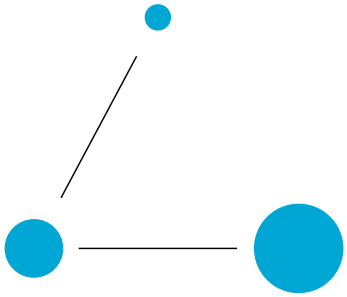
Annales Henri Lebesgue
4 (2021) 227-250



ANNALES
HENRI LEBESGUE

MATTHEW DE COURCY-IRELAND
MICHAEL MAGEE

KESTEN–MCKAY LAW FOR THE MARKOFF SURFACE MOD p

LOI DE KESTEN–MCKAY POUR LA SURFACE DE MARKOFF MODULO p

Abstract. — For each prime p, we study the eigenvalues of a 3-regular graph on roughly p^2 vertices constructed from the Markoff surface. We show they asymptotically follow the Kesten–McKay law, which also describes the eigenvalues of a random regular graph. The proof is based on the method of moments and takes advantage of a natural group action on the Markoff surface.

Résumé. — Pour chaque nombre premier p, on décrit les valeurs propres d'un graphe 3-régulier ayant environ p^2 sommets construit à partir de la surface de Markoff. On montre qu'elles suivent approximativement la loi de Kesten–McKay, qui décrit également les valeurs propres d'un graphe aléatoire régulier. On utilise la méthode des moments et l'action de $\mathrm{GL}_2(\mathbb{Z})$ sur la surface de Markoff.

1. Introduction

The Kesten–McKay Law governs the eigenvalue distribution of a random d-regular graph in the limit of a growing number of vertices [Kes59, McK81]. The limiting

Keywords: Markoff surface, Kesten–McKay law, cubic surfaces, graphs and groups.
2020 *Mathematics Subject Classification:* 11D25, 05C50, 11F72, 37P25.
DOI: `https://doi.org/10.5802/ahl.71`
(*) de Courcy-Ireland's work was supported by the Natural Sciences and Engineering Research Council of Canada through a Postgraduate Scholarships Doctoral grant [PGSD2-471570-2015].

probability density function is

$$\rho_d(\lambda) = \frac{d}{2\pi} \frac{\sqrt{4(d-1)-\lambda^2}}{d^2-\lambda^2} \mathbb{1}_{[-2\sqrt{d-1}, 2\sqrt{d-1}]}(\lambda) \tag{1.1}$$

This spectral density comes from the Plancherel measure on the infinite d-regular tree, and one might expect a similar eigenvalue distribution for non-random d-regular graphs provided they resemble their universal cover closely enough in the sense of having few short cycles. The purpose of this article is to establish such a result for a family of 3-regular graphs constructed from the *Markoff equation*

$$x^2 + y^2 + z^2 = xyz \tag{1.2}$$

modulo large prime numbers $p \to \infty$. The vertices, roughly p^2 in number, are simply the solutions (x, y, z) in $\mathbb{F}_p^3$ excluding $(0, 0, 0)$. The edges connect (x, y, z) to $(x, y, xy - z)$, $(x, xz - y, z)$, and $(yz - x, y, z)$, the Markoff equation being preserved by these operations. If an edge connects a vertex to itself, then it must be counted just once in order for the graph to be 3-regular. We will write $M(\mathbb{F}_p)$ for the vertex set and $\mathfrak{M}_p$ for the graph. The eigenvalues $\{\lambda_j\}$ of the resulting graph can naturally be thought of as a measure on $[-3, 3]$, namely

$$\mu_p = \frac{1}{|M(\mathbb{F}_p)|} \sum_j \delta_{\lambda_j} \tag{1.3}$$

and our main result is that the moments of this measure converge as $p \to \infty$ to those of the Kesten-McKay measure.

THEOREM 1.1. — *There are absolute constants $c > 0$ and $C > 1$ such that for $L \leqslant c \log p$, and with an implicit constant independent of both p and L,*

$$\int x^L d\mu_p = \int x^L \rho_3(x) dx + O\left(\frac{C^L}{p}\right).$$

Thus one can take L to be a small multiple of $\log p$ and the error term C^L/p will remain negligible. More precisely, we require $L < \frac{1}{16 \log 2} \log p - 7 \approx 0.090168 \log p$. Our proof of Theorem 1.1 permits $C = 3 \times 2^{16} = 196608$, which we have not optimized, but an exponential dependence on L is inevitable. As we will explain heuristically at the end of the paper, we do not expect the moments to agree if $L/\log p$ is too large.

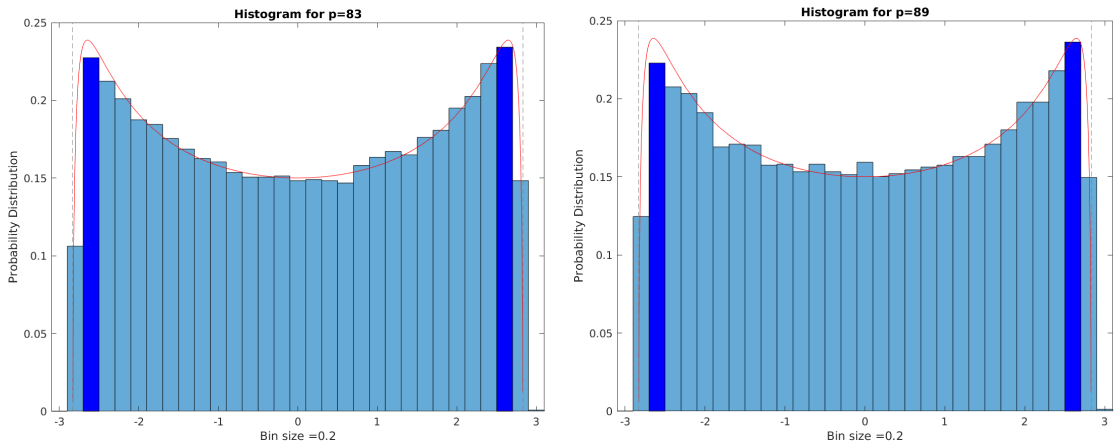
Taking linear combinations and applying Theorem 1.1 for L fixed as $p \to \infty$, we obtain

THEOREM 1.2. — *For any fixed polynomial f, the eigenvalues λ_j of the Markoff graph mod p satisfy*

$$\frac{1}{p^2 \pm 3p} \sum_j f(\lambda_j) = \int_{-2\sqrt{2}}^{2\sqrt{2}} f(\lambda) \rho_3(\lambda) d\lambda + O\left(\frac{1}{p}\right)$$

as $p \to \infty$.

Taking L growing simultaneously with p gives much more information than one could achieve from any fixed L. In particular, we deduce the following bound for the discrepancy between μ_p and ρ_3.



(a) $p = 83$ (b) $p = 89$

Figure 1.1. Histogram of eigenvalues for $p = 83$ and 89 with the density $\rho_3(x)$ shown in red.

COROLLARY 1.3. — *For any interval $J \subseteq [-3, 3]$,*

$$|\mu_p(J) - \rho_3(J)| = O\left(\frac{1}{\log p}\right).$$

Note that the d-regular Kesten–McKay measure is supported on the interval $[-2\sqrt{d-1}, 2\sqrt{d-1}]$. A d-regular graph could in principle have eigenvalues throughout the interval $[-d, d]$, but when the Kesten–McKay law is valid, it implies that most of the eigenvalues lie in a much smaller interval. As an application of Corollary 1.3, we have

COROLLARY 1.4. — *For the Markoff graph mod p, the number of eigenvalues greater than $2\sqrt{2}$ is only $O(p^2/\log p)$ out of a total of $p^2 \pm 3p$.*

It is plausible that one could replace $1/\log p$ by $1/p$ in both of these corollaries. To use our estimates for moments, we approximate the discontinuous indicator function by polynomials, and this entails some loss.

Figure 1.1 shows the histogram of eigenvalues for the Markoff graphs constructed from $p = 83$ and $p = 89$, illustrating the fit to the Kesten–McKay law. For 3-regular graphs, the support is $[-2\sqrt{2}, 2\sqrt{2}]$ and the distribution is bimodal, with maxima at $\pm\sqrt{7}$.

We begin in Section 2 with the overall strategy of comparing the Kesten–McKay moments with those of the graphs we construct. This reduces the problem to counting the fixed points of a natural group action. In Section 3, we compute the fixed points in several examples and outline a heuristic that would give a better dependence on L in Theorem 1.1. In Section 4, we review the connection between the Markoff surface and $\mathrm{GL}_2(\mathbb{Z})$, which is the basis for the actual proof. In Theorem 5.1, we prove that an element has $O(p)$ fixed points with an implicit constant depending on its entries as a matrix in $\mathrm{GL}_2(\mathbb{Z})$. In Section 6, we complete the proof of Theorem 1.1 by noting that the matrix entries are exponential in the length L of the word. In Section 7, we

turn to the proof of Corollaries 1.3 and 1.4. In Section 8, we conclude by comparing the Kesten–McKay law to other (more difficult) questions about the graphs $\mathfrak{M}_p$, in particular their connectedness and spectral gap. We use the rest of this Introduction to summarize some of the recent interest in the Markoff equation and its solutions modulo p.

