Annals of Mathematics, 135 (1992), 1–51

Train tracks and automorphisms of free groups

By Mladen Bestvina and Michael Handel

Introduction

The mapping class group MCG(S) of a compact surface S is the group of isotopy classes of homeomorphisms h: $S \to S$. Thurston introduced the class of *pseudo-Anosov* homeomorphisms in [T2] and used these to choose, for each element $\mathscr{M} \in$ MCG(S), an efficient homeomorphism representing $\mathscr{M}$. He showed that an arbitrary element of MCG(S) can be represented by a homeomorphism h: $S \to S$ that is either of finite order, pseudo-Anosov or built up, in a very simple way, from homeomorphisms of simpler surfaces that are either of finite order or pseudo-Anosov. Thurston's article [T2] contains statements of the main results and a bibliography of the extensive research that this theorem has engendered. For more detailed treatments see [FLP], [BC], [HT] or [BH1].

The outer automorphism group Out(F_n) of the free group F_n is in many ways analogous to MCG(S). The elements of Out(F_n) are equivalence classes of automorphisms Φ: $F_n \to F_n$, where two automorphisms are equivalent if they differ by an inner automorphism. A marked graph G is a graph whose fundamental group has been identified with F_n. Each element $\mathscr{O}$ of Out(F_n) can be represented by a (nonunique) homotopy equivalence f: $G \to G$ of a (nonunique) marked graph. The goal of this paper is to generalize Thurston's theorem by giving an algorithm that produces, for each outer automorphism $\mathscr{O}$, an efficient homotopy equivalence f: $G \to G$ that represents $\mathscr{O}$.

There are several ways to prove Thurston's theorem, each having its own point of emphasis. There is a natural action of MCG(S) on the Teichmüller space of S. Thurston's original proof is based on a compactification of the Teichmüller space to a euclidean ball B in such a way that the natural action of MCG(S) on the Teichmüller space extends over B. Culler and Vogtmann [CV] constructed a free group analogue of the Teichmüller space, called outer space, and a compactification of outer space that is analogous to Thurston's compactification of the Teichmüller space. There is a natural action of Out(F_n) on outer space that extends over this compactification. This compactification is not yet completely understood and is the subject of much current research.

Another approach ([Mi], [HT]) to Thurston's theorem is through the Nielsen setting. The universal cover $\tilde{S}$ of a surface S of negative Euler characteristic can be identified with a convex subset of the hyperbolic plane H. The compactification of H by the "circle at infinity" S_∞ induces a compactification of $\tilde{S}$ by a subset C of S_∞. Each lift $\tilde{h}$: $\tilde{S} \to \tilde{S}$ of a homeomorphism h: $S \to S$ extends over C, and the collection $\{\tilde{h}|C$: $\tilde{h}$ is a lift of $h\}$ depends only on the isotopy class of h. The Nielsen-style approach is to construct a good representative of each mapping class element by considering only this collection of homeomorphisms of C. In a similar way there is a Cantor set K and a compactification $\Gamma \cup K$ of the universal cover Γ of any marked graph G. Each lift $\tilde{f}$: $\Gamma \to \Gamma$ of a homotopy equivalence f: $G \to G$ extends to a homeomorphism $\tilde{f}$: $K \to K$. The collection $\{\tilde{f}$: $K \to K|\tilde{f}$ is a lift of $f\}$ depends only on the outer automorphism determined by f: $G \to G$. This point of view is used in [Co] and [BH2] and influences the work of [CL].

In this paper we adopt the train tracks point of view. Suppose that f: $G \to G$ is a homotopy equivalence of a marked graph. There is no loss in assuming that f maps vertices to vertices and that the restriction of f to the interior of every edge of G is locally injective. The homotopy equivalence f: $G \to G$ is a *train track map* if the restriction of f^k to the interior of every edge of G is locally injective for all $k > 0$. This definition is due to Thurston and is the analog for free group automorphisms of the invariant train tracks that exist for (and completely determine) pseudo-Anosov homeomorphisms of compact surfaces.

If Φ: $F_n \to F_n$ is any positive automorphism, then its usual topological representative on the rose is a train track map. Positive automorphisms have been studied by Gersten [Ge2] and by Cohen and Lustig [CL].

We say that an outer automorphism $\mathcal{O}$ is *reducible* if there are proper free factors $F_1, \ldots, F_k$ of F_n such that $\mathcal{O}$ transitively permutes the conjugacy classes of the F_i's and such that $F_1 * \cdots * F_k$ is a free factor of F_n. For example, if there is a proper free factor whose conjugacy class is preserved by $\mathcal{O}$, then $\mathcal{O}$ is reducible. If $\mathcal{O}$ is not reducible, then we say that it is *irreducible*.

Our first theorem verifies a conjecture of Thurston [T1]. The number λ is the Perron-Frobenius eigenvalue associated to the transition matrix for f: $G \to G$.

THEOREM 1.7. *Every irreducible outer automorphism $\mathcal{O}$ of F_n is topologically represented by a train track map. In fact any irreducible topological representation f: $G \to G$ whose associated eigenvalue λ is minimal (i.e., less than or equal to the eigenvalue associated to any other irreducible topological representation of $\mathcal{O}$) is a train track map. If $\lambda = 1$, then f: $G \to G$ is a finite-order homeomorphism.*

Our proof of Theorem 1.7 is constructive and provides an effective algorithm for finding train track representatives. In [BH1] we modify this algorithm to give an algorithm that produces Thurston homeomorphisms representing elements of MCG(S).

For any Φ: $F_n \to F_n$ let Fix(Φ) be the fixed subgroup Fix(Φ) $= \{x \in F_n : \Phi(x) = x\}$. Dyer and Scott [DS] proved that if Φ is of finite order, then Fix(Φ) is a free factor of F_n; in particular Rank(Fix(Φ)) $\leq n$. Scott later (1978) conjectured that Rank(Fix(Φ)) $\leq n$ for all Φ: $F_n \to F_n$. Gersten [Ge1] proved that Fix(Φ) is always finitely generated (see also [Co], [GT1], [GT2], [St3]). In Section 3 we define *stable train track maps*, prove the existence of such representatives for irreducible outer automorphisms and then use this to prove a strong form of Scott's conjecture for irreducible automorphisms.

THEOREM 3.1. *If* Φ: $F_n \to F_n$ *is an automorphism in an irreducible outer automorphism class, then* Rank(Fix(Φ)) ≤ 1.

In Section 4 we present another application of the existence of stable train track maps for irreducible outer automorphisms. The fundamental group of a compact surface M with nonempty boundary is free and finitely generated. A homeomorphism h: $S \to S$ induces an outer automorphism of $\pi_1(S)$, and hence an outer automorphism of a finitely generated free group. We say that an outer automorphism $\mathcal{O}$ of F_n is *geometric* if it is conjugate to one induced by a surface homeomorphism in this way: we say that h: $S \to S$ *geometrically represents* $\mathcal{O}$. Note that if ∂S is connected and if h: $S \to S$ geometrically represents $\mathcal{O}$, then ∂S determines a conjugacy class s of words in F_n that satisfies $\mathcal{O}(s) = s$ or $\mathcal{O}(s) = \bar{s}$.

THEOREM 4.1. *Suppose that* $\mathcal{O}^l$ *is irreducible for all* $l > 0$ *and that there is a cyclic word* $s \in F_n$ *such that* $\mathcal{O}(s) = s$ *or* $\mathcal{O}(s) = \bar{s}$. *Then* $\mathcal{O}$ *is geometrically realized by a pseudo-Anosov homeomorphism* h: $M \to M$ *of a compact surface with one boundary component.*

Unlike the mapping class group, where one really only needs to understand pseudo-Anosov homeomorphisms, the irreducible elements of Out(F_n) are not the whole story. The algorithm that we use to prove Theorem 1.7 begins with any irreducible homotopy equivalence that is not a train track map and finds either a more efficient irreducible homotopy equivalence or a reduction of the outer automorphism $\mathcal{O}$ represented by the homotopy equivalence. In Section 1 we assume that $\mathcal{O}$ is irreducible so that the algorithm always terminates at a train track map. Section 5 introduces the idea of a *stable relative train track map* and extends the algorithm of Section 1 so that it continues to improve the efficiency of a homotopy equivalence even after reductions are found. The following theorem is the main result of this paper.

THEOREM 5.12. *For every outer automorphism* $\mathcal{O}$ *of* F_n *there exists a stable relative train track map* $f\colon G \to G$ *representing* $\mathcal{O}$.

As an application of Theorem 5.12 we prove the general Scott conjecture.

THEOREM 6.1. *For any automorphism* $\Phi\colon F_n \to F_n$, $\mathrm{Rank}(\mathrm{Fix}(\Phi)) \leq n$.

In [BH2] a slight strengthening of Theorem 5.12 is used to study the dynamical properties of the natural action of an element of $\mathrm{Out}(F_n)$ on the conjugacy classes of words in F_n.

Acknowledgement. In addition to the obvious influence of Thurston on our work we have also been aided by Cohen and Lustig. We proved Lemma 2.1 and Corollary 2.2 after reading their paper [CL]. Thanks also go to the referee, who read our original manuscript with great care and made many helpful suggestions that improved our exposition.

We are grateful to the Institute for Advanced Study for its support and hospitality. The second author also wishes to thank the Sloan Foundation for its support. Both authors were partially supported by grants from the National Science Foundation.

1. Train tracks

The rose with n petals R_n is the graph with one vertex $*$ and n edges. We assume that the free group on n letters F_n is identified with $\pi_1(R_n, *)$. Thus every automorphism $\Phi\colon F_n \to F_n$ can be identified with the automorphism $f_\#\colon \pi_1(R_n, *) \to \pi_1(R_n, *)$ induced by a homotopy equivalence $f\colon (R_n, *) \to (R_n, *)$, and conversely the same holds true.

Example 1.1. Denote the generators of F_4 by $\{a, b, c, d\}$ and define $\Phi\colon F_4 \to F_4$ by $a \mapsto b$, $b \mapsto c$, $c \mapsto d$ and $d \mapsto \bar{a}\bar{d}\bar{c}\bar{b}$, where $\bar{a} = a^{-1}$, $\bar{b} = b^{-1}$, etc. Denote the edges of R_4 by $\{A, B, C, D\}$ and suppose that each of the edges has unit length. Let $f\colon R_4 \to R_4$ be the map that sends A, B and C isometrically onto B, C and D, respectively, and that uniformly expands D over the path $\bar{A}\bar{D}\bar{C}\bar{B}$. □

More generally a marked graph is a graph G along with a homotopy equivalence $\tau\colon R_n \to G$; we denote the vertex set of G by $\mathcal{V}$. A homotopy equivalence $f\colon G \to G$ determines an outer automorphism of $\pi_1(G, \tau(*))$, and hence an outer automorphism $\mathcal{O}$ of F_n. We assume that $f(\mathcal{V}) \subset \mathcal{V}$; if in addition $f|(G \setminus \mathcal{V})$ is locally injective, then we say that $f\colon G \to G$ is a *topological representative* of $\mathcal{O}$. Note that there is no restriction on the valence of the vertices of G.

The *transition matrix* M associated to $f\colon G \to G$ has entries a_{ij} defined as the number of times that the f-image of the j^{th} edge crosses the i^{th} edge in either direction. Thus the transition matrix for Example 1.1 is

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

Recall that a *nonnegative integral matrix M is irreducible* [Se] if for all $1 \le i, j \le \dim(M)$, there exists $N(i,j) > 0$ so that the ij^{th} entry of $M^{N(i,j)}$ is positive. There is a useful equivalent definition of irreducibility in terms of an oriented graph Γ associated to the matrix as follows: Γ has one vertex for each row of M and has a_{ij} oriented edges from the j^{th} vertex to the i^{th} vertex. Note that for each $k > 0$ the ij^{th} entry of M^k is the number of paths in Γ that originate at the j^{th} vertex, terminate at the i^{th} vertex and cross exactly k edges. Thus M is irreducible if and only if there is an oriented path from v_1 to v_2 for every pair of vertices $v_1, v_2 \in \Gamma$. This graph definition will be used throughout the paper to verify that certain matrices are in fact irreducible. We note also that a power of an irreducible matrix need not be irreducible.

A proper subgraph is *nontrivial* if at least one of its components is not a vertex. A subgraph G_0 is *invariant* with respect to a topological representative $f\colon G \to G$ if $f(G_0) \subset G_0$. A topological representative $f\colon G \to G$ is *irreducible* if G does not contain any f-invariant nontrivial subgraphs. Equivalently $f\colon G \to G$ is irreducible if and only if its transition matrix is irreducible.

A subgraph is a *forest* if all of its components are contractible and it is nontrivial. (Our assumption that a forest have at least one component that is not a vertex is not standard.) We say that an outer automorphism $\mathcal{O}$ is *irreducible* if every topological representative $f\colon G \to G$, for which G has no valence-one vertices and no invariant forests, is irreducible. If $\mathcal{O}$ is not irreducible, then we say that it is *reducible*. Thus $\mathcal{O}$ is reducible if there is a graph G with no valence-one vertices and a topological representative $f\colon G \to G$ of $\mathcal{O}$ such that G contains a nontrivial invariant subgraph, but does not contain a nontrivial invariant forest; we call $f\colon G \to G$ a *reduction* for $\mathcal{O}$. We will show that if $\mathcal{O}$ is irreducible, then every topological representative of $\mathcal{O}$ determines, in a natural way, an irreducible topological representative of $\mathcal{O}$.

The following lemma contains a simple criterion for detecting reducibility.

LEMMA 1.2. *If there is a proper free factor F_k of F_n that is invariant, up to conjugacy, under the action of $\mathcal{O}$, then $\mathcal{O}$ is reducible.*

Proof. Choose an automorphism $\Phi\colon F_n \to F_n$ that represents $\mathcal{O}$ such that $\Phi(F_k) = F_k$. Choose a free factor F_{n-k} such that $F_n \cong F_k * F_{n-k}$. Identify $\pi_1(R_n, *)$ with F_n so that the first k edges of R_n correspond to F_k and the remaining $(n-k)$ edges correspond to F_{n-k}. Then $\Phi\colon F_n \to F_n$ is represented by a homotopy equivalence $f\colon R_n \to R_n$ that has R_k as a nontrivial invariant subgraph. Since R_n has no valence-one vertices and no forests, $\mathcal{O}$ is reducible. □

Remark 1.3. Suppose that $f\colon G \to G$ is a reduction for $\mathcal{O}$ and that $G_i = f^i(G_0)$, $0 \le i \le k-1$, are distinct noncontractible components of an f-invariant subgraph. Then each G_i determines a nontrivial free factor F_{n_i} such that $F_{n_1} * \cdots * F_{n_k}$ is a free factor of F_n and such that $\mathcal{O}$ cyclically permutes the conjugacy classes of the F_{n_i}'s. □

The converse to Remark 1.3 is also true. We defer the statement and proof of this to the end of this section, because the proof relies on techniques that we have yet to develop and because the converse plays no direct role in our arguments.

