod p exactly when n divides p+1, p, or p-1. This points to the idea that a fractional linear function can be constructed that will force every possible cycle length on Pp for every p. More concretely, for say P19, this suggests that we could find fractional linear functions that would give:

no fixed points, and all cycles of length either 2, 4, 5, 10, or 20
one fixed point, and a cycle of 19
two fixed points, and all cycles of length either 2, 3, 6, 9, or 18.

Observation also suggests that this is true. That cycles of every possible length can in fact be made on Pp is something that I hope to demonstrate in the near future.

So let's talk about what is meant by "possible" cycle lengths, switching to our second question, "what sized cycles can fractional linear function produce on Pp for some particular p?" Our conjecture is that all cycles made by a fractional linear function on some particular $\mathbf{P}_{\mathrm{p}}$, once we've set aside fixed points, are the same size, and thus this cycle length is a divisor of p+1, p, or p-1, depending on how many fixed points there are. I'll begin by showing that the length of each cycle must divide the number of non-fixed points, and then I'll show how this forces them all to be the same size. In what follows, I've made use of some of the apparatus of linear algebra. As we've seen above, thinking about many iterations of a fractional linear function is made easier by using matrices, and these matrix calculations are made less cumbersome by putting the matrices involved in convenient forms—as close to diagonal as possible. I've used the theory of eigenvalues and eigenvectors to put the matrices in question into that form. You can think of this as changing into and out of bases—like we might to in ordinary arithmetic in order to ease computation. It's worthy of consideration to note the interplay between matrix theory and number theory in the following proofs.

Theorem: The order of any cycle produced by a fractional linear function on Pp divides the number of elements in Pp minus the number of fixed points the fractional linear function produces. This breaks down into three cases:

a) Every cycle's length divides p-1 when there are two fixed points.
b) Every cycle's length divides p when there is one fixed point.
c) Every cycle's length divides p+1 when there are no fixed points.

a) If a fractional linear function F has two fixed points on Pp, then $F^{p-1} = I$.

Proof: Since F is nonsingular, it can be represented as a conjugated diagonal matrix.

$$F = MDM^{-1}$$

Since F has two fixed points, $(d-a)^2 + 4bc$ is a QR (mod p); thus we obtain two distinct, nonzero eigenvalues $\lambda_1, \lambda_2 \in \mathbf{Z}_p$ for F and

$$F = M\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}M^{-1}$$

Raising both sides of the equation to the p-1 power, the interior MM^{-1} pairs cancel out, leaving

$$F^{p-1} = M\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}^{p-1} M^{-1}$$

By Lemma 1,[9]

$$F^{p-1} = M\begin{pmatrix} \lambda_1{}^{p-1} & 0 \\ 0 & \lambda_2{}^{p-1} \end{pmatrix}M^{-1}$$

Since $\lambda_1, \lambda_2 \in \mathbf{Z}_p$ and are nonzero, by Fermat's Little Theorem we have

$$F^{p-1} = M\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}M^{-1} = MM^{-1} = I$$

Q.E.D.

b) If a fractional linear function *F* has one fixed point on Pp, then $F^p = I$.

Proof: Since F is nonsingular, it can be represented as a conjugated diagonal matrix.

$$F = MDM^{-1}$$

Since F has one fixed point, $(d-a)^2 + 4bc = 0$; thus we obtain a single nonzero eigenvalue $\lambda \in \mathbf{Z}_p$ for F and

$$F = M\begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}M^{-1}$$

[9] See Appendix B.

Raising both sides of the equation to the p power, the interior MM^{-1} pairs cancel out, leaving

$$F^{p} = M\begin{pmatrix}\lambda & 1\\ 0 & \lambda\end{pmatrix}^{p} M^{-1}$$

By Lemma 2,[10]

$$F^{p} = M\begin{pmatrix}\lambda^{p} & p\lambda^{p-1}\\ 0 & \lambda^{p}\end{pmatrix}M^{-1}$$

Since $p\lambda^{p-1} \equiv 0 \pmod{p}$,

$$F^{p} = M\begin{pmatrix}\lambda^{p} & 0\\ 0 & \lambda^{p}\end{pmatrix} = M\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}M^{-1} = MM^{-1} = I$$

Q.E.D.

If a fractional linear function F has no fixed points on $\mathbf{P}_{\mathrm{p}}$, then $F^{p+1} = I$.