The original Markoff surface is defined by the cubic equation

$$x^2 + y^2 + z^2 = 3xyz \tag{1.4}$$

and its solutions in nonnegative integers (x, y, z) are called Markoff triples. It differs from our normalization $x^2 + y^2 + z^2 = xyz$ by a scaling $(x, y, z) \mapsto 3(x, y, z)$, which is invertible over $\mathbb{F}_p$ for $p \geqslant 5$. The Markoff equation is a very special case which offers a great simplification compared to other cubic surfaces. The only cubic term in equation (1.4) is $3xyz$, so upon fixing two variables, it is only a quadratic equation for the third. Exchanging the two roots of this quadratic allows us to move from one triple to another. By Vieta's Rule, the two solutions of a quadratic must add up to its middle coefficient, so one such move sends (x, y, z) to another Markoff triple $(x, y, 3xy - z)$. There is another move for each of the variables. Markoff proved in 1880 [Mar80] that any Markoff triple except $(0, 0, 0)$ can be reached starting from the solution $(1, 1, 1)$ by a sequence of Vieta operations and transpositions. In contrast, for a general cubic surface, there is no known method for deciding whether there are integer solutions, let alone finding all of them. For instance, it remains out of reach to determine whether a given number is a sum of three (possibly negative) cubes.

The Markoff triples can be displayed as a 3-regular tree, with $(1, 1, 1)$ as the root and edges giving the action of the Vieta moves. Reducing this Markoff tree modulo a prime p yields a finite graph with cycles, which is one connected component of the graph we study below. In principle, there may be additional solutions over $\mathbb{F}_p$ that do not come from reducing integer solutions mod p. Hence it is no longer guaranteed that all solutions can be found by the Vieta moves, although in practice it seems that they can. If every solution mod p lifts to a solution over the integers, then the same sequence of Vieta moves used to reach the lift will reach its image mod p because the moves are polynomial operations in (x, y, z). Thus the graph of solutions over $\mathbb{F}_p$ will be connected. The connectedness of these graphs for all p is the question of whether *strong approximation* holds for equation (1.2), that is, whether solutions mod p can always be lifted to integer solutions. Baragar was the first to conjecture that this connectedness does hold for all p and he verified it for $p \leqslant 179$ (see [Bar91, p. 124]).

Bourgain–Gamburd–Sarnak [BGS16] proved that, for most primes p, there is only a single component of nonzero solutions $(x, y, z) \neq (0, 0, 0)$. Their method fails in case $p^2 - 1$ has many prime factors, which happens only for rare values of p. Even for these exceptional primes, the Bourgain–Gamburd–Sarnak argument shows that there is a giant component containing, for any given $\varepsilon > 0$, all but p^ε of the vertices, while any putative extra components would have size at least a power of $\log(p)$. On the quantitative level, some improvements have been made by Konyagin–Makarychev–Shparlinski–Vyugin [KMSV20, Theorems 1.3 and 1.4]. Meiri–Puder [MP18] prove that the Markoff action on the largest component is highly transitive: up to grouping solutions by sign changes as in (2.1) below, it is either the full symmetric group or

its alternating subgroup. Cerbu–Gunther–Magee–Peilen [CGMP20] had proposed earlier that the alternating group arises when $p \equiv 3 \bmod 16$, and the full symmetric group otherwise.

2. Method of moments

Let us define the Markoff graph over $\mathbb{F}_p$ more precisely. The vertices are the triples (x, y, z) solving $x^2 + y^2 + z^2 = xyz$, except $(0, 0, 0)$. The most natural graph for our purposes is defined by taking an edge between (x, y, z) and each of its images $(x, y, xy-z)$, $(x, xz-y, z)$, and $(yz-x, y, z)$. We denote the graph by $\mathfrak{M}_p$ and its vertex set by $M(\mathbb{F}_p)$, with edges given by the *Markoff moves* $m_1(x, y, z) = (yz - x, y, z)$, $m_2(x, y, z) = (x, xz - y, z)$, and $m_3(x, y, z) = (x, y, xy - z)$. It has $p^2 \pm 3p$ vertices depending on whether p is congruent to 1 or to 3 modulo 4. The total number of solutions to $x^2 + y^2 + z^2 = xyz \bmod p$ is $p^2 + 3(\frac{-1}{p})p + 1$, but we consider $(0, 0, 0)$ separately from the other solutions because it is in an orbit of its own under the Markoff moves. See Carlitz's note, [Car57, equation (2)] for this count.

At present, we have no guarantee that this graph is connected. Baragar [Bar91] conjectured that $\mathfrak{M}_p$ is connected for any prime p, and Bourgain–Gamburd–Sarnak proved connectedness unless p^2-1 has many small factors in a quantified way [BGS16]. They also prove that, even in a possibly disconnected case, there is a giant component containing at least $p^2 \pm 3p - O(p^\varepsilon)$ vertices for any $\varepsilon > 0$. Our Theorem 1.1 applies both to the whole graph, possibly disconnected, and also to its giant component.

The graphs we study are not simple: Although $\mathfrak{M}_p$ does not contain multiple edges, there are loops at a small fraction of the vertices. On the order of p vertices out of p^2 have loops. We discuss this further in Proposition 3.1, and the presence of loops appears again in Lemma 5.4. It has some importance for our main proofs.

The graph $\mathfrak{M}_p$ is obtained directly from the underlying symmetry of the equation $x^2 + y^2 + z^2 = xyz$ under the Markoff moves m_1, m_2, m_3. Sometimes it may be preferable to take other edges reflecting further symmetries of the Markoff surface. The Markoff equation is preserved by all permutations of (x, y, z) as well as the four *double sign changes* leaving xyz invariant, namely

$$(x, y, z) \mapsto (\sigma_1 x, \sigma_2 y, \sigma_3 z) \tag{2.1}$$

where the signs obey $\sigma_1\sigma_2\sigma_3 = 1$. One could add edges corresponding to any of these. Or one could streamline the graph by first taking the quotient by sign changes, or using alternative generators that combine the Markoff moves with permutations. In this way, one could obtain graphs with fewer loops and a closer fit to the Kesten–McKay law. Nevertheless, the Markoff moves themselves seemed the most natural choice to us.

Let A be the adjacency matrix for the Markoff graph mod p, that is, the matrix indexed by vertices with $A_{ij} = 1$ when there is an edge between i and j and $A_{ij} = 0$ otherwise. Note that the diagonal entries A_{jj} are typically 0, but may be 1 when there is a loop connecting j to j. Permuting the vertices changes the adjacency matrix to $\sigma A \sigma^{-1}$, where σ is the corresponding permutation matrix. Thus the eigenvalues of A do not depend on any choice of ordering. The connectedness of a graph is closely

related to its eigenvalues. Indeed, for a d-regular graph, the number of connected components is the multiplicity of d as an eigenvalue. The Kesten–McKay law is a general theorem about the distribution of eigenvalues for graphs with few short cycles, either random or deterministic. We quote the following theorem of McKay [McK81, Theorem 1.1] to emphasize the generality of the Kesten–McKay law, although we will not be able to use this version to deduce the rate of convergence in Theorem 1.1.

THEOREM 2.1. — *(McKay) If G_i is a sequence of d-regular graphs with n_i vertices such that for each fixed k, the number of k-cycles in G_i is $o(n_i)$ as $n_i \to \infty$, then the eigenvalue counting function*

$$\frac{\#\{j;\ \lambda_j(G_i) \leqslant \lambda\}}{n_i}$$

converges to

$$\int_{-\infty}^{\lambda} \frac{d}{2\pi} \frac{\sqrt{4(d-1)-t^2}}{d^2-t^2} \mathbb{1}_{[-2\sqrt{d-1},2\sqrt{d-1}]}(\lambda) dt$$

as $n_i \to \infty$.

The combinatorial significance of the Kesten–McKay measure is that its moments count walks in a d-regular tree

$$\int_{-2\sqrt{d-1}}^{2\sqrt{d-1}} x^L \rho_d(x) dx = \#\Big(\text{closed walks of length } L\Big)$$

where the walks must start and return at a designated root of the tree. See [McK81, p. 205–213] or [Kes59, p. 14] for more on the origins of ρ_d.

For the case of the Markoff graph mod p, the number of vertices is $p^2 \pm 3p$. Thus all we would have to show to deduce a qualitative result along the same lines as Theorem 1.1, with no explicit error term, is that the number of k-cycles is $o(p^2)$ for each fixed k. For intuition, imagine proving McKay's theorem by the method of moments. The moments are given by

$$\operatorname{tr}(A^L) = \sum \lambda_j^L$$

up to normalization by $p^2 \pm 3p$. On the other hand, there is a combinatorial interpretation. For $L \geqslant 1$, the trace $\operatorname{tr}(A^L)$ counts closed paths of length L in the graph:

$$\begin{aligned} \operatorname{tr}(A^L) &= \sum_j \sum_{k_1} \cdots \sum_{k_{L-1}} a_{jk_1} a_{k_1 k_2} \dots a_{k_{L-1} j} \\ &= \sum_{x \in M(\mathbb{F}_p)} \sum_{x \xrightarrow{L} x} 1 \end{aligned}$$

where the inner sum runs over paths of length L from x to x, and the outer sum runs over all vertices x. Changing the order of summation, we can rewrite this as

$$\operatorname{tr}(A^L) = \sum_w \#\{\text{fixed points of } g_w\}$$

In the summation, w is a (not necessarily reduced) word of length L in the Markoff moves m_1, m_2, m_3 and g_w is the corresponding element of the free product $\mathbb{Z}/2 *$

$\mathbb{Z}/2 * \mathbb{Z}/2$ with generators m_1, m_2, m_3. Note that if $g_w = I$ is the identity, then all of the $p^2 \pm 3p$ vertices are fixed points. These words therefore make a contribution of

(#length L paths beginning and ending at a root in a 3-regular tree)$(p^2 \pm 3p)$.