Example 1.4. If $h\colon M \to M$ is a pseudo-Anosov homeomorphism of a compact surface with nonempty boundary, and if h transitively permutes the components of ∂M, then the outer automorphism $\mathcal{O}$ of $F_n \cong \pi_1(M)$ determined by $h\colon M \to M$ is irreducible. This follows from Remark 1.3 and some well-known facts about pseudo-Anosov homeomorphisms (see [BH2]). On the other hand it is easy to modify the proof of Lemma 1.2 (as we do in Lemma 1.16) and to show that if h does not act transitively on the components of ∂M, then the outer automorphism induced by h is reducible (even though it is irreducible, in the sense of Thurston, as a surface homeomorphism). Note that if ∂M is connected, then $\mathcal{O}^l$ is irreducible for all $l > 0$, and if ∂M is not connected, then $\mathcal{O}$ can be irreducible even though some $\mathcal{O}^l$ are reducible. □

Remark. Other classes of irreducible outer automorphisms have been produced by Gersten and Stallings [GS]. An outer automorphism $\mathcal{O}$ induces an automorphism $\mathcal{O}_{ab}$ of the abelianization $\mathbb{Z}^n$ of F_n. The automorphism $\mathcal{O}_{ab}$ can be realized by a matrix; denote its characteristic polynomial by char $\mathcal{O}$. Gersten and Stallings show that if either (i) char $\mathcal{O}$ is a PV polynomial (i.e., is monic and has precisely one root λ, counted with multiplicity, such that $|\lambda| > 1$, whereas for all other roots μ, $|\mu| < 1$); or (ii) char $\mathcal{O}$ is irreducible and some matrix representative M of $\mathcal{O}_{ab}$ is primitive (i.e., M has nonnegative entries and some M^k have strictly positive entries), then each $\mathcal{O}^l$, $l \ge 1$ is irreducible. □

We say that a homotopy equivalence $f\colon G \to G$ is *tight* if for each edge $e \subset G$ either $f|\mathrm{int}(e)$ is locally injective or $f(e)$ is a vertex. A homotopy equivalence $f\colon G \to G$ can be *tightened* to a tight homotopy equivalence by a homotopy rel $\mathscr{V}$.

Suppose that $f\colon G \to G$ is a tight homotopy equivalence and that $G_0 \subset G$ is an invariant forest. Define $G_1 = G/G_0$ to be the quotient graph obtained by collapsing each component of G_0 to a point and let $\pi\colon G \to G_1$ be the quotient map. Define $f_1\colon G_1 \to G_1$ as the homotopy equivalence $\pi f \pi^{-1}\colon G_1 \to G_1$ (which is well defined because $f(G_0) \subset G_0$). If $e \subset G$ is an edge that is not in G_0, then $f_1(e)$ (thought of as a concatenation of edges in G_1) is obtained from $f(e)$ (thought of as a concatenation of edges in G) by removing all occurrences of the edges in G_0. The endpoints of any immersed path $\sigma \subset G_0$ are distinct. Thus if $e_1\sigma e_2$ is a subpath of $f(e)$, where e_1 and e_2 are edges in $G \setminus G_0$ and σ is a nontrivial path in G_0, then $\bar{e}_1 \neq e_2$. This implies that $f_1\colon G_1 \to G_1$ is tight. The transition matrix for $f_1\colon G_1 \to G_1$ is obtained from the transition matrix for $f\colon G \to G$ by deleting the rows and columns associated to the edges of G_0.

There are two types of forest collapsing that will be of particular interest. A forest $G_0 \subset G$ is pretrivial, with respect to a homotopy equivalence $f\colon G \to G$, if $f^l(G_0)$ is a union of vertices for some $l \geq 0$. Let $G_0 \subset G$ be a maximal (with respect to inclusion) pretrivial forest for $f\colon G \to G$. Note that $f(G_0) \subset G_0$ and that an edge in G is contained in G_0 if and only if it is eventually mapped to a point. Collapse G_0 to obtain $f_1\colon G_1 \to G_1$. We say that $f_1\colon G_1 \to G_1$ is obtained from $f\colon G \to G$ by *collapsing a maximal pretrivial forest*. Since $f_1\colon G_1 \to G_1$ is tight and f_1 does not map any edges to points, $f_1\colon G_1 \to G_1$ is a topological representative.

For the second type of forest collapsing, suppose that $G_0 \subset G$ is a maximal (with respect to inclusion) invariant forest. Collapse G_0 to obtain $f_1\colon G_1 \to G_1$. We say that $f_1\colon G_1 \to G_1$ is obtained from $f\colon G \to G$ by *collapsing a maximal invariant forest*. Since G_0 was maximal, G_1 contains no f_1-invariant forests. Thus if G contains no valence-one vertices, and if $f\colon G \to G$ represents an irreducible outer automorphism $\mathscr{O}$, then $f_1\colon G_1 \to G_1$ is irreducible.

The following standard theorem (see [Se]) is essential to our analysis.

THEOREM 1.5 (Perron–Frobenius). *Suppose that M is an irreducible, non-negative integral matrix. Then there is a unique positive eigenvector $\vec{w}$ of norm one for M, and its associated eigenvalue satisfies $\lambda \geq 1$. If $\lambda = 1$, then M is a transitive permutation matrix. Moreover if $\vec{v}$ is a positive vector and $\mu > 0$ satisfies $(M\vec{v})_i \leq \mu v_i$ for each i and $(M\vec{v})_j < \mu v_j$ for some j, then $\lambda < \mu$.*

Remark 1.6. Theorem 1.5 implies that if $M = (a_{ij})$ and $M' = (a'_{ij})$ are irreducible, nonnegative integral matrices, and if each $a_{ij} \leq a'_{ij}$ with strict inequality holding for at least one choice of (i, j), then the Perron–Frobenius eigenvalue for M' is strictly smaller than the Perron–Frobenius eigenvalue for M. □

To each irreducible topological representative $f\colon G \to G$ of an outer automorphism $\mathcal{O}$, we may assign the Perron–Frobenius eigenvalue λ for the transition matrix of $f\colon G \to G$.

A *turn* in G is an unordered pair of oriented edges of G originating at a common vertex. A *turn is nondegenerate*, if it is defined by distinct oriented edges, and is *degenerate* otherwise. Thus a vertex of valence k is associated to k degenerate turns and $\binom{k}{2}$ nondegenerate turns. A map $f\colon G \to G$ induces a self-map Df on the set of oriented edges of G by sending an oriented edge to the first oriented edge in its f-image; this induces a map Tf on the set of turns in G. In Example 1.1, Df is defined by $A \mapsto B \mapsto C \mapsto D \mapsto \overline{A} \mapsto \overline{B} \mapsto \overline{C} \mapsto \overline{D} \mapsto B$.

A turn is *illegal* with respect to $f\colon G \to G$ if its image under some iterate of Tf is degenerate; a turn is *legal* if it is not illegal. In Example 1.1, $\{A, \overline{D}\}$ is the only nondegenerate turn whose Tf-image is degenerate. Since A is not in the image of Df, $\{A, \overline{D}\}$ is the only nondegenerate illegal turn.

We will reserve the word *path* for a map $\alpha\colon [0, 1] \to G$ that is either locally injective or equal to a constant map; in the latter case we say that α is a trivial path. Every map $\sigma\colon [0, 1] \to G$ is homotopic rel endpoints to a (possibly trivial) path $[\sigma]$. If α is nontrivial, then the points $\alpha^{-1}(\mathscr{V})$ subdivide α into a concatenation of subpaths $\alpha = \alpha_1 \cdot \alpha_2 \cdots \alpha_k$, where each α_i maps onto a single edge $E_i \subset G$. We will not usually distinguish between α and the concatenation of edges $E_1 \cdots E_k$. We therefore speak of α crossing or containing the turns $\{\overline{E}_i, E_{i+1}\}$. A path $\rho \in G$ is *legal* if it does not contain any illegal turns. If the f-image of each edge in G is a legal path, then we say that $f\colon G \to G$ is a *train track map*. Equivalently $f\colon G \to G$ is a train track map if $f^k|(G \setminus \mathscr{V})$ is locally injective for all $k > 0$. In Example 1.1, $f\colon G \to G$ is not a train track map, because $f(D)$ contains the turn $\{A, \overline{D}\}$.

We will reserve the word *loop* for a map $\alpha\colon S^1 \to G$ that is locally injective. A loop is *legal* if it does not contain any illegal turns. Every map $\alpha\colon S^1 \to G$ is homotopic to a (possibly trivial) loop $[\sigma]$. Note that a legal path that begins and ends at the same point need not be legal when thought of as a loop.

We may now state the main theorem of this section. It answers a conjecture of Thurston [T1].

THEOREM 1.7. *Every irreducible outer automorphism $\mathcal{O}$ of F_n is topologically represented by an irreducible train track map. In fact any irreducible topological representative $f\colon G \to G$ whose associated Perron–Frobenius eigenvalue λ is minimal (i.e., less than or equal to the Perron–Frobenius eigenvalue associated to any other irreducible topological representative of $\mathcal{O}$) is a train track map. If $\lambda = 1$, then $f\colon G \to G$ is a finite-order homeomorphism.*

Remark 1.8. For any cyclic word C in F_n and any irreducible outer automorphism $\mathcal{O}$ the exponential growth rate of C with respect to $\mathcal{O}$ is defined as $\mathrm{EGR}(\mathcal{O}, C) = \limsup_{n\to\infty}(\log|\mathcal{O}^n(C)|)/n$, where $|\mathcal{O}^n(C)|$ is the cyclic word length of $\mathcal{O}^n(C)$ with respect to some fixed set of generators of F_n. We can compute $\mathrm{EGR}(\mathcal{O}, C)$ from any topological representative $f\colon G \to G$ of $\mathcal{O}$ as $\limsup_{n\to\infty}(\log L([f^n(\sigma)]))/n$, where $\sigma \subset G$ is the loop determined by C and L is the length function on loops determined by a metric on G.

We will show (see Section 3) that if $f\colon G \to G$ is an irreducible train track map with the Perron–Frobenius eigenvalue λ, then there is a metric on G, with respect to which f expands the length of each edge, and hence each legal path or loop, by the factor λ. It follows immediately that $\mathrm{EGR}(\mathcal{O}, C) \le \lambda$ for all C. On the other hand, since $[f^n(\sigma)]$ has a uniformly bounded number of illegal turns, the "bounded cancellation lemma" ([Co] or [T1]) implies that $L([f^n(\sigma)]) > \lambda L(f^{n-1}(\sigma)) - K$, for all $n > 0$ and some constant K. Thus either $L([f^n(\sigma)])$ is uniformly bounded (in which case $\mathcal{O}$ acts periodically on C) or $\mathrm{EGR}(\mathcal{O}, C) = \log(\lambda)$. This argument is due to Thurston ([T1]) and shows that all irreducible train track representations of $\mathcal{O}$ have the same Perron–Frobenius eigenvalue. □

Remark. By Theorem 1.7 and the Remark 1.8 each irreducible outer automorphism $\mathcal{O}$ has an associated eigenvalue $\lambda(\mathcal{O})$. Lemma 1.16 implies that $\mathcal{O}$ is irreducible if and only if $\mathcal{O}^{-1}$ is irreducible. In general $\lambda(\mathcal{O}) \neq \lambda(\mathcal{O}^{-1})$ (although equality does hold for the cases of Example 1.4). For example, let $\{A, B, C\}$ be the edges of R_3, let $f\colon R_3 \to R_3$ be defined by $B \mapsto A$, $C \mapsto B$ and $A \mapsto C\bar{A}$ and let $g\colon R_3 \to R_3$ be defined by $A \mapsto B$, $B \mapsto C$ and $C \mapsto AB$. Then f and g are irreducible train track maps that are homotopy inverses of each other, but the transition matrices for f and g have different Perron–Frobenius eigenvalues. The elements of $\mathrm{Out}(F_n)$ determined by f and g are irreducible (see Example 3.3 of [GS]). □

Example 1.9. Not every outer automorphism can be represented by a train track map. For example, let $\mathcal{O}$ be the outer automorphism determined by the automorphism $\Phi\colon F_3 \to F_3$ defined by $a \mapsto ba$, $b \mapsto bba$ and $c \mapsto c\bar{a}\bar{b}a\bar{b}$.

Suppose that $f: G \to G$ is a train track representative of $\mathcal{O}$. Let $C \subset G$ and $W \subset G$ be the immersed loops in the free homotopy classes determined by $c \in F_n$ and $w = \bar{a}ba\bar{b} \in F_n$, respectively. Since CW^m tightens to a loop $[f^m(C)]$ that has a uniformly bounded number of illegal turns, W must be entirely legal. Thus $f(W)$ is a loop, and since $\Phi(w) = w$, $f(W)$ must equal W (as loops). This implies that f permutes the edges crossed by W. Choose $l > 0$ so that f^l fixes each edge crossed by W. Since w abelianizes to zero, W is contained in a subgraph $G_1 \subset \mathrm{Fix}(f^l)$ of rank at least two. This implies that all of the eigenvalues of the invertible map $f_*: H_1(G; \mathbb{Z}) \to H_1(G; \mathbb{Z})$ have modulus one, in contradiction to the fact that the element $\hat{\Phi}$ of $\mathrm{GL}(3, \mathbb{Z})$, defined by abelianizing Φ, has eigenvalues with modulus greater than one. □

As a warmup for the proof of Theorem 1.7 we consider Example 1.1. As noted above $f: G \to G$ fails to be a train track map because D is mapped across the illegal turn $\{A, \bar{D}\}$. To improve the situation we perform Stallings' [St1] folding operation, namely identifying all of A with the first quarter of $\bar{D}$. The resulting graph G_1 in Figure 1 is combinatorially the same as G, and we will denote the images in G_1 of A, B, C and the first three quarters of D by A, B, C and D_1. Since A and the first quarter of $\bar{D}$ are both mapped homeomorphically by f to B, there is a quotient homotopy equivalence $f_1: G_1 \to G_1$. Note that $f_1(A) = B$, $f_1(B) = C$, $f_1(C) = D_1\bar{A}$ and $f_1(D_1) = \bar{A}A\bar{D}_1\bar{C}$, because the correspondence between G and G_1 determined by the quotient map is given by $D \leftrightarrow D_1\bar{A}$.

Tighten f_1; $G_1 \to G_1$ to obtain a new map $f_2: G_1 \to G_1$ so that $f_2(D_1) = \bar{D}_1\bar{C}$ and f_2 agrees with f_1 on the rest of G_1.