Proof: Since F has no fixed points, $k = (d - a)^2 + 4bc$ is a QNR (mod p); that is, when we attempt to find eigenvalues, we find none within $\mathbf{Z}_{\mathrm{p}}$. Extend $\mathbf{Z}_{\mathrm{p}}$ to $\mathbf{Z}_{\mathrm{p}}[\sqrt{k}\,]$. Now we have two distinct eigenvalues:

$$\lambda_1 = \frac{a + d + \sqrt{k}}{2a} = r + s\sqrt{k} \in \mathbf{Z}_p[\sqrt{k}] \text{ and } \lambda_2 = \frac{a + d - \sqrt{k}}{2a} = r - s\sqrt{k} \in \mathbf{Z}_p[\sqrt{k}]$$

Note that λ_1 and λ_2 are conjugates. We now have

$$F = M\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}M^{-1}$$

Raising both sides of the equation to the p+1 power, the interior MM^{-1} pairs cancel out, leaving

$$F^{p+1} = M\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}^{p+1} M^{-1}$$

By Lemma 1,

[10] See Appendix B.

$$F^{p+1} = M\begin{pmatrix} \lambda_1{}^{p+1} & 0 \\ 0 & \lambda_2{}^{p+1} \end{pmatrix}M^{-1}$$

Recalling that $\lambda_1, \lambda_2 \in \mathbf{Z_p}[\sqrt{k}]$, I claim $\lambda_1{}^{p+1} = \lambda_2{}^{p+1}$.

$$\begin{aligned}
\lambda_1{}^{p+1} &= (r + s\sqrt{k})^{p+1} \\
&= (r + s\sqrt{k})^p(r + s\sqrt{k}) \\
&= (r^p + s^p\sqrt{k}^{\,p})(r + s\sqrt{k}) && \text{since } (a+b)^p = a^p + b^p \pmod{p}. \\
&= (r + s\sqrt{k}^{\,p})(r + s\sqrt{k}) && \text{by Fermat's Little Theorem.} \\
&= (r + s\sqrt{k}(\sqrt{k})^{p-1})(r + s\sqrt{k}) \\
&= (r + s\sqrt{k}(k)^{\frac{p-1}{2}})r + s\sqrt{k}) \\
&= (r + s\sqrt{k}(-1))(r + s\sqrt{k}) && \text{by Euler's Criterion, since } k \text{ is a QNR (mod p).} \\
&= (r - s\sqrt{k})(r + s\sqrt{k}) \\
&= r^2 - s^2k
\end{aligned}$$

By the same argument, $\lambda_2{}^{p+1} = r^2 - s^2k$. Therefore, $\lambda_1{}^{p+1} = \lambda_2{}^{p+1}$, and

$$F^{p+1} = M\begin{pmatrix} \lambda_1{}^{p+1} & 0 \\ 0 & \lambda_1{}^{p+1} \end{pmatrix}M^{-1} = M\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}M^{-1} = MM^{-1} = I$$

Q.E.D.

Now that I've shown that the length of every cycle divides the number of non-fixed points,

Theorem: All non-fixed-point cycles that a fractional linear function F makes on $\mathbf{P}_\text{p}$ are of the same length.

Proof: Here, too, we'll consider three cases, according to the number of fixed points F produces on $\mathbf{P}_\text{p}$.

First, consider the case where the fractional linear function produces one fixed point on $\mathbf{P}_\text{p}$. Suppose it also produces to cycles of different lengths, and m-cycle and an n-cycle, with m>n. Any other cycles can be what they may. Then consider F^n. It, too, is a fractional linear function, since these are closed under composition. After n applications of F, its fixed points have stayed fixed. Also, each member of F's n-cycle has traversed the cycle and has returned to itself. Thus each of these elements is a fixed point for F^n. But the members of the m-cycle have not yet returned to themselves since *m*>*n*. F^n instead maps them to some different element of the *m*-cycle. Thus F^n is not the identity. But then F^n, a fractional linear function not the identity, has at least n+2 fixed points. This is a contradiction, since we've already shown that a fractional linear function has at most 2 fixed points. Thus F must not have cycles of differing lengths, and so all its non-fixed-point cycles must be of the same length.

Next, consider the case where the fractional linear function *F* produces one fixed point on $\mathbf{P}_p$. The proof runs the same way, only noting that *n* is at least 2—the smallest non-fixed-point cycle. Two unequal cycles entails at least 2+1=3 fixed points for F^n, and this too is a contradiction. Thus *F*'s non-fixed-point cycles are all the same length.

Finally, consider the case where the fractional linear function produces no fixed point on $\mathbf{P}_p$. The proof remains the same when *n* is 3 or more, since then F^n would have 3 or more fixed points—a contradiction. Thus the only case left to consider is *n*=2. Then either all of the remaining cycles are of the same length (other than 2) or of different lengths. They cannot all be of the same length, since if the remainder is composed of m L-cycles, then we have p+1=2+mL. Then

$$p - 1 = mL$$

$$\text{so } L|p - 1$$

$$\text{but we must also have } L|p + 1\text{, by our above theorem.}$$

$$\text{Then } L|2.$$

If p>3, L had better not be 1, since then there are too many fixed points; but if L is 2, then all of the cycles are of length 2, which is what was desired.
The last case is if the remaining cycles are of different lengths. This is bad too, and I don't think hard to show, but to be honest, I haven't thought it all the way out, and I need to turn this in.

Almost QED

These are my major results thus far. There are still some aspects of fractional linear functions that I've yet to understand or fully prove; furthermore, fractional linear functions are just one piece of answering the big question: how can permutations be written down simply? I look forward to continuing to investigate this question.