We divide by $p^2 \pm 3p$ for normalization, and the remaining path-count is exactly the corresponding Kesten–McKay moment. Our task is to show that the remaining contribution, made by words of length L that do not evaluate to the identity, is of a lower order of magnitude as $p \to \infty$.

3. Some examples and heuristics

To argue that the identity contributes the main term, we must study the fixed points of other words in the Markoff moves. Let $w = g_1 \cdots g_L$ be a reduced word of length L where each g_i is one of the Markoff moves m_1, m_2, m_3. Write the fixed point equation as

$$\begin{bmatrix} f(x,y,z) \\ g(x,y,z) \\ h(x,y,z) \\ x^2+y^2+z^2-xyz \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \\ 0 \end{bmatrix} \tag{3.1}$$

where f, g, and h are polynomials that can be computed by successively applying the moves that make up the word w. One might expect this system of four equations in only three unknowns to have no solutions, but there may be redundancy. Indeed, the system always has $(0,0,0)$ as a trivial solution. The extreme case is $w = 1$, for which the first three equations amount to $(x,y,z) = (x,y,z)$ and every point on the Markoff surface is fixed. For nontrivial words, we will use the special structure of the Markoff surface to show that there is at least one nontrivial constraint in addition to the equation $x^2 + y^2 + z^2 = xyz$. First, we consider a few examples of short words.

Proposition 3.1. — *(Fixed points of short words)*

(1) *The number of fixed points of a single Markoff move m_i is $p - 4 - (\frac{-1}{p})$, and in particular is at most p.*
(2) *A reduced word of length 2 has no fixed points.*
(3) *A reduced word of length 3 either has no fixed points or else is conjugate to a single Markoff move.*

Part (1) shows that the graph $\mathfrak{M}_p$ contains loops, but only at a small fraction of the vertices. This example also shows that it is possible for (3.1) to reduce to just one nontrivial constraint in addition to the Markoff equation. The fact that the words of length 1 together have only on the order of p fixed points has some importance for our main proofs and we will revisit it in Lemma 5.4. Part (2) shows that there are never multiple edges joining the same pair of vertices. Part (3) shows that the graph contains no triangles.

Proof of (1). — This count is given in [CGMP20, Lemma 2.3], noting that $p-4-(\frac{-1}{p})$ is $p-5$ when $p \equiv 1 \bmod 4$ and $p-3$ when $p \equiv 3 \bmod 4$. For the reader's convenience, we sketch a similar argument here. If $(x,y,z) = (x,y,xy-z)$, then the Markoff move m_3 connects the vertex (x,y,z) to itself. Substituting $z = xy - z$ into the Markoff equation gives

$$x^2 + y^2 + \left(\frac{xy}{2}\right)^2 = \frac{x^2y^2}{2}.$$

For each $y \in \mathbb{F}_p$, this is a quadratic equation for x, namely

$$(y^2-4)x^2 = (2y)^2$$

which has no solutions if $y^2 = 4$, a unique solution $x = 0$ in case $y = 0$, and otherwise has $1 + (\frac{y^2-4}{p})$ solutions. If $y = 0$, then the fixed point must be $(0,0,0)$, which is not part of our graph. Thus we remove it from the count and find that the number of solutions is

$$\sum_{y \neq \pm 2} \left(1 + \left(\frac{y^2-4}{p}\right)\right) - 1 - \left(\frac{-1}{p}\right) = p - 3 - \left(\frac{-1}{p}\right) + \sum_{y \neq \pm 2} \left(\frac{y^2-4}{p}\right).$$

The character sum can be evaluated by factoring $y^2 - 4$ as $(y-2)(y+2)$, changing variables to $u = y - 2$, and using $(\frac{u^{-1}}{p}) = (\frac{u}{p})$. Note that $v = 1 + 4/u$ assumes all values except 1 and 0 when u is restricted to $u \neq -4, 0$, so that

$$\sum_{y \neq \pm 2} \left(\frac{y^2-4}{p}\right) = \sum_{u \neq -4,0} \left(\frac{u}{p}\right)\left(\frac{u+4}{p}\right) = \sum_{v \neq 1,0} \left(\frac{v}{p}\right) = -1.$$

Our count becomes $p - 4 - (\frac{-1}{p})$ and the result follows. □

For any (x,y) solving $y^2 = x^2(y^2/4 - 1)$, taking $z = xy/2$ gives a point (x,y,z) connected to itself by m_3. In the same way, taking $x = yz/2$ or $y = xz/2$ gives points fixed by m_1 or m_2. All told, there are $3(p - 4 - (\frac{-1}{p}))$ vertices fixed by one of the generators. At each such vertex, there is a single loop. Note, as a special case of part (2), that only $(0,0,0)$ is fixed by multiple generators at once.

Proof of (2). — A word of length 2 has no fixed points. We stated before that the Markoff graph does not contain bigons – that is, multiple edges between the same pair of vertices – and a fixed point x of $m_i m_j$ is equivalent to a bigon between x and $m_j x$. It is easy to see why this does not occur. For example, if the moves m_3 and m_2 define the same edge starting from (x,y,z), then

$$(x, y, xy - z) = (x, xz - y, z).$$

Equivalently, $2y = xz$ and $2z = xy$. Thus $2y = x^2y/2$, which implies that either $y = 0$ or $x = \pm 2$. If $y = 0$, then $2z = xy = 0$ forces $z = 0$, and then the Markoff equation implies that x is also 0. Thus this case arises only for $(x,y,z) = (0,0,0)$, which is not part of our graph. On the other hand, the cases $x = \pm 2$ do not arise at all. Indeed, $2z = xy = \pm 2y$ implies $z = \pm y$. Substituting this into the Markoff equation gives

$$4 + 2y^2 = 2y^2$$

which cannot be. □

Proof of (3). — The words of length 3 are either $m_2m_3m_2$ or $m_2m_3m_1$, up to permuting the variables x, y, z. Note that $m_2m_3m_2$ is conjugate to m_3 since $m_2^{-1} = m_2$, and so it has the same number of fixed points as m_3. For the word $m_2m_3m_1$, all four equations impose nontrivial constraints and we will see that there are no solutions. Composing from left to right, we arrive at

$$\begin{bmatrix} (xz-y)(x(xz-y)-z)-x \\ xz-y \\ x(xz-y)-z \\ x^2+y^2+z^2 \end{bmatrix} = \begin{bmatrix} x \\ y \\ z \\ xyz \end{bmatrix}$$

The second equation implies $y = xz/2$, and substituting this in the third gives

$$2z = \frac{x^2 z}{2}.$$

Hence either $z = 0$ or $x^2 = 4$. If $z = 0$, then the other equations quickly lead us to the trivial solution $(x, y, z) = (0, 0, 0)$. Otherwise, we have $x = \pm 2$, and substituting this in the first equation shows that $z^2 = 4$. Hence any nonzero solutions must be of the form $(x, y, z) = (\pm 2, \pm 2, \pm 2)$. But no such triples solve the Markoff equation $x^2 + y^2 + z^2 = xyz$ since $12 \neq \pm 8 \bmod p$ for any $p \neq 2$. □

As an example involving a word of length 4, consider $m_2m_3m_2m_3$. The equation $f = x$ becomes vacuous because the word does not involve m_1. The remaining equations $g = y$ and $h = z$ can both be solved by taking $x = 0$. Taking $x = 0$ in the Markoff equation, we see that every solution of $y^2 + z^2 = 0$ leads to a fixed point. If $p \equiv 1 \bmod 4$, then -1 is a square mod p and any point $(0, y, \sqrt{-1}y)$ is fixed on the Markoff surface. Thus the system (3.1) can have on the order of p solutions even for a word that is not conjugate to any of the Markoff moves.

In all of these examples, there are on the order of p fixed points at the most. Now we present a heuristic suggesting why this trend should continue for longer words, so that the system (3.1) has only $O(p)$ solutions. Note first that applying a move such as $h \mapsto fg - h$ at most doubles the overall degree of the polynomials in the sense that, with respect to any of the variables x, y, or z,

$$\max(\deg(f), \deg(g), \deg(fg-h)) \leqslant 2\max(\deg(f), \deg(g), \deg(h)).$$

It could conceivably leave the degree the same if $\deg(h) \geqslant \deg(f) + \deg(g)$. In any case, for a word of length L, the final f, g, and h have degree at most 2^L in any of the variables x, y, z.

Fix $z \in \mathbb{F}_p$. We expect that (3.1) has only $O(1)$ solutions for x, y. It might happen that two of the equations are redundant, say $g = y$ and $h = z$, as for a single Markoff move. Thus we consider only $f = x$ together with the Markoff equation itself. The latter is quadratic in x and y, while the equation $f = x$ has degree at most 2^L. By Bézout's theorem, there are at most 2^{L+1} common solutions in an algebraic closure, and perhaps even fewer in the ground field $\mathbb{F}_p$ itself. However, it might happen for some values of z that the locus $f = x$ is contained entirely within $x^2 + y^2 + z^2 = xyz$. If there were no such z, we could conclude that the number of fixed points is at most $2^{L+1}p$. Instead, our bound will lead to $C^L p$ for some constant $C > 2$. In particular, the number of fixed points is at most $2^{17L+10}p$.