The map Df_2 is defined by $\bar{A} \mapsto \bar{B} \mapsto \bar{C} \mapsto A \mapsto B \mapsto C \mapsto D_1 \mapsto \bar{D}_1 \mapsto C$. Under forward iteration Df_2 identifies the oriented edges into three classes $\{\bar{A}, A, D_1\}$, $\{\bar{B}, B, \bar{D}_1\}$ and $\{C, \bar{C}\}$. Thus a nondegenerate turn is illegal if and

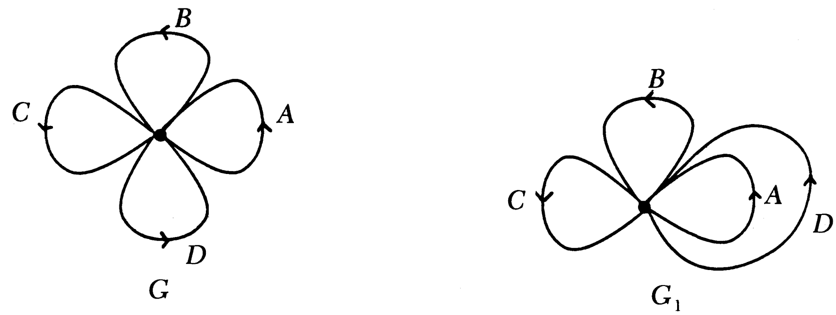

FIGURE 1

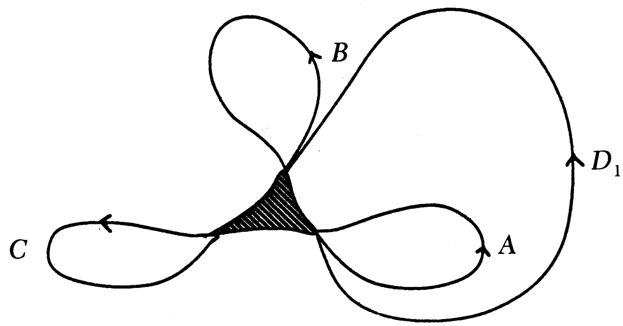

FIGURE 2

only if it is determined by two oriented edges in the same grouping. Figure 2 shows G_1 with the illegal turns depicted as sharp corners and the vertex blown up to a solid triangle. A legal path in G_1 is simply one that stays tangent to the graph and does not make any 180° turns; it is a path that a train could take as it moves along its tracks.

It is easy to see that $f_2\colon G_1 \to G_1$ is a train track map. The only turns that occur in the images of edges are $\{\overline{D}_1, \overline{A}\}$ and $\{D_1, \overline{C}\}$, which are both legal. The lesson here is that $f\colon G \to G$ fails to be a train track map because some f^k (edge) contains a degenerate turn, and that k can sometimes be reduced by folding. When $k = 1$, tightening will remove the degenerate turn. (In the formal proof we will fold one more time to produce a valence-one vertex, which is then removed by a homotopy.)

In general it takes many steps to pass from an arbitrary, irreducible, topological representative of $\mathcal{O}$ to one that is a train track map. Progress is measured by a decrease in the eigenvalue λ associated to the transition matrix.

We now define four operations on topological representatives $f\colon G \to G$ of an outer automorphism $\mathcal{O}$ and note their effect on λ. The marking on G is changed in an obvious way. For example, in the fold done on the previous page, the marking on the new graph is obtained from the marking on the original graph by identifying $D_1\overline{A}$ with D.

Suppose that w is not a vertex of G, but $f(w)$ is. Without changing the underlying space of G, we give it a new graph structure G_1 by adding w to the previous vertices. We say that $f_1 = f\colon G_1 \to G_1$ is obtained from $f\colon G \to G$ by *subdivision*.

LEMMA 1.10. *If $f\colon G \to G$ is an irreducible topological representative of $\mathcal{O}$ and $f_1\colon G_1 \to G_1$ is obtained from $f\colon G \to G$ by subdivision, then $f_1\colon G_1 \to G_1$ is an irreducible topological representative of $\mathcal{O}$, and the associated eigenvalues, $\lambda_1 = \lambda$ are equal.*

Proof. We may assume that the edges of G are $e_1, \dots, e_m$ and that e_m is being subdivided into two edges e'_m and e'_{m+1}. A segment of $f(e_i)$ that crossed e_m now crosses both e'_m and e'_{m+1}; the total number of times that $f(e_m)$ crosses a given edge is equal to the sum of the total number of times $f_1(e'_m)$ crosses that edge plus the total number of times $f_1(e'_{m+1})$ crosses that edge. Thus if $M = (a_{ij})$ and $M_1 = (b_{ij})$ are the transition matrices for $f\colon G \to G$ and $f_1\colon G_1 \to G_1$, respectively, then

$$\begin{aligned} b_{ij} &= a_{ij}, && 1 \le i \le m;\ 1 \le j \le m-1, \\ b_{i,m} + b_{i,m+1} &= a_{i,m}, && 1 \le i \le m, \\ b_{m+1,j} &= b_{m,j}, && 1 \le j \le m+1. \end{aligned}$$

Choose a positive vector $\vec{w}$ such that $M\vec{w} = \lambda\vec{w}$ and define $\vec{v}$ by $v_i = w_i$ for $1 \le i \le m$ and $v_{m+1} = v_m$. Then $M_1\vec{v} = \lambda\vec{v}$, and by Theorem 1.5 it suffices to show that M_1 is irreducible.

Let Γ be the oriented graph with m vertices and a_{ij} edges from the j^{th} vertex to the i^{th} vertex. Define Γ_1 for M_1 similarly. Then Γ_1 is obtained from Γ by replacing the m^{th} vertex x by a pair of vertices y and z. Every edge into x corresponds to a pair of edges (with the same initial point), one into each of y and z. Every edge out of x corresponds to an edge (with the same terminal point) out of y or out of z, and there is at least one edge out of each. One easily checks that every pair of vertices of Γ_1 can be connected by an oriented path, using the assumption that the corresponding statement is true for Γ. □

Suppose that $f\colon G \to G$ is a topological representative of $\mathcal{O}$ and that v is a valence-one vertex of G with an incident edge e. Let $G(1)$ be the subgraph of G obtained by removing v and the interior of e, let $\pi_1\colon G \to G(1)$ be the deformation retraction that collapses e to a vertex in $G(1)$ and let $f_1\colon G_1 \to G_1$ be the topological representative of $\mathcal{O}$ obtained from $\pi_1 f|G(1)\colon G(1) \to G(1)$ by tightening and collapsing a maximal pretrivial forest. We say that $f_1\colon G_1 \to G_1$ is obtained from $f\colon G \to G$ by a *valence-one homotopy*. Note that G_1 has fewer vertices than G.

LEMMA 1.11. *If $f\colon G \to G$ is an irreducible topological representative and $f_1\colon G_1 \to G_1$ is an irreducible topological representative obtained from $f\colon G \to G$ by a sequence of valence-one homotopies followed by the collapse of a maximal invariant forest, then the associated eigenvalues satisfy $\lambda_1 < \lambda$.*

Proof. The r-dimensional transition matrix M_1 for $f_1\colon G_1 \to G_1$ is obtained from the m-dimensional transition matrix M for $f\colon G \to G$ by removing the rows and columns corresponding to some of the edges of G. After renumbering, we may assume that these edges are labelled $e_{r+1}, \dots, e_m$. Thus M_1 is the square submatrix determined by $e_1, \dots, e_r$. It is easy to see (by the graph definition of irreducibility) that there are $1 \le i_0 \le r$ and $r + 1 \le j_0 \le m$ such that $a_{i_0 j_0} > 0$.

Choose a positive vector $\vec{w}$ so that $M\vec{w} = \lambda\vec{w}$ and let $\vec{v}$ be the vector defined by $v_i = w_i$ for $1 \le i \le r$. Then $(M_1\vec{v})_i \le \lambda v_i$ for all $1 \le i \le r$ and $(M_1\vec{v})_{i_0} < \lambda v_{i_0}$. By Theorem 1.5, $\lambda_1 < \lambda$. □

Remark 1.12. During the proof of Lemma 1.11 we showed that if M is irreducible and M_1 is an irreducible square submatrix of M, then the Perron–Frobenius eigenvalue for M_1 is strictly less than the Perron–Frobenius eigenvalue for M. □

Suppose next that $v \in \mathscr{V}$ has valence two; after renumbering and reorienting the edges $e_1, \dots, e_m$ of G if necessary, we may assume that v is the initial vertex of e_m and the terminal vertex of e_{m-1}. Let g_t: $G \to G$ be a homotopy with support in $e_{m-1} \cup e_m$ from $g_0 =$ identity to a map g_1 that collapses e_m to a point and stretches e_{m-1} across e_m. Define $f(1)$: $G \to G$ by tightening $g_1 f$: $G \to G$ and note that v is not in the $f(1)$-image of $\mathscr{V}$. Define $f(2)$: $G(2) \to G(2)$ from $f(1)$: $G \to G$ by performing the inverse of a subdivision, removing v from the vertices of G and then tightening the image of the new edge (called e'_{m-1}) obtained from $e_{m-1} \cup e_m$ if it is not already tight. If E is an edge of G other than e_{m-1} or e_m, then E is naturally an edge of $G(2)$ and $f(2)(E)$ is obtained from $f(E)$ by removing all occurrences of e_m and $\bar{e}_m$ and by replacing each e_{m-1} and $\bar{e}_{m-1}$ with e'_{m-1} and $\bar{e}'_{m-1}$, respectively. The $f(2)$ image of e'_{m-1} is obtained from the concatenation $f(e_{m-1}) \cdot f(e_m)$ by tightening, removing all occurrences of e_m and $\bar{e}_m$ and replacing each e_{m-1} and $\bar{e}_{m-1}$ with e'_{m-1} and $\bar{e}'_{m-1}$, respectively. Finally define f_1: $G_1 \to G_1$ to be the topological representative that is obtained from $f(2)$: $G(2) \to G(2)$ by collapsing a maximal pretrivial forest. We say that f_1: $G_1 \to G_1$ is obtained from f: $G \to G$ by a *valence-two homotopy of v across e_m*. Note that G_1 has fewer vertices than G_2 and that no new valence-one vertices have been produced.

LEMMA 1.13. *Suppose that f: $G \to G$ is an irreducible topological representative and that f_2: $G_2 \to G_2$ is an irreducible topological representative that is obtained from f: $G \to G$ by a valence-two homotopy of v across e_m followed by the collapse of a maximal invariant forest. Let M be the transition matrix for f: $G \to G$ and let $\vec{w}$ be a positive vector such that $M\vec{w} = \lambda\vec{w}$. If $w_{m-1} \le w_m$, then $\lambda_2 \le \lambda$; if $w_{m-1} < w_m$, then $\lambda_2 < \lambda$.*

Remark 1.14. If $f: G \to G$ is an irreducible topological representative of $\mathcal{O}$ whose associated eigenvalue λ is already minimal, then Lemma 1.13 implies that the eigenvector coefficients for the edges adjacent to a valence-two vertex must be equal. In particular a valence-two homotopy can be performed across either adjacent edge without changing λ. $\square$

Proof of Lemma 1.13. Define an $(m-1)$-dimensional matrix M' by adding the m^{th} column of $M = (a_{ij})$ to the $(m-1)^{\text{st}}$ column of M and then removing the m^{th} row and m^{th} column of M. If f is locally injective at v, then M' equals the transition matrix $M(2)$ for the map called $f(2)$: $G(2) \to G(2)$ in the definition of valence-two homotopy. If f is not locally injective at v, then nontrivial tightening occurs when $f(2)$ is obtained from $f(1)$; so $M(2)$ is obtained from this new matrix by reducing some of the entries in the $(m-1)^{\text{st}}$ column.

Define the $(m-1)$-dimensional vector $\vec{v}$ by $v_i = w_i$ for $1 \le i \le m-1$. Note that $(M_2\vec{v})_i \le \lambda w_i - a_{im}(w_m - w_{m-1}) \le \lambda w_i = \lambda v_i$ for all $1 \le i \le m-1$. Since M is irreducible, there exists $1 \le i_0 \le m-1$ such that $a_{i_0 m} > 0$. If $w_m - w_{m-1} > 0$, then $(M_2\vec{v})_{i_0} < v_{i_0}$. If $M(2)$ is irreducible, then $M_1 = M(2)$ and Theorem 1.5 completes the proof.

If $M(2)$ is not irreducible, then M_2, the transition matrix for f_2: $G_2 \to G_2$, is obtained from $M(2)$ by removing the rows and columns corresponding to edges in a pretrivial and an invariant forest. If the $(m-1)^{\text{st}}$ row and column of $M(2)$ are removed, then M_2 is a square submatrix of M. Remark 1.12 completes the proof in this case. Suppose then that the $(m-1)^{\text{st}}$ row and column of $M(2)$ are not removed. The edges in the pretrivial forest F_1 can be thought of as being contained in G. The f-image of any edge in F_1 is contained in $F_1 e_m$. Similarly the edges in the invariant forest F_2 can be thought of as being in G; the f-image of each edge in F_2 is contained in $F_2 \cup F_1 \cup e_m$. Since M is irreducible, there exists an edge e_{i_0} in G such that $e_{i_0} \not\subset F_2 \cup F_1 \cup e_m$ and such that $a_{i_0 m} > 0$. The proof now concludes as in the case where $M(2)$ is irreducible. $\square$

Finally suppose that some pair of edges originating at a common vertex of G have the same images under f; i.e., their images cross the same edges of G in the same order. Let G_1 be the graph obtained by identifying these two edges in such a way that $f: G \to G$ descends to a well-defined map f_1: $G_1 \to G_1$. We say that the homotopy equivalence f_1: $G_1 \to G_1$ is obtained from $f: G \to G$ by an *elementary fold* (see [St1]). More generally suppose that e_i and e_j are edges of G with a common initial vertex v and that there are nontrivial, maximal, initial segments $e_i' \subset e_i$ and $e_j' \subset e_j$ with endpoints in $f^{-1}(\mathcal{V})$ such that f maps e_i' and e_j' over the same edges in the same order. Subdivide e_i and e_j at the endpoints of e_i' and e_j' if necessary. Then perform an elementary fold of e_i' and e_j'. The resulting homotopy equivalence $f_1: G_1 \to G_1$ is obtained from $f: G \to G$ by *folding* e_i and e_j. It is a *partial fold* if both e_i and e_j are subdivided, and a *full fold* otherwise. A partial fold increases the number of vertices in the graph, but a full fold does not.

LEMMA 1.15. *Suppose that $f: G \to G$ is an irreducible topological representative of $\mathcal{O}$ and that $f_1: G_1 \to G_1$ is obtained from $f: G \to G$ by folding a pair of edges. If $f_1: G_1 \to G_1$ is a topological representative, then it is irreducible, and the associated eigenvalues $\lambda_1 = \lambda$ are equal. If $f_1: G_1 \to G_1$ is not a topological representative, let $f_2: G_2 \to G_2$ be an irreducible topological representative obtained from $f_1: G_1 \to G_1$ by tightening, collapsing a maximal pretrivial forest and collapsing a maximal invariant forest. Then the associated eigenvalues satisfy $\lambda_2 < \lambda$.*

Proof. In light of Lemma 1.10 we may assume that the fold is an elementary one. In particular we may assume that the oriented edges being folded do not determine the same unoriented edge. Label the edges of G by $e_1, \dots, e_m$ so that e_{m-1} and e_m are the edges folded together. Let $M = (a_{ij})$ be the m-dimensional transition matrix for $f: G \to G$. Define the $(m-1)$-dimensional matrix $M_1 = (b_{ij})$ by

$$b_{ij} = a_{ij}, \qquad 1 \le i \le m-2, 1 \le j \le m-1,$$

$$b_{m-1,j} = a_{m-1,j} + a_{m,j}, \qquad 1 \le j \le m-1.$$

The matrix M_1 is irreducible; the proof is similar to the one given for Lemma 1.10 and is left to the reader.