4. The Fricke–Klein trace identity and its consequences

Let $P_M \in \mathbb{Z}[x, y, z]$ denote the polynomial that defines the Markoff surface:

$$P_M(x, y, z) = x^2 + y^2 + z^2 - xyz.$$

Write F_2 for the free group on two generators X, Y. We first explain that the outer automorphism group $\mathrm{Out}(F_2)$ has a natural action on $\mathbb{C}^3$ by polynomial maps, *defined over* $\mathbb{Z}$, and moreover preserves the polynomial P_M. Although we are ultimately interested in solutions over $\mathbb{F}_p$, we use the complex numbers in this section in order to explain this action of $\mathrm{Out}(F_2)$.

Write $\mathrm{Hom}(F_2, \mathrm{SL}_2(\mathbb{C}))$ for the set of homomorphisms $\theta : F_2 \to \mathrm{SL}_2(\mathbb{C})$. Since conjugation preserves traces, we have a well-defined map on the quotient by SL_2-conjugacy:

$$\Phi : \mathrm{Hom}(F_2, \mathrm{SL}_2(\mathbb{C}))/\mathrm{SL}_2(\mathbb{C}) \to \mathbb{C}^3$$

given by

$$\Phi : \theta \mapsto (\mathrm{tr}\,\theta(X), \mathrm{tr}\,\theta(Y), \mathrm{tr}\,\theta(XY)).$$

By work of Fricke [Fri96] and Fricke–Klein [FK65] – paraphrased in modern language – Φ is an isomorphism of schemes, provided that the quotient $\mathrm{Hom}(F_2, \mathrm{SL}_2(\mathbb{C}))/\mathrm{SL}_2(\mathbb{C})$ is understood in the sense of geometric invariant theory. The group of automorphisms $\mathrm{Aut}(F_2)$ acts on $\mathrm{Hom}(F_2, \mathrm{SL}_2(\mathbb{C}))$ by composing $\theta : F_2 \to \mathrm{SL}_2(\mathbb{C})$ with $\sigma \in \mathrm{Aut}(F_2)$. This gives a well-defined action of outer automorphisms $\mathrm{Out}(F_2)$ on the quotient of $\mathrm{Hom}(F_2, \mathrm{SL}_2)$ by conjugation and hence, via Φ, an action on $\mathbb{C}^3$ by polynomial maps. These polynomial maps are defined over $\mathbb{Z}$, as one verifies on generators of $\mathrm{Out}(F_2)$. We give examples below, and in the process see how the Markoff moves act in this representation.

To see that this action preserves P_M, we use Fricke's trace identity. This states that for matrices $A, B \in \mathrm{SL}_2(\mathbb{C})$

$$\mathrm{tr}(A)^2 + \mathrm{tr}(B)^2 + \mathrm{tr}(AB)^2 = \mathrm{tr}(A)\,\mathrm{tr}(B)\,\mathrm{tr}(AB) + \mathrm{tr}([A, B]) + 2,$$

where $[A, B]$ denotes $ABA^{-1}B^{-1}$. This identity can be proved using the Cayley–Hamilton theorem and other properties of the trace. See, for instance, [Aig13, Proposition 4.3].

We also note that the action of $\mathrm{Out}(F_2)$ on $\mathrm{Hom}(F_2, \mathrm{SL}_2(\mathbb{C}))/\mathrm{SL}_2(\mathbb{C})$ preserves the function $\theta \mapsto \mathrm{tr}([\theta(X), \theta(Y)])$. We rely here on an important fact about F_2 which has no counterpart for free groups of higher rank: given a basis X, Y for F_2, every outer automorphism preserves the conjugacy class of $XYX^{-1}Y^{-1}$ up to inversion [Nie17]. This, together with the fact that $\mathrm{tr}(A) = \mathrm{tr}(A^{-1})$ for $A \in \mathrm{SL}_2(\mathbb{C})$, implies that $\mathrm{tr}([\theta(X), \theta(Y)])$ is an invariant function for $\mathrm{Out}(F_2)$. Putting this fact together with the Fricke–Klein identity implies that the polynomial action of $\mathrm{Out}(F_2)$ on $\mathbb{C}^3$ preserves the polynomial P_M.

Since $\mathrm{Out}(F_2)$ acts on $\mathbb{C}^3$ by polynomial maps defined over $\mathbb{Z}$, and preserves the polynomial P_M, also defined over $\mathbb{Z}$, we obtain by base change an action of $\mathrm{Out}(F_2)$ by polynomial maps on the Markoff surface $M(\mathbb{F}_p)$ for any prime p.

We now explain the relationship between $\mathrm{Out}(F_2)$ and $\mathrm{GL}_2(\mathbb{Z})$. Any automorphism of F_2 preserves the commutator subgroup, and in particular $\mathrm{Out}(F_2)$ acts on the abelianization

$$F_2^{\mathrm{ab}} = F_2/[F_2, F_2] \cong \mathbb{Z}^2$$

which is a free abelian group of rank 2. This action induces a map

$$\mathrm{Out}(F_2) \to \mathrm{Aut}(\mathbb{Z}^2) = \mathrm{GL}_2(\mathbb{Z})$$

and it is a theorem of Nielsen that this map is an isomorphism (see, for instance, [Aig13, Theorem 6.24] or [Nie17] for the original article). Thus $\mathrm{GL}_2(\mathbb{Z})$ acts on the Markoff surface via the action of $\mathrm{Out}(F_2)$.

To show that the Markoff generators are induced by the $\mathrm{Out}(F_2)$ action, and find specific matrix representatives for them, we argue as follows. Given an element $\theta \in \mathrm{Hom}(F_2, \mathrm{SL}_2(\mathbb{C}))$, write $A = \theta(X)$ and $B = \theta(Y)$ in $\mathrm{SL}_2(\mathbb{C})$.

By the Cayley–Hamilton theorem, A solves its own characteristic polynomial, so

$$A^2 - \mathrm{tr}(A)A + 1 = 0.$$

Multiplying by BA^{-1}, we obtain

$$BA - \mathrm{tr}(A)B + BA^{-1} = 0.$$

Taking the trace of both sides gives

$$\mathrm{tr}(BA) = \mathrm{tr}(A)\,\mathrm{tr}(B) - \mathrm{tr}\left(BA^{-1}\right).$$

This has the same form as a Markoff move on the vector of traces

$$(\mathrm{tr}(A), \mathrm{tr}(B), \mathrm{tr}(BA)),$$

with the other solution for the third coordinate being $\mathrm{tr}(BA^{-1})$. To keep the third matrix equal to the product of the first two, we use $\mathrm{tr}(A^{-1}) = \mathrm{tr}(A)$ to rewrite the vector of traces as

$$\left(\mathrm{tr}(A), \mathrm{tr}(B), \mathrm{tr}\left(BA^{-1}\right)\right) = \left(\mathrm{tr}(A), \mathrm{tr}\left(B^{-1}\right), \mathrm{tr}\left(AB^{-1}\right)\right)$$

Thus the third move m_3 arises from the element of $\mathrm{Aut}(F_2)$ that sends X to X and Y to Y^{-1}, which corresponds to the matrix

$$[m_3] = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

Equally well, since we work in PGL_2, m_3 could be represented by $\left[\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, which would correspond to writing the trace vector as

$$\left(\mathrm{tr}(A), \mathrm{tr}(B), \mathrm{tr}\left(BA^{-1}\right)\right) = \left(\mathrm{tr}\left(A^{-1}\right), \mathrm{tr}(B), \mathrm{tr}\left(A^{-1}B\right)\right)$$

by cyclicity of trace. In the same way, we find that the first move m_1 arises from the element of $\mathrm{Aut}(F_2)$ that sends (X, Y) to (XY^2, Y^{-1}). The second move arises from the element of $\mathrm{Aut}(F_2)$ that sends X to X^{-1} and Y to X^2Y. In terms of $\mathrm{GL}_2(\mathbb{Z})$, the Markoff moves therefore correspond to the matrices

$$[m_1] = \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix},\ [m_2] = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix},\ [m_3] = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{4.1}$$

In particular, the group generated by these matrices acts on the Markoff surface. One also has permutations of the three coordinates. For instance, the transposition τ_{23} acts by

$$(\mathrm{tr}(A), \mathrm{tr}(B), \mathrm{tr}(AB)) \mapsto (\mathrm{tr}(A), \mathrm{tr}(AB), \mathrm{tr}(B)) = \left(\mathrm{tr}(A), \mathrm{tr}\left(A^{-1}B^{-1}\right), \mathrm{tr}\left(B^{-1}\right)\right)$$

so that, in matrix form,

$$[\tau_{23}] = \begin{bmatrix} 1 & -1 \\ 0 & -1 \end{bmatrix}.$$

Likewise, $[\tau_{13}] = \left[\begin{smallmatrix} 1 & 0 \\ 1 & -1 \end{smallmatrix}\right]$. As a consistency check, one must have $[m_2] = [\tau_{23}][m_3][\tau_{23}]$ and $[m_1] = [\tau_{13}][m_3][\tau_{13}]$ up to sign.

Let $G \leqslant \mathrm{PGL}_2(\mathbb{Z})$ be the group generated by $[m_1], [m_2], [m_3]$. As before, these correspond to the Markoff moves

$$\begin{aligned} m_1(x, y, z) &= (yz - x, y, z), \\ m_2(x, y, z) &= (x, xz - y, z), \\ m_3(x, y, z) &= (x, y, xy - z). \end{aligned} \tag{4.2}$$

As an abstract group, $G \cong \mathbb{Z}/2\mathbb{Z} * \mathbb{Z}/2\mathbb{Z} * \mathbb{Z}/2\mathbb{Z}$ with the m_i the generators of the factors in the free product [ÈH74, Theorem 1], [CL09, Theorem 3.1]. To conclude this section, we note a property of G that will be used in the sequel:

Lemma 4.1. — *The only torsion elements in the Markoff group are the Markoff moves themselves and their conjugates.*

Proof. — This follows from the fact that finite-order elements of a free product are conjugate into one of the factors, which in turn follows for example from Kurosh's theorem (see [MKS04, Corollary 4.9.1]). □

5. Bounds for the number of fixed points of words

The goal of this section is to prove the following Theorem 5.1.

Theorem 5.1. — *If $g \in G \leqslant \mathrm{PGL}_2(\mathbb{Z})$ is not the identity, and is represented by a matrix $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathrm{GL}_2(\mathbb{Z})$ with $\max(|a|, |b|, |c|, |d|) < (p/128)^{1/8}$, then g has at most $1024p \max(|a|, |b|, |c|, |d|)^8$ fixed points on $M(\mathbb{F}_p)$.*

The exponent 8 on $\max(|a|, |b|, |c|, |d|)$ can likely be replaced by 1. Similarly, the assumption $\max(|a|, |b|, |c|, |d|) \leqslant (p/128)^{1/8}$ can most likely be loosened. To keep the arguments to their simplest and most readable, and since the bound above is enough for our qualitative result, we chose not to pursue the optimal constants here.

After raising g to a small power, three natural cases arise, and we will give a different bound in each case.

Lemma 5.2. — *For any element $g \in \mathrm{GL}_2(\mathbb{Z})$, there is a power $1 \leqslant K \leqslant 8$ of g such that one of the following holds.*

1. *All the entries of g^K have absolute value at least 2.*
2. *g^K is a torsion element of $\mathrm{GL}_2(\mathbb{Z})$. In this case, g is already torsion.*

(3) *g^K is one of the following types of matrices*

$$\pm\begin{bmatrix}1 & n\\ 0 & 1\end{bmatrix}, \pm\begin{bmatrix}1 & 0\\ n & 1\end{bmatrix}, \quad n \in \mathbb{Z} - \{0\}. \tag{5.1}$$

Proof. — We first show that one may take $K \leqslant 4$ in the case $\det(g) = 1$. To avoid considering the case $\det(g) = -1$ separately, we replace g by g^2 and double K if necessary. Assuming $\det(g) = 1$, the Cayley–Hamilton theorem implies

$$g^2 - \operatorname{tr}(g)g + I = 0.$$

If $\operatorname{tr}(g) = 0$, we then have $g^4 = I$ so that g is torsion. If $\operatorname{tr}(g) = \pm 1$, then multiplying by g gives

$$g^3 = g(\pm g - I) = \pm(\pm g - I) - g = \mp I$$

and hence $g^6 = I$. Therefore g is torsion if $|\operatorname{tr}(g)| < 2$. Otherwise, $|\operatorname{tr}(g)| \geqslant 2$ and we use $ad - bc = 1$ to write

$$\begin{aligned} g &= \begin{bmatrix}a & b\\ c & d\end{bmatrix}\\ g^2 &= \begin{bmatrix}a(a+d)-1 & b(a+d)\\ c(a+d) & d(a+d)-1\end{bmatrix}\\ g^4 &= \begin{bmatrix}(a(a+d)-1)^2 + bc(a+d)^2 & b(a+d)((a+d)^2-2)\\ c(a+d)((a+d)^2-2) & (d(a+d)-1)^2 + bc(a+d)^2\end{bmatrix}\end{aligned}$$

If $bc = 0$, then $ad = 1$ and g must be of the form (5.1). Otherwise, we have $|b| \geqslant 1$, $|c| \geqslant 1$, and $(a+d)^2 \geqslant 4$. It follows that all entries of g^4 are at least 2 in absolute value (moreover, at least 3). The entries of g^2 might not be, for instance if $a = 0$. □

5.1. Fixed points of generic elements of G

The "generic" case is when all the entries of h have absolute value $\geqslant 2$. In this case, we use the following bound of Cerbu–Gunther–Magee–Peilen ([CGMP20, Lemma 3.9]).

Lemma 5.3. — *(Cerbu–Gunther–Magee–Peilen) If $g = \left[\begin{smallmatrix}a & b\\ c & d\end{smallmatrix}\right] \in \mathrm{GL}_2(\mathbb{Z})$ has*

$$|a|, |b|, |c|, |d| \geqslant 2,$$

then g has fewer than $2p(|a| + |b| + \big||d| - |c|\big|)$ fixed points on $M(\mathbb{F}_p)$.

We refer to [CGMP20] for the proof. The assumption that all the entries have absolute value at least 2 makes it possible to implement a rigorous version of the heuristic in Section 3.

5.2. Fixed points of torsion elements of G

The next Lemma 5.4 bounds the number of fixed points of torsion elements of G.

LEMMA 5.4. — *If g is a non-identity torsion element of $G \leqslant \mathrm{PGL}_2(\mathbb{Z})$ then g has fewer than p fixed points on $M(\mathbb{F}_p)$.*

Proof. — By Lemma 4.1, any non-identity torsion g is conjugate in G to a Markoff move m_i. Therefore it has the same number of fixed points as m_i on $M(\mathbb{F}_p)$. This number is either $p-3$ or $p-5$, by part (1) of Proposition 3.1, and the result follows. □

5.3. Fixed points of standard parabolic elements in G

Finally, we estimate the number of fixed points of the standard parabolic elements in the list (5.1).

PROPOSITION 5.5. — *If $g \in \mathrm{GL}_2(\mathbb{Z})$ is one of the matrices $\pm\left[\begin{smallmatrix}1 & n\\ 0 & 1\end{smallmatrix}\right], \pm\left[\begin{smallmatrix}1 & 0\\ n & 1\end{smallmatrix}\right]$ where $0 \neq |n| < p$, then g has fewer than $2|n|p$ fixed points on $M(\mathbb{F}_p)$.*

Consider the "rotation" $\mathrm{rot} = \left[\begin{smallmatrix}1 & 1\\ 0 & 1\end{smallmatrix}\right]$, following the notation of Bourgain–Gamburd–Sarnak [BGS16, p. 3]. From the matrices in Section 4, especially,

$$[m_3] = \begin{bmatrix}1 & 0\\ 0 & -1\end{bmatrix}, \quad [\tau_{23}] = \begin{bmatrix}1 & -1\\ 0 & -1\end{bmatrix}$$

we see that $\mathrm{rot} = [\tau_{23}]\circ[m_3]$. Again, working in PGL_2, this is to be understood modulo sign. This element $\mathrm{rot} : (x,y,z) \to (x, xy-z, y)$ thus combines a Markoff move on the third coordinate with a transposition of the second and third coordinates. Although rot itself is not a word in the Markoff moves (its top-right entry is 1, whereas words in the Markoff matrices have even off-diagonal entries), we do have

$$\mathrm{rot}^2 = m_2 \circ m_3 \tag{5.2}$$

as one sees using $\mathrm{rot} = \tau_{23} \circ m_3$ and $m_2 = \tau_{23} \circ m_3 \circ \tau_{23}$, or observing that both sides send (x,y,z) to $(x, x(xy-z)-y, xy-z)$. We will prove Proposition 5.5 by carefully examining the orbit structure of rot on $M(\mathbb{F}_p)$. For any fixed $x \in \mathbb{F}_p$, the Markoff equation defines a conic section

$$C(x): \quad y^2 - xyz + z^2 = -x^2. \tag{5.3}$$

Since rot does not change the first coordinate of (x,y,z), it preserves each of these conics. Its action is given by

$$\mathrm{rot}\begin{bmatrix}y\\ z\end{bmatrix} = \begin{bmatrix}z\\ xz-y\end{bmatrix} = \begin{bmatrix}0 & 1\\ -1 & x\end{bmatrix}\cdot\begin{bmatrix}y\\ z\end{bmatrix} \tag{5.4}$$

If $x = \pm 2$, then $C(x)$ degenerates to

$$C(\pm 2): \quad \left(\frac{y \mp z}{2}\right)^2 = -1$$

which is either empty if $p \equiv 3 \bmod 4$ or a pair of lines if $p \equiv 1$. Hence

$$\#C(\pm 2) = p\left(1 + \left(\frac{-1}{p}\right)\right).$$

For the remaining values of x, we have

$$\#C(x) = p - \left(\frac{x^2-4}{p}\right)$$

as we will see by an explicit parametrization. It can also be shown by direct manipulations with the Legendre symbol.

One can think of the conic sections $C(x)$ either as ellipses or hyperbolas modulo p according to whether x^2-4 is a square. Following [BGS16], we say $x \in \mathbb{F}_p$ is *hyperbolic* if x^2-4 is a nonzero square in $\mathbb{F}_p$. We say $x \in \mathbb{F}_p$ is *elliptic* if x^2-4 is nonzero and not a square. We say $x \in \mathbb{F}_p$ is *parabolic* if $x^2-4=0$, i.e. $x = \pm 2$. Note that the parabolic case only arises for $p \equiv 1 \bmod 4$, and that the conic section in such a case is not a parabola but something degenerate. The behaviour of rot on $C(x)$ was described by Bourgain, Gamburd, and Sarnak in [BGS16] using this classification of values of x. They state their results for the surface $X^2+Y^2+Z^2 = 3XYZ$, although in many of the proofs they use the same normalization $x^2+y^2+z^2 = xyz$ as in the present article. The two surfaces are equivalent over $\mathbb{F}_p$ for $p \geqslant 5$ by a scaling $(X,Y,Z) = (x,y,z)/3$, and we review the corresponding parts of [BGS16] for the reader's convenience.