If $f_1: G_1 \to G_1$ is a topological representative, then M_1 is its transition matrix, because any interval in G that is mapped by f over e_{m-1} or e_m corresponds to an interval in G_1 that is mapped by f_1 over the image $e'_{m-1} \subset G_1$ of e_{m-1} and e_m. Choose a positive vector $\vec{w}$ such that $M\vec{w} = \lambda\vec{w}$ and define $\vec{v}$ by $v_i = w_i$ for $1 \le i \le m-2$ and $v_{m-1} = w_{m-1} + w_m$. A calculation shows that $M_1\vec{v} = \lambda\vec{v}$. By Theorem 1.5, $\lambda_1 = \lambda$.

If $f_1: G_1 \to G_1$ is not a topological representative, then the transition matrix $M_2 = (c_{ij})$ for $f_2: G_2 \to G_2$ is obtained from M_1 by decreasing at least one entry and possibly removing the columns and rows corresponding to certain edges of G_1. If no rows and columns are removed, then M_2 is $(m-1)$-dimensional and $c_{ij} \le b_{ij}$ for all $1 \le i, j \le m-1$, with strict inequality holding for at least one pair (i_1, j_1). In this case Remark 1.6 completes the proof.

Suppose then that at least one row and column of M_1 is collapsed. After relabelling the edges of G_1, we may assume that the first r rows and columns of

M_1 are removed for some $1 \le r \le m-2$. The entries of M_2 are less than or equal to the entries of the square submatrix of M_1 given by the last $m-r-1$ rows and columns of M_1. Since M_1 is irreducible, there are $1 \le j_0 \le r$ and $r+1 \le i_0 \le m-1$ such that $b_{i_0 j_0} > 0$. Let $\vec{u}$ be the positive vector defined by $u_i = v_{r+i}$ for $1 \le i \le m-r-1$. Then $(M_2\vec{u})_i \le \lambda_1 u_i$ for all $1 \le i \le m-r-1$ and $(M_2\vec{u})_{i_0-r} < \lambda_1 u_{i_0-r}$. Once again Theorem 1.5 completes the proof. $\square$

We have now finished analyzing the four basic operations and can turn to proving this section's main theorem.

Proof of Theorem 1.7. Suppose that $f\colon G \to G$ is an irreducible topological representative of $\mathcal{O}$ and that G has no valence-one or valence-two vertices. If $f\colon G \to G$ fails to be a train track map, we will find another such irreducible topological representative $f_4\colon G_4 \to G_4$ of $\mathcal{O}$ such that $1 \le \lambda_4 < \lambda$. An easy Euler characteristic argument shows that the number of edges in G is uniformly bounded by $3n-3$. The Perron–Frobenius eigenvalue is bounded below by the minimum value of the sum of the entries in a row of the transition matrix (see [Se]). Since the size of the matrix is bounded, there is a uniform choice of k such that each row sum of M^k is at least as big as the largest entry of M. Thus our matrices take on only finitely many values less than any given constant K, and iteration of our procedure eventually terminates at an irreducible topological representative that is also a train track map.

If $\lambda = 1$, then M is a permutation matrix and $f\colon G \to G$ is a homeomorphism. We may therefore assume that $\lambda > 1$. We may also assume that f maps each component of $G \setminus f^{-1}(\mathcal{V})$ linearly onto an edge (with respect to some metric on G); since M is irreducible, $\bigcup_{l=1}^{\infty} f^{-1}(\mathcal{V})$ is dense in G.

If $f\colon G \to G$ is not a train track map, then there exists $P \in G \setminus \mathcal{V}$ such that $f(P) \in \mathcal{V}$ and f^k is not locally injective at P for some $k > 1$. Let U be a neighborhood of P such that:

(i) $\partial U = \{s, t\} \subset f^{-l}(\mathcal{V})$ for some $l > 0$;
(ii) $f^i|U$ is one to one, $1 \le i \le k-1$;
(iii) $f^k|(U \setminus P)$ is two to one onto a subset of a single edge;
(iv) $P \notin f^i(U)$, $1 \le i \le k$.

In Figure 3 we are suppressing all but two of the oriented edges incident to the vertices $f^i(P)$.

Perform a subdivision at P. Then subdivide at $\{f^j(s), f^j(t)\colon 0 \le j \le l-1\}$ in reverse order, i.e., first on $f^{l-1}(s)$ and $f^{l-1}(t)$ (which may be the same point), then on $f^{l-2}(s)$ and $f^{l-2}(t)$, etc.

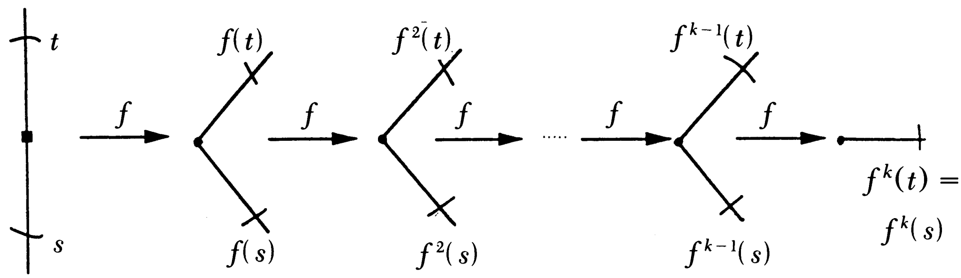

FIGURE 3

Denote the resulting irreducible topological representative of $\mathcal{O}$ by f_1: $G_1 \to G_1$ and note that, by Lemma 1.10, $\lambda_1 = \lambda$. The vertex P has valence two in G_1; denote the edges incident to P by α and β. Fold $Df_1^{k-1}(\alpha)$ and $Df_1^{k-1}(\beta)$ to obtain $f(1)$: $G(1) \to G(1)$. If $f(1)$: $G(1) \to G(1)$ is not a topological representative, define f_2: $G_2 \to G_2$ as the irreducible topological representative obtained from $f(1)$: $G(1) \to G(1)$ by tightening, collapsing a maximal pretrivial forest and collapsing a maximal invariant forest. In this case Lemma 1.15 implies that $\lambda_2 < \lambda$. If $f(1)$: $G(1) \to G(1)$ is a topological representative, then P is a valence-two vertex of $G(1)$ with incident edges α and β; fold $Df(1)^{k-2}(\alpha)$ and $Df(1)^{k-2}(\beta)$ to obtain $f(2)$: $G(2) \to G(2)$. Repeating this dichotomy k times if necessary, we arrive at an irreducible topological representative f_2: $G_2 \to G_2$ such that either $\lambda_2 < \lambda$ or $\lambda_2 = \lambda$ and P is a valence-one vertex of G_2; see Figure 4.

Perform a sequence of valence-one homotopies to remove all valence-one vertices followed by the collapse of a maximal invariant forest; denote the resulting irreducible topological representative by f_3: $G_3 \to G_3$. Lemma 1.11 implies that $\lambda_3 < \lambda$.

If v is a valence-two vertex of G_3 with edges e_i and e_j, perform a valence-two homotopy of v across e_j, where $w_i \leq w_j$ for the positive eigen-

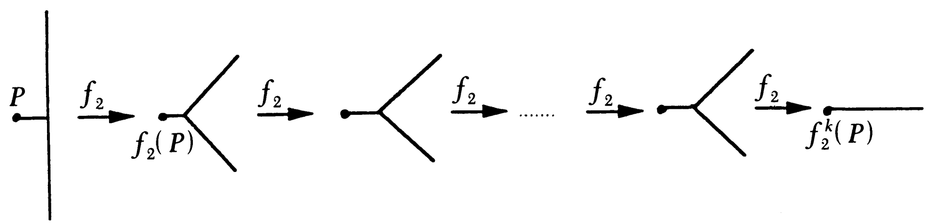

FIGURE 4

vector $\vec{w}$ associated to the transition matrix for f_3: $G_3 \to G_3$. Collapse a maximal invariant forest if necessary to maintain an irreducible transition matrix. Continue this process until there are no valence-two vertices left and denote the resulting irreducible topological representative of $\mathcal{O}$ by f_4: $G_4 \to G_4$. Lemma 1.13 implies that $\lambda_4 < \lambda$. □

To conclude this section we give the converse to Remark 1.3.

LEMMA 1.16. *If there are free factors F_{n_i}, $1 \leq i \leq k$, $n_1 < n$, such that $F_{n_1} * \cdots * F_{n_k}$ is a free factor of F_n and $\mathcal{O}$ cyclically permutes the F_{n_i}'s, then $\mathcal{O}$ is reducible.*

Proof. Label the generators of F_{n_i} by $a_1^i, \ldots, a_{n_i}^i$. Let R_i be the rose with n_i petals $A_1^i, \ldots, A_{n_i}^i$ and vertex v_i. Choose automorphisms Φ_i: $F_n \to F_n$ representing $\mathcal{O}$ such that $\Phi_i(F_{n_i}) = F_{n_{i+1}}$ for $1 \leq i \leq k-1$ and such that $\Phi_k(F_{n_k}) = F_{n_1}$; let f_i: $R_i \to R_{i+1(\text{mod } k)}$ be the corresponding homotopy equivalences. Choose a free factor $F_{n_{k+1}}$ so that $F_n \cong F_{n_1} * \cdots * F_{n_{k+1}}$ and let $R_{n_{k+1}}$ be the rose on n_{k+1} petals with vertex v_{k+1}. Define G as the union of the R_{n_i}'s with edges E_i, $1 \leq i \leq k$, connecting v_{k+1} to v_i. Collapsing the E_i's to points gives a homotopy equivalence from G to the rose on n petals; this identifies $\pi_1(G, v_{k+1})$ with $F_{n_1} * \cdots * F_{n_{k+1}}$ and thereby gives a marking for G. Note that G has no valence-one vertices.

We will use Φ_1 to extend $\bigcup_{i=1}^k f_i$: $R_i \to R_{i+1(\text{mod } k)}$ to a topological representative f: $G \to G$. For $1 \leq i \leq k$ there exists $c_i \in F_n$ such that $\Phi_1(x) = c_i \Phi_i(x) \bar{c}_i$ for all $x \in F_n$. Let $\gamma_i \subset G$ be the path originating and terminating at v_{k+1} that is determined by c_i. Define $f(E_i) = \gamma_i E_{i+1(\text{mod } k)}$. Finally define f on the edges of R_{k+1} according to Φ_1 and the marking on G.

This is a reduction for $\mathcal{O}$ unless $n_{k+1} = 0$ and each c_i is the identity element. In that case we define a new topological representative f_1: $G_1 \to G_1$ as follows: First perform a homotopy with support in E_1 so that the image of E_1 is now $E_2 f(A_1^1) f(\bar{A}_1^1)$; then fold the part of $\bar{E}_1$ that maps to $f(A_1^1)$ with all of A_1^1; call the resulting map f_1: $G_1 \to G_1$. The graphs G and G_1 differ by a change in marking. The maps f and f_1 agree on all edges except E_1 and E_k; $f_1(E_1) = E_2 f(A_1^1)$ and $f_1(E_k) = E_1 \bar{A}_1^1$. The E_i's no longer form an invariant forest so that f_1: $G_1 \to G_1$ is a reduction of $\mathcal{O}$. □

2. A fixed point lemma

In this section we use standard topological techniques to show that $\text{Rank}(\text{Fix}(\Phi)) \leq 1$ for many automorphisms Φ: $F_n \to F_n$.

To begin with some standard facts, let f: $G \to G$ be a topological representative of $\mathcal{O}$. The marking τ: $(R_n, *) \to (G, x)$ identifies F_n with $\pi_1(G, x)$. Any path u from x to $f(x)$ determines an automorphism $(f_u)_\#\colon \pi_1(G, x) \to \pi_1(G, x)$ by $(f_u)_\#(\langle \alpha \rangle) = \langle u \cdot f(\alpha) \cdot u^{-1} \rangle$, where $\alpha\colon [0,1] \to G$ satisfies $\alpha(0) = \alpha(1) = x$ and $\langle \alpha \rangle$ is the element of $\pi_1(G, x)$ determined by α. As u varies among all homotopy classes of paths between x and $f(x)$, $(f_u)_\#$ varies among all automorphisms in the outer automorphism class $\mathscr{O}$.

Let $p\colon \Gamma \to G$ be the projection of the universal cover Γ of G onto G and let T be the group of covering translations of Γ. Choose a lift $\tilde{x} \in \Gamma$ of x. For each $t \in T$ there is a unique path γ_t connecting $\tilde{x}$ to $t\tilde{x}$; the map $t \mapsto \langle p\gamma_t \rangle$ identifies T with $\pi_1(G, x)$. For each u: $[0,1] \to G$, with $u(0) = x$ and $u(1) = f(x)$, there is a unique lift $\tilde{u}$: $[0,1] \to \Gamma$ with $\tilde{u}(0) = \tilde{x}$. Let $\tilde{f}_u\colon \Gamma \to \Gamma$ be the lift of $f\colon G \to G$ satisfying $\tilde{f}_u(\tilde{x}) = \tilde{u}(1)$.

We claim that, under the identification of T with $\pi_1(G, x)$, the automorphism $(f_u)_\#\colon\ \pi_1(G, x) \to \pi_1(G, x)$ is identified with the automorphism $(\tilde{f}_u)_\#\colon T \to T$ defined by $t \mapsto \tilde{f}_u t \tilde{f}_u^{-1}$. To see this, note that $\tilde{f}_u(\gamma_t)$ is the lift of $f(p(\gamma_t))$ that originates at $\tilde{f}_u(\tilde{x})$. Since the terminal endpoint of $\tilde{f}_u(\gamma_t)$ equals $\tilde{f}_u(t\tilde{x}) = \tilde{f}_u t \tilde{f}_u^{-1}(\tilde{f}_u \tilde{x})$, and since $\tilde{u}$ connects $\tilde{x}$ to $\tilde{f}_u(\tilde{x})$, the lift of $u \cdot f(p(\gamma_t)) \cdot u^{-1}$ that originates at $\tilde{x}$ is $\tilde{u} \cdot \tilde{f}_u(\gamma_t) \cdot \tilde{f}_u t \tilde{f}_u^{-1}(\tilde{u}^{-1})$. The endpoint of this lifted path is $\tilde{f}_u t \tilde{f}_u^{-1}(\tilde{x})$, as desired.

We note that $\mathrm{Fix}((\tilde{f}_u)_\#) = \{t \in T\colon t\tilde{f}_u = \tilde{f}_u t\}$.

Each $t \in T$ has a unique axis $A(t) \subset \Gamma$, i.e., a t-invariant subgraph, which is isometric to the real line and on which t acts by translation. We will use two simple standard facts about axes. The first is that if $C \subset \Gamma$ is connected and if $tC \cap C \neq \emptyset$, then $C \cap A(t) \neq \emptyset$. The second is that if s and t do not commute, then $A(s) \cap A(t)$ is compact.

Lemma 2.1. *If $\tilde{f}\colon \Gamma \to \Gamma$ is a lift of a topological representative $f\colon G \to G$, and if* $\mathrm{Rank}(\mathrm{Fix}(\tilde{f}_\#)) \geq 2$, *then* $\mathrm{Fix}(\tilde{f}) \neq \emptyset$.

Proof. Choose $t \in \mathrm{Fix}(\tilde{f}_\#)$ and let A be the axis of t. Since $\tilde{f}(A)$ is connected and t-invariant, $\tilde{f}(A) \cap A \neq \emptyset$. Choose $a \in A$ such that $\tilde{f}(a) \in A$; replacing t by t^{-1} if necessary, we may assume that $\tilde{f}(a)$ lies in the component of $A \setminus \{a\}$ that is mapped into itself by t. Since $\mathrm{Rank}(\mathrm{Fix}(\tilde{f}_\#)) \geq 2$, there exists $s \in \mathrm{Fix}(\tilde{f}_\#)$, whose axis B intersects A in a compact (possibly empty) set. As above, there exists $b \in B$ such that $\tilde{f}(b)$ lies in the component of $B \setminus \{b\}$ that is mapped into itself by s. Replacing a and b by $t^n(a)$ and $s^m(b)$ for large values of m and n if necessary, we may assume that the oriented path connecting $\tilde{f}(a)$ to $\tilde{f}(b)$ contains the oriented path γ connecting a to b as an oriented subpath. (In Figure 5 we are assuming that A and B intersect nontrivially, but the claim is just as obvious in the case where A and B are disjoint.) If $x \in \gamma$ is close to a, then x separates $\tilde{f}(x)$ from b. This fails for $x \in \gamma$ close to b. Thus $\tilde{f}(x) = x$ for some $x \in \gamma$. $\square$

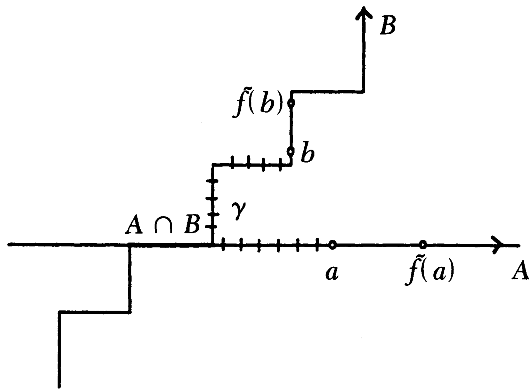

FIGURE 5

COROLLARY 2.2. *Suppose that* $f\colon G \to G$ *is a topological representative of the outer automorphism* $\mathcal{O}$ *determined by the automorphism* $\Phi\colon F_n \to F_n$. *If* $\mathrm{Rank}(\mathrm{Fix}(\Phi)) \geq 2$, *then* Φ *is conjugate to the automorphism* $f_\#\colon \pi_1(G, y) \to \pi_1(G, y)$ *defined by the trivial path at* y *for some* $y \in \mathrm{Fix}(f)$.

Proof. We have established a conjugacy between $\Phi\colon F_n \to F_n$ and $(\tilde{f})_\#\colon T \to T$, where $\tilde{f}\colon \Gamma \to \Gamma$ is some lift of $f\colon G \to G$. The same arguments also establish, for any $\tilde{y} \in \Gamma$, a conjugacy between $(\tilde{f})_\#\colon T \to T$ and the automorphism $(f_v)_\#\colon \pi_1(G, y) \to \pi_1(G, y)$ determined by f and the path $v = p(\gamma)$, where $\gamma \subset \Gamma$ is the path connecting $\tilde{y}$ to $\tilde{f}(\tilde{y})$. Lemma 2.1 implies that there exists $\tilde{y} \in \mathrm{Fix}(\tilde{f})$. In this case v is the trivial path. □

Remark 2.3. For use in Section 6 we can restate Corollary 2.2 as follows: Suppose that $v \subset G$ is a path from some y to $f(y)$ and that the fixed subgroup $\mathrm{Fix}((f_v)_\#)$ for the automorphism $(f_v)_\#\colon \pi_1(G, y) \to \pi_1(G, y)$ determined by f and v has rank at least two. Then there is a path α from y to some $z \in \mathrm{Fix}(f)$ such that $[f(\alpha)] = [\bar{v}\alpha]$. □

3. The irreducible case

In this section we prove a strong version of the Scott conjecture for irreducible automorphisms.

THEOREM 3.1. *If* $\Phi\colon F_n \to F_n$ *is irreducible*, *then* $\mathrm{Rank}(\mathrm{Fix}(\Phi)) \leq 1$.

Although the Scott conjecture concerns the automorphisms $\Phi\colon F_n \to F_n$, it is more convenient for us to consider the outer automorphism $\mathcal{O}$ determined by Φ. Unless specifically stated otherwise, $f\colon G \to G$ will be an irreducible topological representative of $\mathcal{O}$ that is a train track map and is normalized as follows. The eigenvalue λ associated to $\mathcal{O}$ is the Perron–Frobenius eigenvalue for the transition matrix M of $f\colon G \to G$. The graph G is equipped with a metric in which the i^{th} edge is isometric to an interval of length v_i, where v_i is the i^{th} coordinate of a positive vector $\vec{v}$ satisfying $\vec{v}M = \lambda\vec{v}$. With respect to this metric, f linearly expands each edge by the factor λ. We refer to such an $f\colon G \to G$ as an *irreducible train track representative* of $\mathcal{O}$. We denote the associated length function on paths $\sigma \subset G$ by $L(\sigma)$. If σ is legal, then $L(f(\sigma)) = \lambda L(\sigma)$.

We say that two irreducible train track representatives $f_1\colon G_1 \to G_1$ and $f_2\colon G_2 \to G_2$ are *projectively equivalent* if there is a homeomorphism $h\colon G_1 \to G_2$ that maps edges to edges, linearly expanding each by the same factor, such that $hf_1 = f_2 h$. Note in particular that if $f_1\colon G_1 \to G_1$ and $f_2\colon G_2 \to G_2$ differ only by the choice of the metric on $G_1 = G_2$, then they are projectively equivalent.

We assume without loss that $n > 1$. If $\lambda = 1$, then by Theorem 1.5, M is a permutation matrix and $f\colon G \to G$ is a homeomorphism. Since M is irreducible, f does not setwise fix any edges. Thus for any $y \in \mathrm{Fix}(f)$ the automorphism $f_\#$ of $\pi_1(G, y)$, determined by f and the trivial path at y, has a trivial fixed subgroup. By Corollary 2.2 this implies Theorem 3.1. We may therefore assume that $\lambda > 1$.

Recall that a path $\sigma\colon [0,1] \to G$ is assumed to be an immersion. Given $f\colon G \to G$, every path σ has a unique decomposition into maximal length legal subpaths $\sigma = \sigma_1 \cdots \sigma_k$, where each σ_i is a legal path and the initial edges of $\bar{\sigma}_i$ and σ_{i+1} form an illegal turn. (In the future we will abuse notation and refer to this turn as $\{\bar{\sigma}_i, \sigma_{i+1}\}$.) The image $f(\sigma_1) \cdots f(\sigma_k)$ may contain degenerate turns, and so can be written $\mu_1 \cdot \tau_1 \cdot \bar{\tau}_1 \cdot \mu_2 \cdots \bar{\tau}_{l-1} \cdot \mu_l$, where each μ_i is a legal path and the turns $\{\bar{\mu}_i, \mu_{i+1}\}$ are nondegenerate, but not necessarily illegal. Since the image of a legal path is legal, $l \le k$. We say that the concatenation $\mu_1 \cdots \mu_l$ is the *path determined by* $f(\sigma)$ and denote it by $[f(\sigma)]$.

A path ρ between $x, y \in \mathrm{Fix}(f)$ is a *Nielson path* if $f(\rho) \simeq \rho$ rel endpoints; it is *indivisible* if it cannot be written as a nontrivial concatenation $\rho = \rho_1 \cdot \rho_2$, where ρ_1 and ρ_2 are subpaths of ρ that are Nielsen paths.

Remark 3.2. If $y \in \mathrm{Fix}(f)$ and $f_\#$ is the automorphism of $\pi_1(G, y)$ determined by f and the trivial path at y (i.e., $f_\#(\langle \alpha \rangle) = \langle f(\alpha) \rangle$), then every element of $\mathrm{Fix}(f_\#)$ is represented by a Nielsen path from y to itself. □

We will establish the following result through a series of lemmas. The second part of the proposition is used in Section 4.

PROPOSITION 3.3. *There is an irreducible train track representative $f: G \to G$ of $\mathcal{O}$ that supports, at most, one indivisible Nielsen path ρ (up to reversal of orientation of ρ). If the endpoints of ρ are equal, then it forms a loop that crosses every edge of G exactly twice.*

Choose $f: G \to G$ as in Proposition 3.3. An immediate corollary of this proposition and Remark 3.2 is that if $y \in \mathrm{Fix}(f)$ and $f_\#$ is the automorphism of $\pi_1(G, y)$ determined by f and the trivial path at y, then $\mathrm{Rank}(\mathrm{Fix}(f_\#)) \leq 1$. Combined with Corollary 2.2, this implies Theorem 3.1. It therefore suffices to prove Proposition 3.3. The first step is to analyze indivisible Nielsen paths.

LEMMA 3.4. *An indivisible Nielsen path ρ contains exactly one illegal turn. In particular there are unique, nontrivial, legal paths α, β and τ such that $\rho = \alpha\beta$, $f(\alpha) = \alpha\tau$, $f(\beta) = \bar{\tau}\beta$ and such that $\{\bar{\alpha}, \beta\}$ is a nondegenerate turn; see Figure 6.*

Proof. Decompose $\rho = \rho_1 \cdot \cdots \cdot \rho_k$ into maximal length legal segments. As discussed above, $f(\rho) = \mu_1 \cdot \tau_1 \cdot \bar{\tau}_1 \cdot \mu_2 \cdots \bar{\tau}_{l-1} \cdot \mu_l$ for legal paths μ_i and τ_i. Since the image of a legal path of length L is a legal path of length $\lambda L > L$, ρ is not legal; thus $k > 1$. Suppose $k > 2$. Since $[f(\rho)] = \rho$, $l = k$, $\rho_i = \mu_i$ and $f(\rho_i) = \bar{\tau}_{i-1}\rho_i\tau_i$ for $1 \leq i \leq k$, where τ_0 and τ_k are trivial paths and the other τ_i's are nontrivial. For $1 < i < k$ and all $j > 0$ inductively define $\rho_i^0 = \rho_i$ and ρ_i^{j+1} to be the subpath of ρ_i^j that satisfies $f(\rho_i^{j+1}) = \rho_i^j$. Then each $P_i = \bigcap_{j=0}^{\infty} \rho_i^j \in \rho_i$ is a fixed point for f, and the P_i's divide ρ into Nielsen subpaths. This contradicts the assumption that ρ is indivisible. $\square$

COROLLARY 3.5. *An irreducible train track map supports only finitely many indivisible Nielsen paths.*

Proof. For any indivisible Nielsen path $\rho = \alpha\beta$ and for τ as in Lemma 3.4, $L(\rho) = \lambda L(\rho) - 2L(\tau)$. The "bounded cancellation lemma" ([T1] or [Co]) im-

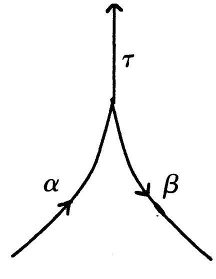

FIGURE 6

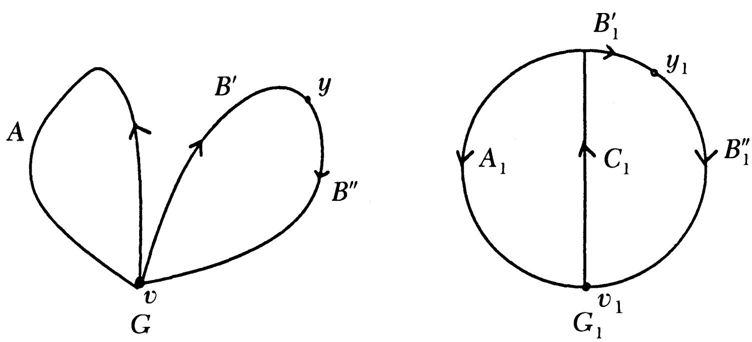

FIGURE 7

plies that $L(\tau)$ is bounded independently of the choice of ρ. Thus $L(\rho)$ is uniformly bounded. Since $\mathrm{Fix}(f)$ is finite, this proves the corollary. □

Theorem 1.7, Remark 1.8 and Lemmas 1.10 and 1.15 imply that if $f_1\colon G_1 \to G_1$ is obtained from $f\colon G \to G$ by folding a pair of edges e_i and e_j, then $f_1\colon G_1 \to G_1$ is an irreducible train track representation. (Note that, with respect to the quotient metric on G_1, f_1 linearly expands each edge by the factor λ.) When e_i is the initial edge of $\bar{\alpha}$ and e_j is the initial edge of β, where α and β are as in Lemma 3.4, we refer to this operation as *folding the indivisible Nielsen path* $\rho = \alpha \cdot \beta$.

Example 3.6. Let $G = R_2$ be the rose, with edges $\{A, B\}$ and vertex v, and let $f\colon G \to G$ be defined by $A \mapsto BA$, $B \mapsto BBA$. Then $\mathrm{Fix}(f) = \{v, y\}$, where $y \in \mathrm{int}(B)$. Subdivide B at y into $B'B''$ and note that $A \mapsto B'B''A$, $B \mapsto B'B''B'$ and $B'' \mapsto B''A$. Then $\rho = \overline{B}'A$ is an indivisible Nielsen path connecting y and v, as shown in Figure 7.

To fold ρ we identify the closure of the union of the first two components of $A \setminus f^{-1}(\mathscr{V})$ with the closure of the union of the first two components of $B' \setminus f^{-1}(\mathscr{V})$. This is a partial fold. The resulting graph G_1 has edges $\{A_1, B'_1, B''_1, C_1\}$, and $f_1\colon G_1 \to G_1$ is defined by $A_1 \mapsto C_1A_1$, $B'_1 \mapsto C_1B'_1$, $B''_1 \mapsto B''_1C_1A_1$ and $C_1 \mapsto C_1B'_1B''_1$. □

Suppose that $f_1\colon G_1 \to G_1$ is obtained from $f\colon G \to G$ by folding a pair of edges. The quotient map $p\colon G \to G_1$ restricts to a bijection between $\mathrm{Fix}(f)$ and $\mathrm{Fix}(f_1)$, because distinct elements of $\mathrm{Fix}(f)$ have distinct f-images and so cannot be identified by a fold; if ρ is an indivisible Nielsen path connecting x to y, then $[p(\rho)]$ is an indivisible Nielsen path connecting $x_1 = p(x)$ to $y_1 = p(y)$. We say that $\rho \subset G$ *determines* $\rho_1 \subset G_1$. Conversely for any indivisible Nielsen path $\rho_1 \subset G_1$ there is an indivisible Nielsen path $\rho \subset G$ such that $\rho_1 = [p(\rho)]$.

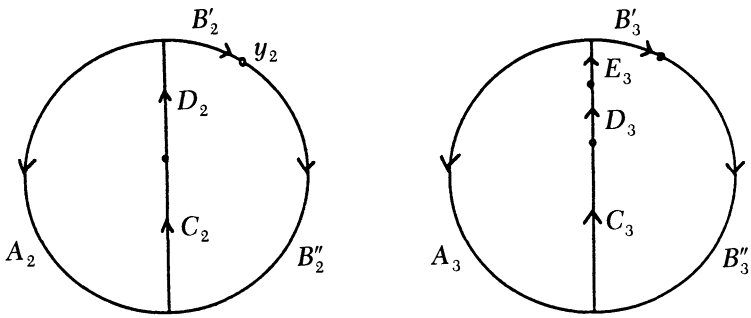

FIGURE 8

Any indivisible Nielsen path for $f: G \to G$ can be folded. Any indivisible Nielsen path for the resulting irreducible train track representative can also be folded. Denote by $\mathscr{W}(f)$ the set of irreducible train track representatives that can be realized by a finite sequence of such foldings.