A convenient change of variable toward parametrizing $C(x)$ is

$$x = \xi + \xi^{-1} \tag{5.5}$$

where $\xi \neq 0$ lies in $\mathbb{F}_p$ if x^2-4 is a square, and otherwise in a quadratic extension $\mathbb{F}_{p^2}$. Let

$$\kappa = \kappa(x) = \frac{x^2}{x^2-4} = \left(\frac{\xi+\xi^{-1}}{\xi-\xi^{-1}}\right)^2. \tag{5.6}$$

Then

$$(x,y,z) = \left(x, t + \frac{\kappa}{t}, t\xi + \frac{\kappa}{t\xi}\right)$$

solves the Markoff equation for any $t \neq 0$. Note that multiplying (5.5) by ξ gives $\xi^2 - x\xi + 1 = 0$, and this equation simplifies the verification that $(x, t+\kappa t^{-1}, t\xi + \kappa t^{-1}\xi^{-1})$ solves the Markoff equation. The action of rot is to multiply the parameter t by ξ^{-1}. Indeed, from the definition $x = \xi + \xi^{-1}$ and $\mathrm{rot}(y,z) = (xy-z, y)$, we calculate that

$$\begin{aligned}
\mathrm{rot}\left(t + \frac{\kappa}{t}, t\xi + \frac{\kappa}{t\xi}\right) &= \left(xt + \frac{x\kappa}{t} - t\xi - \frac{\kappa}{t\xi}, t + \frac{\kappa}{t}\right) \\
&= \left((\xi+\xi^{-1})t + (\xi+\xi^{-1})\frac{\kappa}{t} - t\xi - \frac{\kappa}{t\xi}, t + \frac{\kappa}{t}\right) \\
&= \left(t\xi^{-1} + \frac{\kappa}{t\xi^{-1}}, t\xi^{-1}\xi + \frac{\kappa}{t\xi^{-1}\xi}\right)
\end{aligned}$$

that is, t has been multiplied by ξ^{-1}. These considerations can be summarized in the following Lemma 5.6, due to Bourgain–Gamburd–Sarnak [BGS16].

LEMMA 5.6. — *(Bourgain–Gamburd–Sarnak)*

- *(*[BGS16, Lemma 3]*) Let x be parabolic. If $p \equiv 3 \bmod 4$ then $C(x)$ is empty. If $p \equiv 1 \bmod 4$ then $C(x)$ consists of two lines. Letting i be such that $i^2 \equiv -1 \bmod p$, the conic sections are parametrized by*

$$C(2) = (2, t, t \pm 2i)$$
$$C(-2) = (-2, t, -t \pm 2i).$$

The action of rot *is given by*

$$\mathrm{rot}(2, t, t \pm 2i) = (2, t \pm 2i, t \pm 4i),$$
$$\mathrm{rot}(-2, t, -t \pm 2i) = (-2, -t \pm 2i, t \mp 4i).$$

- *(*[BGS16, Lemma 4]*) Let x be hyperbolic. Write $x = w + w^{-1}$ with $w \in \mathbb{F}_p^*$. Let*

$$\kappa(x) = \frac{x^2}{x^2 - 4}.$$

Then $C(x)$ is parametrized by $\mathbb{F}_p^*$ via the map

$$t \in \mathbb{F}_p^* \mapsto \left(x, t + \frac{\kappa(x)}{t}, tw + \frac{\kappa(x)}{tw} \right).$$

As a consequence, $|C(x)| = p - 1$. After this identification, rot *acts on $C(x) \cong \mathbb{F}_p^*$ by multiplication by w^{-1}.*

- *(*[BGS16, Lemma 5]*) Let x be elliptic. Write $x = v + v^{-1}$ where $v \in \mathbb{F}_{p^2} - \mathbb{F}_p$ and $v^{p+1} = 1$. Let $\kappa(x) = \frac{x^2}{x^2-4}$. Let $E(x) \subset \mathbb{F}_{p^2}$ be the set of t such that $t^{p+1} = \kappa(x)$. Then $C(x)$ is parametrized by $E(x)$ via the map*

$$t \in E(x) \mapsto \left(x, t + \frac{\kappa(x)}{t}, tv + \frac{\kappa(x)}{tv} \right).$$

As a consequence, $|C(x)| = p + 1$. After this identification, rot *acts on $C(x) \cong E(x)$ by multiplication by v^{-1}.*

Proof of Proposition 5.5. By multiplying by $-I$, taking inverses, or conjugating by $\left[\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right]$, all the matrices of the proposition can be brought into the form $\left[\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right]$ where $n > 0$. None of these operations change the number of fixed points of g on $M(\mathbb{F}_p)$, or the bound for the number of fixed points claimed in the lemma. So it suffices to prove the proposition for $\left[\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right]$. This matrix acts on $M(\mathbb{F}_p)$ by rot^n. We split up the fixed points of rot^n depending on whether they belong to $C(x)$ with x parabolic, hyperbolic, or elliptic. If x is parabolic, Lemma 5.6 implies that for $n < p$, rot^n has no fixed points on $C(x)$. For fixed hyperbolic x, let $x = w + w^{-1}$ as in Lemma 5.6. Lemma 5.6 implies that rot^n has a fixed point in $C(x)$ if and only if $w^n = 1$, and this happens if and only if every element of $C(x)$ is fixed by rot^n. The number of fixed points of rot^n contained in $C(x)$ with x hyperbolic is therefore bounded by

$$\sum_{w \in \mathbb{F}_p^* : w^n = 1} \left| C\left(w + w^{-1}\right) \right| = (p-1) \left| \left\{ w \in \mathbb{F}_p^* : w^n = 1 \right\} \right| \leqslant (p-1)n.$$

When x is elliptic, a similar argument using Lemma 5.6 shows that the number of fixed points of rot^n contained in $C(x)$ is bounded by

$$\sum_{\substack{v \in \mathbb{F}_{p^2} - \mathbb{F}_p: \\ v^{p+1}=1,\, v^n=1}} \left|C\left(v + v^{-1}\right)\right| = (p+1)\left|\left\{v \in \mathbb{F}_{p^2} - \mathbb{F}_p : v^{p+1} = 1, v^n = 1\right\}\right| \leqslant (p+1)n.$$

Therefore, when $n < p$, adding our previous bounds together, rot^n has at most $2pn$ fixed points on $M(\mathbb{F}_p)$. This concludes the proof. □

We have used the bound that there are at most n solutions to $w^n = 1$, as a polynomial cannot have more roots than its degree. For many values of n, the only solution is $w = 1$. Extra solutions arise only if n and $p-1$ have a common factor. This is related to the difficulties encountered in [BGS16] when $p^2 - 1$ has many factors.

5.4. Proof of Theorem 5.1.

Consider any $g \neq 1$ in the Markoff group $G \leqslant \mathrm{PGL}_2(\mathbb{Z})$. Let $h = g^K$ where $K \leqslant 8$ is the power from Lemma 5.2. Any fixed point of g is also a fixed point of its powers, so it suffices to bound the number of fixed points of h on $M(\mathbb{F}_p)$. Note that if g is represented by $\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$ in $\mathrm{GL}_2(\mathbb{Z})$, and h by $\left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$, then

$$\max(|A|, |B|, |C|, |D|) \leqslant 128 \max(|a|, |b|, |c|, |d|)^8.$$

Indeed, we have

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^2 = \begin{bmatrix} a^2 + bc & ab + bd \\ ac + dc & d^2 + bc \end{bmatrix}$$

so that the entries of g^2 are at most $2\max(|a|, |b|, |c|, |d|)^2$. By repeated squaring, the entries of g^{2^n} are at most $2^{2^n-1}\max(|a|, |b|, |c|, |d|)^{2^n}$. In particular, for the eighth power we have the bound $128\max(|a|, |b|, |c|, |d|)^8$, as claimed. We have assumed that $\max(|a|, |b|, |c|, |d|) < (p/128)^{1/8}$ for the express purpose of ensuring that the entries of h are less than p. Thus we may apply Proposition 5.5. This implies that if h is a standard parabolic element, then it has at most

$$2p\max(|A|, |B|, |C|, |D|) \leqslant 256p\max(|a|, |b|, |c|, |d|)^8$$

fixed points. Otherwise, one of the other cases in Lemma 5.2 pertains. In the torsion case, g has at most p fixed points by Lemma 5.4. This is smaller than the previous bound because

$$1 \leqslant \max(|A|, |B|, |C|, |D|)$$

since the entries of h are integers, not all zero. In the generic case, Lemma 5.3 shows that the number of fixed points of h is at most

$$2p(|A| + |B| + ||C| - |D||) \leqslant 8p\max(|A|, |B|, |C|, |D|).$$

Thus in all cases, h has at most

$$8p\max(|A|, |B|, |C|, |D|) \leqslant 1024p\max(|a|, |b|, |c|, |d|)^8$$

fixed points, and therefore so does g.