Returning to Example 3.6, we see $\mathrm{Fix}(f_1) = \{v_1, y_1\}$ and $\rho_1 = \bar{B}_1' A_1$ is an indivisible Nielsen path between y_1 and v_1. Fold ρ_1 to obtain $f_2: G_2 \to G_2$, where G_2 has edges $\{A_2, B_2', B_2'', C_2, D_2\}$ and f_2 is defined by $A_2 \mapsto D_2 A_2$, $B_2' \mapsto D_2 B_2'$, $B_2'' \mapsto B_2'' C_2 D_2 A_2$, $C_2 \mapsto C_2 D_2 B_2' B_2''$ and $D_2 \mapsto C_2$. Now $\mathrm{Fix}(f_2) = \{v_2, y_2\}$ and $\rho_2 = \bar{B}_2' A_2$ is an indivisible Nielsen path between y_2 and v_2; see Figure 8. This can be repeated infinitely to produce $f_k: G_k \to G_k$, where G_k has k valence-two vertices. Thus $\mathscr{W}(f)$ determines infinitely many projective equivalence classes.

LEMMA 3.7. *If no partial folds occur in the construction of $\mathscr{W}(f)$, then $\mathscr{W}(f)$ contains only finitely many projective equivalence classes.*

Proof. If G_1 and G_2 are combinatorially equivalent, and if after identifying G_1 and G_2, we find $f_1(E) = f_2(E)$ for each edge E of the identified graph, then $f_1: G_1 \to G_1$ and $f_2: G_2 \to G_2$ are projectively equivalent. It therefore suffices to show that the graphs and transition matrices that occur in $\mathscr{W}(f)$ take on only finitely many values. This follows from the fact that there is a uniform bound for the number of vertices, and hence also for the number of edges, in the graphs that occur in $\mathscr{W}(f)$, and that there are only finitely many irreducible matrices of a given size and Perron–Frobenius eigenvalue. $\square$

Motivated by Lemma 3.7 we say that $f: G \to G$ is *stable* if no partial folds occur in the construction of $\mathscr{W}(f)$. Let $N(f)$ be the number of indivisible Nielsen paths supported by $f: G \to G$. The following lemma implies the existence of stable train track representatives of $\mathcal{O}$. The second part of the lemma will be used in Section 4.

LEMMA 3.8. *If $f\colon G \to G$ is an irreducible train track representative of $\mathcal{O}$ that is not stable, then there is an irreducible train track representative $f'\colon G' \to G'$ of $\mathcal{O}$ with $N(f') < N(f)$. If $f\colon G \to G$ supports a closed Nielsen path σ, then $f'\colon G' \to G'$ supports a closed Nielsen path in the free homotopy class determined by σ.*

Proof. Since $f\colon G \to G$ is not stable, we may assume, after some preliminary folding, that there is an indivisible Nielsen path $\rho \subset G$ so that the fold occurring for ρ is a partial fold.

Let α, β and τ be as in Lemma 3.4 and let v be the vertex from which $\bar{\alpha}$ and β originate. Suppose that α is not contained in a single edge of G. Since f linearly expands α over $\alpha\tau$, and since $f(\mathcal{V}) \subset \mathcal{V}$, the last edge in α maps entirely into τ. The only way that this edge will be subdivided during the folding of ρ is if the first edge of β maps entirely into τ and is shorter than the last edge of α. In that case the first edge of β is not subdivided. This contradicts the assumption that the fold is partial, and we conclude that α must be contained in a single edge. A similar argument holds for β.

If $\alpha\beta$ is a loop, then α and β are nonseparating, and so $\alpha\beta$ determines a free factor F_1 of F_n. Since $f(\alpha\beta) \simeq \alpha\beta$, F_1 is $\mathcal{O}$-invariant. This contradicts Lemma 1.2 and the assumption that $\mathcal{O}$ is irreducible. We conclude that $\alpha\beta$ is not a loop.

After subdividing if necessary, we may assume that α and β are edges in G. Fold ρ to obtain $f_1\colon G_1 \to G_1$ and let $\rho_1 \subset G_1$ be the indivisible Nielsen path determined by ρ. Fold ρ_1 to obtain $f_2\colon G_2 \to G_2$ and $\rho_2 \subset G_2$. Decompose

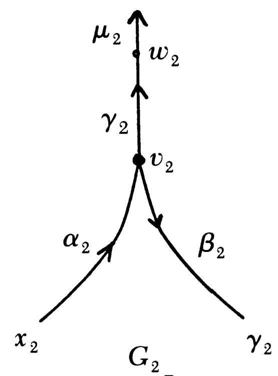

FIGURE 9

ρ_2 into maximal length legal paths $\rho_2 = \alpha_2\beta_2$. The vertex v_2, from which $\bar{\alpha}_2$ and β_2 originate, has valence three; let γ_2 be the third edge originating at v_2. The terminal vertex w_2 of γ_2 has valence two; let μ_2 be the second edge originating at w_2. The map f_2 satisfies $f_2(\alpha_2) = \alpha_2\gamma_2$, $f_2(\beta_2) = \bar{\gamma}_2\beta_2$ and $f_2(\gamma_2) = \mu_2$, as can be seen in Figure 9. Moreover $w_2 \notin f_2(\mathscr{V}_2 \setminus v_2)$ and $v_2 \notin f_2(\mathscr{V}_2)$.

We claim that the square matrix M_2', obtained from the transition matrix M_2 for $f_2: G_2 \to G_2$ by deleting the rows and columns corresponding to α_2 and β_2, is irreducible. To see this, let Γ_2 be the usual graph associated to M_2. Denote the vertices of Γ_2 corresponding to α_2, β_2 and γ_2 by v_α, v_β and v_γ. If v is a vertex of Γ_2 and there is an oriented edge connecting v to either v_α or v_β, then there is also an oriented edge connecting v to v_γ. Moreover v_γ is the only vertex, other than v_α and v_β, that is the terminal endpoint of an oriented edge originating at v_α or v_β. Since there are oriented paths connecting any oriented pair of vertices in Γ_2, there are oriented paths in $\Gamma_2 \setminus \{v_\alpha, v_\beta\}$ connecting any ordered pair of vertices in $\Gamma_2 \setminus \{v_\alpha, v_\beta\}$. This justifies our claim.

Recall that a valence-two homotopy has three steps: a homotopy followed by a tightening; the inverse of a subdivision; and the collapse of a pretrivial forest. Define $f_3: G_2 \to G_2$ by performing the first part of a valence-two homotopy and sliding w_2 across $\bar{\gamma}_2$ to v_2. Since v_2 is the only vertex that is mapped by f_2 to w_2, f_3 agrees with f_2 on all edges other than α_2, β_2 and γ_2. By construction, $f_3(\alpha_2) = \alpha_2$, $f_3(\beta_2) = \beta_2$ and $f_3(\gamma_2) = \gamma_2\mu_2$. The irreducibility of M_2' implies the irreducibility of the square submatrix M_3' of M_3 obtained by deleting the rows and columns corresponding to α_2 and β_2.

Since M_3' is irreducible, $\alpha_2 \cup \beta_2$ is the only proper invariant subgraph of G_2. By the irreducibility of $\mathscr{O}$, $\alpha_2 \cup \beta_2$ is a forest. Define $f': G' \to G'$ by collapsing $\alpha_2 \cup \beta_2$. The transition matrix M' equals M_3' and hence is irreducible. By Lemmas 1.15, 1.10 and Remark 1.14, $\lambda' = \lambda_2 = \lambda$; by Theorem 1.7 and Remark 1.8, $f': G' \to G'$ is an irreducible train track map.

We claim that $N(f) = N(f_1) = N(f_2) > N(f')$. The first two equalities follow from our previous observation that folding induces a bijection between Nielsen paths. For the final inequality note the homotopy that defines f_3 from f_2 is the identity when restricted to $\mathrm{Fix}(f_2)$ and, in particular, when restricted to the endpoints of any Nielsen path in G_2. It follows that there is a bijection between the indivisible Nielsen paths of f_3 and the indivisible Nielsen paths of f_2 other than $\alpha_2\bar{\beta}_2$. Since f' is obtained from f_3 by collapsing the Nielsen path $\alpha_2\bar{\beta}_2$, the desired inequality is established. This argument also shows that if $f: G \to G$ supports a closed Nielsen path σ, then $f': G' \to G'$ supports a closed Nielsen path in the free homotopy class determined by σ. $\square$

The following lemma completes the proof of Proposition 3.3.

LEMMA 3.9. *A stable train track map $f\colon G \to G$ supports at most one indivisible Nielsen path ρ (up to reversal of orientation of ρ). If $f\colon G \to G$ supports an indivisible Nielsen path ρ, then the illegal turn in ρ is the only illegal turn in G. If the endpoints of ρ are equal, then, as a loop, ρ crosses every edge of G exactly twice.*

Proof. We may assume that $f\colon G \to G$ supports at least one indivisible Nielsen path ρ. Let L be the length function on paths determined by the metric on G. Let $L(G)$ be the sum of the lengths of the edges of G.

If $f_1\colon G_1 \to G_1$ is obtained from $f\colon G \to G$ by folding ρ, then G_1 has an induced metric and an induced length function L_1. Since G_1 is obtained from G by identifying a pair of intervals of some equal length x, $L_1(G_1) = L(G) - x$. The indivisible Nielsen path $\rho_1 \subset G_1$ determined by $\rho \subset G$ satisfies $L_1(\rho_1) = L(\rho) - 2x$. If $L(\rho)/L(G) \neq 2$, then

$$\begin{aligned} |(L_1(\rho_1)/L_1(G_1)) - 2| &= |((L(\rho) - 2x)/(L(G) - x)) - 2| \\ &= |(L(\rho) - 2L(G))/(L(G) - x)| \\ &> |(L(\rho) - 2L(G))/L(G)| \\ &= |(L(\rho)/L(G)) - 2|. \end{aligned}$$

Iterating this, we see that there exist $f_i\colon G_i \to G_i$ in $\mathscr{W}(f)$ and indivisible Nielsen paths $\rho_i \subset G_i$ such that $L_i(\rho_i)/L_i(G_i)$ takes on infinitely many values.

On the other hand, for any $f'\colon G' \to G'$ in $\mathscr{W}(f)$, the set of ratios $\{L'(\rho')/L'(G')\colon \rho'$ is an indivisible Nielsen path for $f'\colon G' \to G'\}$ is a finite set that depends only on the projective equivalence class of $f'\colon G' \to G'$. By Lemma 3.7 there are only finitely many projective equivalence classes in $\mathscr{W}(f)$. This contradiction implies that $L(\rho) = 2L(G)$.

Suppose that T is an illegal turn in G that is distinct from the illegal turn in ρ. Let l be the smallest value for which $Tf^l(T)$ is degenerate. Suppose at first that $l = 1$. Fold T to obtain a train track map $\hat{f}\colon \hat{G} \to \hat{G}$ and let $\hat{\rho} \subset \hat{G}$ be the indivisible Nielsen path determined by ρ. Then $\hat{L}(\hat{G}) < L(G)$ and $\hat{L}(\hat{\rho}) = L(\rho)$ so that $\hat{L}(\hat{\rho})/\hat{L}(\hat{G}) > 2$. If we fold $\hat{\rho}$ to obtain $\hat{f}_1\colon \hat{G}_1 \to \hat{G}_1$ and $\hat{\rho}_1$, then the calculation in the second paragraph of this proof shows that $\hat{L}_1(\hat{\rho}_1)/\hat{L}_1(\hat{G}_1) > 2$. In particular $\hat{\rho}_1$ contains at least three edges. As we showed in the proof of Lemma 3.8, this implies that the fold at $\hat{\rho}_1$ is not partial. Iterating this argument proves that $\hat{f}\colon \hat{G} \to \hat{G}$ is stable. But this contradicts our previous arguments and the fact that $\hat{L}(\hat{\rho})/\hat{L}(\hat{G}) \neq 2$. We conclude that $l \neq 1$ and $Tf^{l-1}(T)$ is the illegal turn in ρ. Replacing T by $Tf^{l-2}(T)$, we may assume that $l = 2$. Fold ρ to obtain a stable train track map $f_1\colon G_1 \to G_1$. Let ρ_1 be the indivisible Nielsen path in G_1 determined by ρ and let T_1 be the turn in G_1 determined by T. Let

l_1 be the smallest value for which $Tf_1^{l_1}(T_1)$ is degenerate. Then $l_1 = 1$, and we reach a contradiction as in the previous case. We have now shown that the illegal turn in ρ is the only illegal turn in G.

Inductively define f_k: $G_k \to G_k$ and ρ_k by folding $\rho_{k-1} \subset G_{k-1}$, starting with f_0: $G_0 \to G_0$ equal to f: $G \to G$ and $\rho_0 = \rho$. For all $k \geq 0$, $L_{k+1}(G_{k+1}) = L_k(G_k) - x_k$. Since $\mathscr{W}(f)$ contains only finitely many projective equivalence classes, $x_k/L_k(G_k)$ is uniformly bounded below. It follows that $L_k(G_k) \to 0$, and hence that $L_k(e_k) \to 0$ for each edge $e_k \subset G_k$. In particular each edge of G must be crossed at least once by ρ.

Suppose that ρ is a loop. If ρ crosses some edge only once, then it must be a nonseparating edge. In this case ρ determines a free factor for F_n that is preserved, up to conjugacy, by $\mathscr{O}$. This contradicts Lemma 1.2. Thus ρ crosses every edge at least twice. But $L(\rho) = 2L(G)$. Thus ρ crosses every edge exactly twice.

Suppose that $\rho' = \alpha'\beta'$ is another indivisible Nielsen path for f: $G \to G$. It remains to show that $\rho = \rho'$. Since there is only one illegal turn in G, ρ and ρ' have the same illegal turn. After reorienting ρ' if necessary, we may assume that the initial edge of $\bar{\alpha}$ equals that of $\bar{\alpha}'$ and that the initial edge of β equals that of β'. We denote by ρ_k' the indivisible Nielsen path in G_k determined by ρ'. Let $\rho_k = \alpha_k\beta_k$ and $\rho_k' = \alpha_k'\beta_k'$ be as in Lemma 3.4.

Suppose that $\alpha \neq \alpha'$. If the initial endpoints of α and α' are distinct, then the initial endpoints of α_k and α_k' are distinct. If the initial endpoints of α and α' are equal, then $\alpha\bar{\alpha}'$ is a nontrivial loop, and each $\alpha_k\bar{\alpha}_k'$ is a nontrivial loop. In either case $\alpha_k \neq \alpha_k'$ for all k. For each $k \geq 0$ let ν_k be the maximum, common, terminal subinterval of α_k and α_k' and let μ_k and μ_k' be the remaining initial segments of $\alpha_k = \mu_k\nu_k$ and $\alpha_k' = \mu_k'\nu_k$.