6. Proof of the Kesten–McKay Law

In this section, we prove Theorem 1.1. Let A be the adjacency matrix of the Markoff graph and λ_j its eigenvalues. By definition of the empirical measure $\mu_p = \sum \delta_{\lambda_j}$, we have

$$\int x^L d\mu_p(x) = \sum_j \lambda_j^L = \operatorname{tr}\left(A^L\right).$$

On the other hand, expanding the trace as in Section 2 gives

$$\operatorname{tr}\left(A^L\right) = \sum_{j_1} \cdots \sum_{j_L} a_{j_1,j_2}\, a_{j_2,j_3}\, \ldots\, a_{j_L,j_1}$$

The product $a_{j_1,j_2}\, a_{j_2,j_3}\, \ldots\, a_{j_L,j_1}$ is 0 unless there is a cycle

$$j_1 \to j_2 \to \ldots \to j_L \to j_1$$

where each arrow represents a Markoff move m_1, m_2, or m_3. In such a case the product is 1 and the vertex labeled j_1 is fixed by some word of length L. The trace is obtained by summing over all words

$$\operatorname{tr}(A^L) = \sum_w \operatorname{Fix}(w) = \sum_{i_1} \cdots \sum_{i_L} \operatorname{Fix}(m_{i_1} \cdots m_{i_L})$$

where $\operatorname{Fix}(w)$ denotes the number of fixed points of w acting on $M(\mathbb{F}_p)$, and the indices $i_1, \ldots, i_L$ take the values 1, 2, 3. The words that reduce to the identity fix all of $M(\mathbb{F}_p)$ and contribute the main term:

$$\sum_{\substack{j_1,\ldots j_L \text{s.t.} \\ m_{j_1}\cdots m_{j_L}=1}} |M(\mathbb{F}_p)| = |M(\mathbb{F}_p)| \int x^L d\rho_3(x) = \left(p^2 \pm 3p\right) \int x^L d\rho_3$$

From the combinatorial interpretation noted in Section 2, the Kesten–McKay moment $\int x^L d\rho_3$ is exactly this count of paths in a tree returning to the starting point. We will use Theorem 5.1 to show that the remaining words make a negligible contribution, together with the following preparations. If

$$g = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

is a word of length L in the generators m_1, m_2, m_3, then the entries a, b, c, d are at most exponential in L. As an explicit upper bound, we have

Proposition 6.1. — *The entries of a word of length L in the Markoff moves m_1, m_2, m_3 are at most 4^L in absolute value.*

Proof. — The generators themselves have entries of absolute value at most 2, namely

$$[m_1] = \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}, \quad [m_2] = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}, \quad [m_3] = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

from equation (4.1). This confirms the base case $L = 1$ (and would even allow a better exponential rate than 4^L). For the induction step, consider

$$\begin{bmatrix} a_{k+1} & b_{k+1} \\ c_{k+1} & d_{k+1} \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} a_k & b_k \\ c_k & d_k \end{bmatrix} = \begin{bmatrix} aa_k + bc_k & ab_k + bd_k \\ ca_k + dc_k & cb_k + dd_k \end{bmatrix}$$

We have $\max(|a|,|b|,|c|,|d|) \leqslant 2$ from the base case, $\max(|a_k|,|b_k|,|c_k|,|d_k|) \leqslant 4^k$ from the induction hypothesis, and therefore

$$\max(|a_{k+1}|,|b_{k+1}|,|c_{k+1}|,|d_{k+1}|) \leqslant 2\cdot 4^k + 2\cdot 4^k = 4^{k+1}.$$

□

Corollary 6.2. — *There is an absolute exponent $\alpha > 0$ such that if $g \in \mathrm{GL}_2(\mathbb{Z})$ with $|\operatorname{tr}(g)| > 2$ is a word of length L in the matrices representing the Markoff moves m_1, m_2, m_3, then g has at most $e^{\alpha L}p$ fixed points.*

Proof. — By the previous Proposition 6.1, the entries of g are at most 4^L in absolute value. Combining this with Theorem 5.1, provided L is small enough that $4^L < (p/128)^{1/8}$ we find that the number of fixed points of g is at most

$$1024p\left(4^L\right)^8 = p2^{16L+10}.$$

Thus we can take $\alpha = 26\log 2 = 18.0218\ldots$ and have the result for all $L \geqslant 1$ obeying $4^L < (p/128)^{1/8}$. For larger L, note that the conclusion holds trivially once $e^{\alpha L}p > p^2 + 3p$, there being at most p^2+3p points on the Markoff surface. If L is so large that $4^L \geqslant (p/128)^{1/8}$, then $e^{\alpha L}p \geqslant (p/128)^{\alpha/16\log 2}p$. This can be made to exceed p^2+3p for all $p \geqslant 5$ by taking α large enough. □

There are at most 3^L words $m_{j_1}\cdots m_{j_L}$ of length L since each index is either 1, 2, or 3. Using the previous corollary over each of these terms leads to

$$\sum_{i_1}\cdots\sum_{i_L}\operatorname{Fix}(m_{i_1}\cdots m_{i_L}) \leqslant 3^L \times \left(p\times 2^{16L+10}\right)$$

where the sum is over the remaining words, that is, those that do not reduce to the identity. Combining this with the main term from the words that do reduce to 1, we have

$$\operatorname{tr}\left(A^L\right) = |M(\mathbb{F}_p)|\int x^L\rho_3(x)dx + O\left(C^Lp\right)$$

where $C = 3\times 2^{16} = 196608$ and the implicit constant could be taken as $2^{10} = 1024$ independent of both p and L. We have $|M(\mathbb{F}_p)| = p^2 \pm 3p$, and normalizing by p^2 gives

$$\int x^L d\mu_p(x) = \int x^L\rho_3(x)dx + O\left(\frac{C^L}{p}\right).$$

The error term is negligible provided that

$$L - \frac{\log p}{\log C} \to -\infty.$$

This allows for $L \sim c\log p$ for a sufficiently small $c > 0$, namely

$$c < \frac{1}{\log C} = 0.082041\ldots$$

and one could also take, for instance, $L \sim \frac{1}{\log C}\log p - \sqrt{\log p}$.

7. Proof of Corollary 1.3

Corollary 1.3 compares the measure of an interval under the empirical distribution of eigenvalues as against the limiting Kesten–McKay law, whereas Theorem 1.1 gives information about moments. A natural bridge between these is to approximate the given interval's indicator function by polynomials. If we had estimates for the Fourier transform $\sum_j e^{i\xi\lambda_j}$, then we could try to bound the discrepancy using the Erdős –Turán inequality (see [Mon94, Corollary 1.1] or the original articles [ET48a, ET48b]). However, Theorem 1.1 only allows us to take moments of the form $\sum_j \lambda_j^L$ with L on the order of $\log p$. There are standard arguments to pass from moments to discrepancy, and in particular Gamburd–Jakobson–Sarnak faced the same problem in a setting very close to ours [GJS99]. What we state below as Lemma 7.1 is a summary of facts given in equations (55) and (57) of [GJS99, Proof of Theorem 1.3]. It is based on Selberg polynomials after Selberg [Sel91, p. 213–219] and Vaaler [Vaa85]. We also recommend Montgomery's treatment [Mon94, p. 5–15].

Lemma 7.1. — *(Gamburd–Jakobson–Sarnak, after Selberg and Vaaler) For any interval $I \subseteq [-1,1]$ and $m \in \mathbb{N}$, there exist polynomials $f_m^{\pm}$ of degree m such that*

- $f_m^- \leqslant \chi_I \leqslant f_m^+$ *on* $[-1,1]$,
- *There is an absolute constant $B > 0$, independent of I, such that the coefficients of $f_m^{\pm}$ have absolute value $\leqslant B^m$.*
- $$\int_{-\frac{1}{\sqrt{2}}}^{\frac{1}{\sqrt{2}}} (f_m^+ - f_m^-)(y)dy = O\left(\frac{1}{m}\right).$$

Proof of Corollary 1.3. The Markoff eigenvalues lie in $[-3,3]$, so we first rescale so that Lemma 7.1 applies. Given any subinterval J of $[-3,3]$, let

$$I = K^{-1}J \subseteq \left[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]$$

where $K = 3\sqrt{2}$. Let $f_m^{\pm}$ be the polynomials from Lemma 7.1 applied to I, where m will be a small multiple of $\log p$, and let $g_m^{\pm}(x) = f_m^{\pm}(x/K)$ be the rescaled polynomials on $[-3,3]$. Then we can write

$$g_m^{\pm}(y) = \sum_{i=0}^{m} a_{m,i}^{\pm} y^i,$$

where $|a_{m,i}^{\pm}| \leqslant B^m K^{-i} \leqslant B^m$, noting that $K > 1$. We write μ_∞ for the measure with density $\rho_3(x)$ and μ_p for the eigenvalue counting measure (normalized to have total mass 1).