Suppose that some $\{\bar{\mu}_k, \mu_k'\}$ is not equal to $\{\bar{\alpha}_k, \beta_k\}$. Then $L_{k+1}(\mu_{k+1}) = L_k(\mu_k)$. But $\{\bar{\alpha}_k, \beta_k\} \neq \{\bar{\mu}_k, \mu_k'\}$ implies that $\{\bar{\alpha}_{k+1}, \beta_{k+1}\} \neq \{\bar{\mu}_{k+1}, \mu_{k+1}'\}$. This

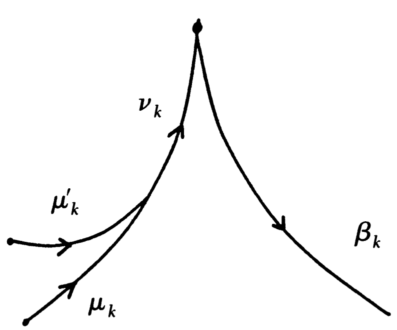

FIGURE 10

implies that $L_{k+i}(\mu_{k+i}) = L_k(\mu_k)$ for all $i > 0$, in contradiction to the fact that $L_k(\rho_k) = 2L_k(G_k) \to 0$; see Figure 10.

We may therefore assume that, for all $k \geq 0$, the turns $\{\bar{\alpha}_k, \beta_k\}$ and $\{\bar{\mu}_k, \mu'_k\}$ are equal. Then ν_{k+1} is constructed from ν_k by deleting a terminal interval of length x_k and adding an initial interval of length x_k. In particular $L_{k+1}(\nu_{k+1}) = L_k(\nu_k)$. As above, this contradicts the fact that $L_k(\rho_k) \to 0$. □

4. Geometric automorphisms

The fundamental group of a compact surface M with nonempty boundary is free and finitely generated. A homeomorphism $h: M \to M$ induces an outer automorphism of $\pi_1(M)$, and hence an outer automorphism of a finitely generated free group. We say that an outer automorphism $\mathcal{O}$ of F_n is *geometric* if it is induced by a surface homeomorphism in this way: we say that $h: M \to M$ *geometrically represents* $\mathcal{O}$. Stallings [St2] and Fried [Fr] gave the first examples of nongeometric outer automorphisms. Gersten [Ge3] showed that, in a certain sense, most automorphisms are not geometric.

We will restrict our attention to the surfaces M with connected boundary and to pseudo-Anosov homeomorphisms ([T2], [FLP], [BH1]). The boundary curve $\sigma \subset M$ satisfies $h(\sigma) = \sigma$ or $h(\sigma) = \bar{\sigma}$. If $h: M \to M$ geometrically represents $\mathcal{O}$, then σ determines a conjugacy class s of words in F_n that satisfies $\mathcal{O}(s) = s$ or $\mathcal{O}(s) = \bar{s}$. As we remarked (Example 1.4) in Section 1, each $\mathcal{O}^l$ is irreducible. Our next theorem states that the converse holds; it was suggested to us by Bill Thurston after learning of Proposition 3.3.

THEOREM 4.1. *Suppose that each $\mathcal{O}^l$ is irreducible and that there is a cyclic word $s \in F_n$ such that $\mathcal{O}(s) = s$ or $\mathcal{O}(s) = \bar{s}$. Then $\mathcal{O}$ is geometrically realized by a pseudo-Anosov homeomorphism $h: M \to M$ of a compact surface with one boundary component.*

Remark 4.2. The hypothesis on $\mathcal{O}$ can be weakened as follows: Given an irreducible train track representative $f: G \to G$ of $\mathcal{O}$, let k be the number of illegal turns contained in the closed loop $\sigma \subset G$ that represents s. If $\mathcal{O}^k$ is irreducible, then $\mathcal{O}$ can be geometrically realized. If, in addition, $\mathcal{O}^l$ is irreducible for each $1 \leq l \leq n-1$, then the geometric realization must be pseudo-Anosov. □

Remark 4.3. A pseudo-Anosov mapping class (see [T2], [FLP] or [BH1]) does not act periodically on any nonperipheral, free homotopy class of closed curves. Thus s and its multiples are the only cyclic words in F_n that satisfy $\mathcal{O}^k(s) = s$ for some $k > 0$. It follows that $\mathcal{O}$ cannot be geometrically realized on a surface with more than one boundary component. □

Remark 4.4. Implicit in our construction is an algorithm to decide if a given outer automorphism is geometrically realized by a pseudo-Anosov homeomorphism of a compact surface with one boundary component. □

We recognize geometric automorphisms via the following proposition:

PROPOSITION 4.5. *Suppose that $f\colon G \to G$ is a homotopy equivalence representing an irreducible outer automorphism $\mathcal{O}$ and suppose that there is a closed path $\sigma \subset G$ that crosses every edge of G exactly twice and satisfies $f(\sigma) \simeq \sigma$ or $f(\sigma) \simeq \bar{\sigma}$. Then there is a closed surface M, with one boundary component, and a homeomorphism $h\colon M \to M$ that geometrically represents $\mathcal{O}$. If each $\mathcal{O}^l$ is irreducible, then h is pseudo-Anosov.*

Proof. Let M be the space obtained from G and an annulus A by gluing one boundary component of A to G via the path σ. Projecting along the radial lines of A gives a deformation retraction of M onto G. The map $f\colon G \to G$ extends to a homotopy equivalence $F\colon M \to M$ that restricts to a homeomorphism on the unattached component $\partial_1 A$ of A. We will show that M is a surface with boundary equal to $\partial_1 A$. It is well known (see [BC]) that $F\colon M \to M$ is homotopic to a homeomorphism $h\colon M \to M$. Obviously $h\colon M \to M$ geometrically represents $\mathcal{O}$. If h is not pseudo-Anosov, then some iterate h^l fixes, up to isotopy, either a simple, closed, nonseparating curve τ or a proper incompressible subsurface $M_0 \subset M$ with connected boundary. Since both τ and M_0 determine proper free factors of $\pi_1(M)$, h^l is reducible. It therefore suffices to show that M is a surface.

Since σ crosses every edge of G exactly twice, M is locally euclidean away from the vertices of G. There is an equivalence relation on the oriented edges originating at a vertex v generated by the rule that $E_1 \sim E_2$ if $\{\bar{E}_1, E_2\}$ is a turn in σ. Let l_v be the number of equivalence classes that occur at the vertex v. Then a neighborhood $U(v) \subset M$ of v is a cone on l_v circles. Thus $(U(v) \setminus v)$ is homeomorphic to the product of a half-open interval with a disjoint union of l_v circles. If we make $(U(v) \setminus v)$ compact again by adding in l_v points, then the result is l_v disjoint disks. Carrying out this operation at each vertex v with $l_v > 1$, we produce a compact surface S. In other words M is obtained from a surface S by identifying finite collections of points to single points.

Let G' be the graph that is obtained from G by replacing each vertex $v \in G$ with l_v vertices; each equivalence class of oriented edges originating at v is incident to its own vertex in G'. The graph G' has the same set of edges as G and, by definition of the equivalence relation, σ lifts into G'. Since σ crosses every edge of G', G' is connected. It is easy to check that S deformation retracts onto a copy of G'. In particular S is connected.

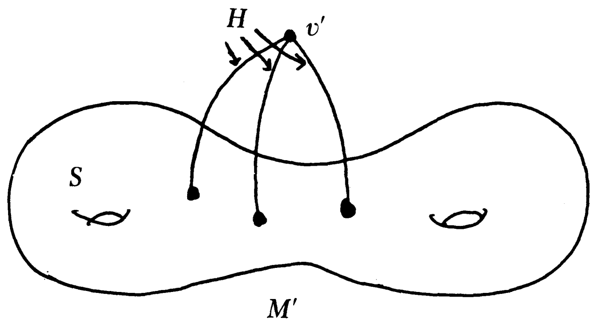

FIGURE 11

For notational simplicity we will assume that there is exactly one vertex v with $l_v > 1$. The more general case is left as a straightforward exercise for the reader. Thus M is obtained from S by identifying l_v points to a single point. Instead of actually identifying these l_v points together, we add an extra vertex v' and connect each of the l_v points in S to v' by an edge. We denote the resulting space by M' and the union of the new edges by H, as in Figure 11. Then H is a tree, and collapsing it to a point gives M. The homotopy equivalence $F\colon M \to M$ induces a homotopy equivalence $F'\colon M' \to M'$ whose restriction to $\partial M' = \partial S = \partial_1 A$ is a homeomorphism.

For each edge $e \subset H$ choose a point p_e in the interior of e. We may assume that $(F')^{-1}(p_e)$ is a union of finitely many points in H with a finite collection of simple closed curves in S. Since F' is a homotopy equivalence, each simple closed curve in $(F')^{-1}(p_e)$ is null homotopic in S. We may therefore perform a homotopy to produce $F''\colon M' \to M'$ with each $(F'')^{-1}(p_e) \subset H$. Performing a further homotopy, we can arrange it so that $F''(S) \subset S$. But S is connected and carries σ; so $\pi_1(S)$ is an invariant, nontrivial, free factor of $\pi_1(M')$. This contradicts Lemma 1.2, and we conclude that each $l_v = 1$. □

The hypotheses of Proposition 4.5 coincide with the conclusions of Lemma 3.9. In order to apply Lemma 3.9, we will use the following lemma:

LEMMA 4.6. *If $f\colon G \to G$ is an irreducible train track map and $\sigma \subset G$ is a loop that satisfies $f(\sigma) \simeq \sigma$, then there exist $k > 0$ and a choice of basepoint in σ so that σ is a closed Nielsen path for $f^k\colon G \to G$.*

Proof. Since f expands the length of legal paths, σ contains an illegal turn. Choose a cyclic decomposition $\sigma = \sigma_1 \cdots \sigma_k$ into maximal length legal subpaths, where (σ_k, σ_1) is an illegal turn. This decomposition is unique up to rotation, i.e., up to renaming σ_i by σ_{i+r} for some $0 \le r \le k-1$, where indices are mod k. There are subintervals $\sigma_i' \subset \sigma_i$ and legal paths μ_i such that $f(\sigma) = \bar{\mu}_1 \cdot f(\sigma_1') \cdot \mu_2 \cdot \bar{\mu}_2 \cdot f(\sigma_2') \cdots f(\sigma_k') \cdot \mu_1$. Since $[f(\sigma)] = \sigma$, $f(\sigma_i') = \sigma_{i+r}$ for some $0 \le r \le k-1$. As in the proof of Lemma 3.4 there are fixed points for f^k in each σ_i that divided σ into a concatenation of indivisible Nielsen subpaths for f^k. □

Proof of Theorem 4.1. Choose an irreducible train track representative $f\colon G \to G$ for $\mathcal{O}$. Let $\sigma \subset G$ be the closed loop in the free homotopy class determined by $s \in F_n$. Apply Lemma 4.6 to $f\colon G \to G$ to find $f^k\colon G \to G$ so that σ is a closed Nielsen path for $f^k\colon G \to G$. Apply Lemmas 3.8 and 3.9 to find an irreducible train track representative $f'\colon G' \to G'$ of $\mathcal{O}^k$ that supports exactly one indivisible Nielsen path ρ' and that supports a closed Nielsen path σ' in the free homotopy class determined by $s \in F_n$. Obviously ρ' must be a closed path and σ' is a multiple of ρ' or $\bar{\rho}'$. By Lemma 3.9, ρ' crosses every edge of G' exactly twice.

Let $r \in F_n$ be a representative of the conjugacy class determined by ρ'. Replacing s by some conjugate word if necessary, we have $s = r^m$ for some $m \neq 0$. Since $\mathcal{O}(s) = s$ or $\mathcal{O}(s) = \bar{s}$, we conclude that $\mathcal{O}(r) = r$ or $\mathcal{O}(r) = \bar{r}$. Let $f''\colon G' \to G'$ be a homotopy equivalence that represents $\mathcal{O}$. Then $f''(\rho')$ is freely homotopic to either ρ' or $\bar{\rho}'$, and Theorem 4.1 follows from Proposition 4.5. □

5. Relative train track maps

In Section 1 we defined an algorithm which produces either a train track map or a reduction for a given outer automorphism $\mathcal{O}$. There we assumed that $\mathcal{O}$ was irreducible, so we always found a train track map. In this section we extend the algorithm so that it finds a good representative of $\mathcal{O}$ in all cases. The exact choice of a good representative is tailored to our needs in the Scott conjecture. For other problems the algorithm can be refined (see [BH2] where we are more specific about the nonexponentially growing strata).

A *filtration* for a topological representative $f\colon G \to G$ is an increasing sequence of (not necessarily connected) invariant subgraphs $\emptyset = G_0 \subset \cdots \subset G_m = G$. The subcomplex $\mathrm{cl}(G_i \setminus G_{i-1})$ is denoted by H_i and referred to as the i^{th} *stratum*. A turn with one edge in H_i and one edge in G_{i-1} is said to be a *mixed turn* in (G_i, G_{i-1}). We may assume that the edges of G are labelled so that those in H_i precede those in H_{i+1}. The edges of H_i determine a square submatrix M_i of the transition matrix M called the *transition submatrix* for H_i. We say that the *filtration is maximal* if each M_i is either irreducible or the zero matrix. We will assume, unless otherwise stated, that all filtrations are maximal.

A topological representative $f\colon G \to G$ determines a maximal filtration $\emptyset = G_0 \subset \cdots \subset G_m = G$ as follows. Let $M = (a_{ij})$ be the transition matrix for $f\colon G \to G$ and let Γ be the graph with a vertex for each edge in G and a_{ij} oriented edges from the j^{th} vertex v_j to the i^{th} vertex v_i. Then two edges $e_1, e_2 \subset G$, with corresponding vertices $v_1, v_2 \subset \Gamma$, belong to the same stratum of the filtration if and only if there is an oriented path in Γ originating at v_1 and terminating at v_2 and an oriented path in Γ originating at v_2 and terminating at v_1. There is a partial ordering on the strata defined by $H_j > H_i$ if there is a vertex v_j corresponding to an edge in H_j, a vertex v_i corresponding to an edge in H_i and an oriented path in Γ originating at v_j and terminating at v_i. Reorder the edges of G so that if $H_j > H_i$, then the edges of H_i precede those of H_j. The reordering is not always uniquely defined, but our constructions depend only on the set of M_i's that occur and on the partial ordering on the H_i's.

To measure the efficiency of $f\colon G \to G$ we assign an invariant Λ to $\emptyset = G_0 \subset \cdots \subset G_m = G$ as follows: If the transition submatrix M_i corresponding to H_i is irreducible, then it has a Perron–Frobenius eigenvalue λ_i. If $\lambda_i > 1$, then we say that H_i is *exponentially growing*. Define Λ to be the nonincreasing sequence $\lambda_{i_1} \geq \lambda_{i_2} \geq \cdots \geq \lambda_{i_r}$ of Perron–Frobenius eigenvalues that occur for the exponentially growing strata of the filtration. We order the Λ's lexicographically; thus Λ is reduced when, for example, some λ_i is replaced by any number of smaller eigenvalues and all other λ_j's are unchanged.