From Lemma 7.1, we have

$$\int g_m^- d\mu_p \leqslant \mu_p(J) \leqslant \int g_m^+ d\mu_p. \tag{7.1}$$

By Theorem 1.1,

$$\int x^i d\mu_p = \int x^i d\mu_\infty + O\big(p^{-1}C^i\big)$$

Therefore

$$\int g_m^{\pm} d\mu_p = \int g_m^{\pm} d\mu_\infty + O\left(\sum_{i=0}^{m} \left|a_{m,i}^{\pm}\right| C^i p^{-1}\right)$$

Since the coefficients $a_{m,i}^{\pm}$ are at most B^m, we have

$$\int g_m^{\pm} d\mu_p = \int g_m^{\pm} d\mu_\infty + O\left((BC)^m p^{-1}\right)$$

Therefore we can replace μ_p by μ_∞ in (7.1):

$$\int g_m^{-} d\mu_\infty + O\left((BC)^m p^{-1}\right) \leqslant \mu_p(J) \leqslant \int g_m^{+} d\mu_\infty + O\left((BC)^m p^{-1}\right).$$

Using $g_m^- \leqslant \chi_J \leqslant g_m^+$ gives

$$\int g_m^{-} d\mu_\infty \leqslant \mu_\infty(J) \leqslant \int g_m^{+} d\mu_\infty$$

and since the Kesten–McKay density ρ_3 is bounded, Lemma 7.1 also implies that

$$\int (g_m^+ - g_m^-) d\mu_\infty = O\left(\frac{1}{m}\right).$$

It follows that

$$|\mu_p(J) - \mu_\infty(J)| \lesssim \frac{1}{m} + \frac{(BC)^m}{p}.$$

If we choose $m = \lfloor c \log p \rfloor$ for a small constant $c > 0$, we obtain

$$|\mu_p(J) - \mu_\infty(J)| \lesssim \frac{1}{c \log p} + p^{-1 + c \log(BC)}$$

By choosing $c > 0$ such that $-1 + c \log(BC) < 0$, the $\frac{1}{c \log p}$ is eventually the larger term above. Thus

$$\mu_p(J) = \int_J \rho_3(x) dx + O\left(\frac{1}{\log p}\right)$$

as required. □

8. Conclusion

We have argued that nontrivial words of length L have at most $pe^{O(L)}$ fixed points, while the identity has $p^2 + O(p)$. Thus, for any fixed L or even up to a small multiple of $\log p$, the path-count will approximately match what one would get in the process of computing a Kesten–McKay moment. The error term $O(p)$ cannot be improved because some words, such as the Markoff moves themselves, do have on the order of p fixed points. There is room for improvement in taking longer words, namely allowing L to be a larger multiple of $\log p$. This would lead to a more refined scale at which the Kesten–McKay holds. Beyond the scale $\log p$, the Markoff graph no longer resembles a tree in the same statistical sense that we have proved for smaller L. To see this, start from the 3-regular tree of integer solutions and reduce mod p. There are only $p^2 \pm 3p$ nonzero solutions mod p (and it is not even known whether all of them appear from integer solutions reduced mod p). On the other hand, the first n layers in a 3-regular tree comprise $3 \times 2^n - 2$ nodes. Once $3 \times 2^n - 2 > p^2 + 3p$,

there must be distinct Markoff triples over $\mathbb{Z}$ that coincide mod p. This gives a cycle in $M(\mathbb{F}_p)$ of length at most $2n$ (to the root and back). The same argument produces a closed path starting from any solution mod p that lifts to $\mathbb{Z}$, which Bourgain–Gamburd–Sarnak [BGS16] prove is the vast majority of them. Thus many cycles of length $4\log_2(p)$ or shorter form as the tree collapses on itself mod p. We would not expect it to be possible to take $L > \frac{4}{\log 2}\log p = (5.77078\ldots)\log p$ and still have agreement with the Kesten–McKay moments. At that scale, if not sooner, cycles appear at a positive proportion of the vertices.

The Kesten–McKay law leaves open the question of whether the Markoff graphs are connected for each prime p, and the even harder question of whether they form an expander family. The number of connected components of a 3-regular graph is the multiplicity of $\lambda = 3$ as an eigenvalue. Corollary 1.4 implies that the number of eigenvalues in an interval $[3-\varepsilon, 3]$ is $O(p^2/\log p)$, which is well short of proving even that the number of components is exactly 1 or even $O(1)$ independent of p. To prove a spectral gap, even if the interval contained a bounded number of eigenvalues, one would need a further argument to rule out some eigenvalues being $3+o(1)$ as $p \to \infty$. The bulk distribution of eigenvalues we have proved here is a coarser property.

Acknowledgments

We thank Seungjae Lee for Figure 1.1, the first numerical evidence in favour of the Kesten–McKay law for graphs constructed from the Markoff equation, and many helpful conversations. We thank Peter Sarnak for his encouragement in this project and Nick Katz for helpful discussions. We thank the anonymous referee for a careful reading and many helpful suggestions.

BIBLIOGRAPHY

[Aig13] Martin Aigner, *Markov's Theorem and 100 Years of the Uniqueness Conjecture: A Mathematical Journey from Irrational Numbers to Perfect Matchings*, Springer, 2013. ↑236, 237

[Bar91] Arthur Baragar, *The Markoff equation and equations of Hurwitz*, Ph.D. thesis, Brown University, USA, 1991. ↑230, 231

[BGS16] Jean Bourgain, Alexander Gamburd, and Peter Sarnak, *Markoff Surfaces and Strong Approximation: 1*, `https://arxiv.org/abs/1607.01530`, 2016. ↑230, 231, 240, 241, 242, 243, 248

[Car57] Leonard Carlitz, *The number of points on certain cubic surfaces over a finite field*, Boll. Unione Mat. Ital. **12** (1957), 19–21. ↑231

[CGMP20] Alois Cerbu, Elijah Gunther, Michael Magee, and Luke Peilen, *The cycle structure of a Markoff automorphism over finite fields*, J. Number Theory **211** (2020), 1–27. ↑231, 234, 239

[CL09] Serge Cantat and Frank Loray, *Dynamics on character varieties and Malgrange irreducibility of Painlevé VI equation.*, Ann. Inst. Fourier **59** (2009), no. 7, 2927–2978. ↑238

[ÈH74] M. H. Èl'-Huti, *Cubic surfaces of Markov type*, Math. USSR, Sb. **22** (1974), no. 3, 333–348, translated by R. Lenet. ↑238

[ET48a] Pál Erdős and Pál Turán, *On a problem in the theory of uniform distribution. I*, Proc. Akad. Wet. Amsterdam **51** (1948), 1146–1154. ↑246

[ET48b] _____, *On a problem in the theory of uniform distribution. II*, Proc. Akad. Wet. Amsterdam **51** (1948), 1262–1269. ↑246

[FK65] Robert Fricke and Felix Klein, *Vorlesungen über die Theorie der automorphen Funktionen. Band 1: Die gruppentheoretischen Grundlagen. Band II: Die funktionentheoretischen Ausführungen und die Andwendungen*, Bibliotheca Mathematica Teubneriana, Bände 3, vol. 4, Johnson Reprint Corp.; Teubner, 1965. ↑236

[Fri96] Robert Fricke, *Über die Theorie der automorphen Modulgrupper*, Nachr. Ges. Wiss. Göttingen, Math.-Phys. Kl. **1896** (1896), 91–101. ↑236

[GJS99] Alexander Gamburd, Dmitry Jakobson, and Peter Sarnak, *Spectra of elements in the group ring of SU(2)*, J. Eur. Math. Soc. **1** (1999), no. 1, 51–85. ↑246

[Kes59] Harry Kesten, *Symmetric random walks on groups*, Trans. Am. Math. Soc. **92** (1959), 336–354. ↑227, 232

[KMSV20] Sergeĭ V. Konyagin, Sergey V. Makarychev, Igor E. Shparlinski, and Ilya V. Vyugin, *On the Structure of Graphs of Markoff Triples*, Q. J. Math. **71** (2020), no. 2, 637–648. ↑230

[Mar80] Andreĭ Markoff, *Sur les formes quadratiques binaires indéfinies*, Math. Ann. **17** (1880), no. 3, 379–399. ↑230

[McK81] Bredan D. McKay, *The expected eigenvalue distribution of a large regular graph*, Linear Algebra Appl. **40** (1981), 203–216. ↑227, 232

[MKS04] Wilhelm Magnus, Abraham Karrass, and Donald Solitar, *Combinatorial Group Theory: Presentations of Groups in Terms of Generators and Relations*, reprint of the 1976 ed., Dover Publications, 2004. ↑238

[Mon94] Hugh L. Montgomery, *Ten lectures on the interface between analytic number theory and harmonic analysis*, Regional Conference Series in Mathematics, vol. 84, American Mathematical Society, 1994. ↑246

[MP18] Chen Meiri and Doron and Puder, *The Markoff Group of Transformations in Prime and Composite Moduli*, Duke Math. J. **167** (2018), no. 14, 2679–2720. ↑230

[Nie17] Jakob Nielsen, *Die Isomorphismen der allgemeinen, unendlichen Gruppe mit zwei Erzeugenden*, Math. Ann. **78** (1917), 385–397. ↑236, 237

[Sel91] Atle Selberg, *Collected Papers. Vol. II*, Springer, 1991, Lectures on sieves, p. 65–247. ↑246

[Vaa85] Jeffrey D. Vaaler, *Some extremal functions in Fourier analysis*, Bull. Am. Math. Soc. **12** (1985), no. 2, 183–216. ↑246

Manuscript received on 4th December 2018,
revised on 4th November 2019,
accepted on 1st May 2020.

Recommended by Editor S. Cantat.

Matthew DE COURCY-IRELAND
Department of Mathematics,
Princeton University,
Princeton, NJ, 08544, (USA)
mdc4@math.princeton.edu

Michael MAGEE
Department of Mathematical Sciences,
Durham University,
Durham, DH1 3LE, (UK)
michael.r.magee@durham.ac.uk