We will consider various operations that transform a topological representative $f\colon G \to G$ into another topological representative $f'\colon G' \to G'$. If $\Lambda' \leq \Lambda$, then we say the operation is *safe*. For example (see Lemma 5.1), subdivision is a safe operation. If $\Lambda' < \Lambda$, then we say the operation is *beneficial*; certain valence-one homotopies are beneficial (Lemma 5.2). If the operation is not safe, then we say it is *dangerous*. As we shall see (Lemma 5.4), certain valence-two homotopies are dangerous.

Our first lemma shows that subdivision is safe.

LEMMA 5.1. *If $f'\colon G' \to G'$ is obtained from $f\colon G \to G$ by subdivision, then* $\Lambda' = \Lambda$.

Proof. Suppose that an edge e in H_i is being subdivided into e_1 and e_2. This has no effect on the transition submatrices M_j except possibly for M_i. We may assume that M_i is irreducible. If neither e_1 nor e_2 maps entirely into G_{i-1}, then arguing exactly as in Lemma 1.10, we see that M_i is replaced by an irreducible matrix M_i' with $\lambda_i' = \lambda_i$. If e_1 or e_2, say, e_1, maps entirely into G_{i-1}, then e_1 determines a new stratum whose transition submatrix is the zero matrix; the union of e_2 with the remaining edges of H_2 determines a stratum whose transition submatrix is exactly M_i. □

The next lemma states that valence-one homotopies are safe and are sometimes beneficial.

LEMMA 5.2. *Suppose that f': $G' \to G'$ is obtained from f: $G \to G$ by a valence-one homotopy at a vertex v. Then $\Lambda' \leq \Lambda$. If the stratum H_r, containing the edge E that is incident to v, is exponentially growing, then $\Lambda' < \Lambda$.*

Proof. A valence-one homotopy has two parts. First a homotopy, followed by a tightening, is performed so that the edge E is no longer in the image of the map. Then a pretrivial forest is collapsed, and if E has not been collapsed, then the map is restricted to the complement of E. Orient E so that v is its initial endpoint. For any edge e of G other than E, the effect of the homotopy and tightening on the path $f(e)$ is to remove the initial edge if it is E and the terminal edge if it is $\overline{E}$ (because these are the only ways that E and $\overline{E}$ can occur in a path). When the pretrivial forest is collapsed, paths are changed by removing all occurrences of the edges in the forest. We conclude that the transition matrix M' for f': $G' \to G'$ is obtained from the transition matrix M for f: $G \to G$ by removing rows and columns corresponding to E and to the pretrivial forest that is collapsed.

The total set of edges to be collapsed can be described from f: $G \to G$ as follows: First E is collapsed. Then we inductively collapse any edge that maps entirely into the subgraph consisting of edges that are already collapsed. Note that if e is an edge that is not in H_r, and if there exists $k > 0$ such that the path determined by $f^k(e)$ contains e, then e will not be collapsed during this process. Thus only edges in H_r or in strata with transition submatrices equal to zero may be collapsed. We conclude for $l \neq r$ that: if M_l is the zero matrix, then the edges in H_l (if any) that are not collapsed determine a stratum with transition submatrix equal to zero; if M_l is irreducible, then the edges in H_l determine a stratum with transition submatrix equal to M_l.

The edges of H_r that are not collapsed may determine one or more strata. The key point is that if the transition submatrices associated to any of these strata are irreducible, then Remark 1.12 implies that their associated Perron–Frobenius eigenvalues are strictly less than λ_i. Lemma 5.2 now follows from the definition of Λ and the lexicographical ordering. □

Our next lemma states that folding is safe and sometimes beneficial.

LEMMA 5.3. *Suppose that f': $G' \to G'$ is obtained from the topological representative f: $G \to G$ by an elementary fold of edges $E_i \subset H_i$ and $E_j \subset H_j$, $i \leq j$, and that the topological representative f'': $G'' \to G''$ is obtained from f': $G' \to G'$ by tightening and the collapse of a pretrivial forest. If $H_i = H_j$ is* *exponentially growing, and if nontrivial tightening occurs for some edge in the stratum of G' determined by H_i, then $\Lambda'' < \Lambda$. In all other cases $\Lambda'' = \Lambda$.*

Proof. To avoid special notational cases we denote the image of E_i and E_j in G' by E_i; thus the edges in G' and in G'' have the same labels as edges in G.

We consider first an edge $e \subset H_l$, where $l \neq i, j$. The effect on $f(e)$ of folding and tightening is first to remove all occurrences of $\overline{E}_i E_j$ and $\overline{E}_j E_i$ and then to change all remaining E_j's and $\overline{E}_j$'s to E_i's and $\overline{E}_i$'s, respectively. No further tightening is required, because the endpoints of $\overline{E}_i E_j$ are distinct (c.f. the collapse of a pretrivial forest). Since $l \neq i, j$, these changes have no effect on the transition submatrix for the edges of H_l. If M_l is the zero matrix, then the edges of H_l that are not part of the pretrivial forest determine a stratum of G'' with transition submatrix equal to zero. If M_l is irreducible, then (c.f. the proof of Lemma 5.2) no edge of H_l will be part of the pretrivial forest; we conclude that the edges of H_l determine an irreducible stratum in G'' with transition submatrix equal to M_l.

If $j > i$, then M_j must be the zero matrix, since $f(e_j) \subset G_i$. If any edges of H_j are not folded away or collapsed, then they determine a stratum of G'' with transition submatrix equal to zero. Moreover if $j > i$, then H_i is unaffected by the entire process; we conclude that $\Lambda'' = \Lambda$.

It remains to analyze the effect on H_i when $i = j$. If H_i is not exponentially growing, then it does not produce any exponentially growing strata of G''. We may therefore assume that H_i is exponentially growing. Arguing exactly as in Lemma 1.15 (where the irreducible case is considered), we conclude that the edges of H_i determine a unique stratum H_i' of G' and that the transition matrix M_i' for H_i' is irreducible with $\lambda_i' = \lambda_i$. If the f' images of edges in H_i' are tight, then no edges in H_i' are collapsed, and $M_i'' = M_i'$. In this case $\lambda_i'' = \lambda_i$. If tightening does occur, then M_i'' is obtained from M_i' by lowering some entries and possibly removing some rows and columns. Remarks 1.12 and 1.6 imply that any exponentially growing stratum of G'' determined by edges in H_i' will have Perron–Frobenius eigenvalues that are strictly smaller than λ_i'. The lemma now follows from the definition of Λ and the lexicographical ordering. □

We now come to the one dangerous operation, the valence-two homotopy. Suppose that $E_i \subset H_i$ and $E_j \subset H_j$ are the edges incident to a valence-two vertex v and that $i \leq j$. If $i = j$ and H_i is exponentially growing, then we relabel if necessary so that the eigenvector coefficient of E_i (see Lemma 1.13) is greater than or equal to that of E_j. In all cases we perform the homotopy across E_i. The following lemma shows that a valence-two homotopy is dangerous if $i = j$ and H_i is exponentially growing, but is safe otherwise.

LEMMA 5.4. *Suppose that* $f'\colon G' \to G'$ *is obtained from* $f\colon G \to G$ *by a valence-two homotopy with restrictions as above. If* H_i *is not exponentially growing, then* $\Lambda' = \Lambda$. *If* $i < j$ *and* M_i *is exponentially growing, then* $\Lambda' < \Lambda$. *If* $i = j$ *and* H_i *is exponentially growing, then* Λ' *is obtained from* Λ *by replacing* λ_i *with some number of* λ'_{ij}*'s, each satisfying* $\lambda'_{ij} \le \lambda_i$.

Proof. The proof is similar to that of Lemmas 5.2 and 5.3. There are three steps involved in constructing $f'\colon G' \to G'$. They are: a homotopy with support in $E_i \cup E_j$ followed by a tightening; an inverse subdivision (producing an edge that we call E_j) followed by a tightening; and the collapse of a pretrivial forest.

These steps have no effect at all on G_{i-1}. It is easy to check that strata with transition submatrices equal to zero do not produce strata with irreducible transition matrices.

Choose an edge $e \subset H_l$, where $l > i$, $l \neq j$, and H_l is exponentially growing. The effect of the first two steps on $f(e)$ is to remove all occurrences of E_i and $\overline{E}_i$. Since edges in strata with irreducible transition submatrices are not contained in a pretrivial forest, $f''(e)$ is obtained from $f(e)$ by erasing all occurrences of E_i, $\overline{E}_i$ and the edges that are collapsed in the third step. This does not change the transition submatrix and so has no effect on λ_l or Λ.

Suppose that $i < j$, that M_j is irreducible and that e is an edge in H_j. If $e \neq E_j$, then the analysis of the previous case applies to $f(e)$; i.e., $f''(e)$ is obtained from $f(e)$ by erasing all occurrences of E_i, $\overline{E}_i$ and the edges that are collapsed in the third step. If f is locally injective at v, then the effect of steps one and two on $f(E_j)$ is first to replace $f(E_j)$ with $f(E_i)f(E_j)$ (which adds no new edges in H_j) and then to remove all occurrences of E_i and $\overline{E}_i$. If f is not locally injective at v, then some further tightening may be required, but only edges in G_i can be cancelled, since $f(E_j)$ is tight and $f(E_i) \subset G_i$. We conclude that, after the first two steps, the edges of H_j determine a stratum with transition submatrix equal to M_j. As above, the third step does not collapse any edges of this stratum and does not change the transition submatrix. Thus the edges of H_j determine a stratum H'_j with $\lambda'_j = \lambda_j$.

It remains to consider H_i. If $i < j$, then the effect on the edges of H_i is exactly that of a valence-one homotopy, because v is a valence-one vertex of G_i. As we argued in Lemma 5.2, if M_i is exponentially growing, then λ_i is replaced by some number of λ'_{ij}'s with $\lambda'_{ij} < \lambda_i$; in this case Λ is decreased. If M_i is not exponentially growing, then Λ is unchanged.

Finally suppose that $i = j$ and that M_i is irreducible. We follow the proof of Lemma 1.13. After relabelling, we may assume that E_i and E_j are the m^{th} and $(m-1)^{\text{st}}$ edges of H_i, respectively. Define an $(m-1)$-dimensional matrix M'_i by adding the m^{th} column of $M_i = (a_{jk})$ to the $(m-1)^{\text{st}}$ column of M_i and by removing the m^{th} row and m^{th} column of M_i. If f is locally injective at v, then M_i' equals the transition submatrix $M_i(2)$ for the map called $f(2)$: $G(2) \to G(2)$ in the definition of valence-two homotopy. If f is not locally injective at v, then nontrivial tightening occurs when $f(2)$ is obtained from $f(1)$; so $M_i(2)$ is obtained from M_i' by reducing some of the entries in the $(m-1)^{\text{st}}$ column.

Choose a positive vector $\vec{w}$ satisfying $M_i\vec{w} = \lambda_i\vec{w}$ and define the $(m-1)$-dimensional vector $\vec{v}$ by $v_l = w_l$ for $1 \le l \le m-1$. Recall that, by our restriction on valence-two homotopies, $w_{m-1} \le w_m$. Then $(M_i(2)\vec{v})_l \le \lambda w_l - a_{lm}(w_m - w_{m-1}) \le \lambda w_l = \lambda v_l$ for all $1 \le l \le m-1$. If $M_i(2)$ is irreducible, then it is the transition submatrix for the stratum of G' determined by H_i; in this case Theorem 1.5 implies that $\lambda_i' \le \lambda_i$. If $M_i(2)$ is not irreducible, then the transition submatrix M_{ij}' for any stratum H_{ij}' of G', determined by some of the edges in H_i, is (after relabelling edges) a square submatrix of $M_i(2)$. Since $(M_i(2)\vec{v})_l \le \lambda_i v_l$ for all $1 \le l \le m-1$, Theorem 1.5 implies that each $\lambda_{ij} \le \lambda_i$. $\square$

This completes our analysis of the four basic operations.

A simple Euler characteristic shows that if f: $G \to G$ represents $\mathcal{O}$, and if G has no valence-one or valence-two vertices, then G has at most $L = 3n - 3$ edges. We say that f: $G \to G$ is *bounded* if there are at most L exponentially growing strata H_r, and if, for each such H_r, λ_r is the Perron–Frobenius eigenvalue for some matrix with no more than L rows and columns. Note that there is no restriction on the number of edges in the stratum H_r. Note also that Λ completely determines whether or not f: $G \to G$ is bounded.

Since the λ_i's can take on only finitely many values less than any given constant, Λ takes on a minimum value $\Lambda_{\min}$ among all bounded f: $G \to G$.

In Section 1 we used valence-one and valence-two homotopies to prove that, for any irreducible topological representative f: $G \to G$, there is an irreducible topological representative f': $G' \to G'$ such that G' has no valence-one or valence-two vertices and such that $\lambda' \le \lambda$. Since some valence-two homotopies are dangerous, we cannot prove the exact analogue of that statement. The following lemma is sufficient for our purposes.

LEMMA 5.5. *Suppose that f: $G \to G$ is a bounded topological representative and that f': $G' \to G'$ is obtained from f: $G \to G$ by a sequence of safe moves discussed in Lemmas* 5.1–5.4. *If $\Lambda' < \Lambda$, then there is a bounded topological representative f'': $G'' \to G''$ such that $\Lambda'' < \Lambda$.*

Proof. Our goal is to change f': $G' \to G'$ until the resulting topological representative is bounded, without raising Λ' so much as to become as large as Λ.

After performing valence-one homotopies and safe valence-two homotopies, we may assume that G' has no valence-one vertices and that if v is a valence-two vertex of G', then both edges of G' incident to v lie in the same stratum. The filtration associated to $f'\colon G' \to G'$ has at most L strata. Our analysis of safe operations shows that the set Λ' is obtained from Λ by replacing certain of the λ_i's, say, $\lambda_1, \ldots, \lambda_v$, with collections $\{\lambda'_{i1}, \ldots, \lambda'_{il_i}\}$, where each $\lambda'_{is} < \lambda_i$. We call the strata of G' corresponding to the λ'_{is}'s *new strata* and the strata of G' corresponding to $\{\lambda_i\colon i > v\}$ *old strata*.

If there is a valence-two vertex with both incident edges in a new stratum, perform a valence-two homotopy at such a vertex. By Lemma 5.4 this results in some λ'_{is} being replaced by a collection $\{\lambda''_{isu}\}$, where each $\lambda''_{isu} \leq \lambda'_{is} < \lambda_i$. Since the last inequality is strict, $\Lambda'' < \Lambda$. Iterating this, we obtain $f''\colon G'' \to G''$ such that $\Lambda'' < \Lambda$, the associated filtration has at most L strata and each new strata has at most L edges. To prove that $f''\colon G'' \to G''$ is bounded we need only check that the λ''_k's corresponding to the old strata are the Perron–Frobenius eigenvalues for matrices with at most L rows and columns. This follows from the definition of old strata and the assumption that Λ is bounded. □

COROLLARY 5.6. *If $f'\colon G' \to G'$ is obtained from $f\colon G \to G$ by a sequence of safe moves discussed in Lemmas* 5.1–5.4, *and if $\Lambda = \Lambda_{\min}$, then $\Lambda' = \Lambda_{\min}$.*

For the remainder of this section we will restrict our attention to those topological representations with $\Lambda = \Lambda_{\min}$. If $f'\colon G' \to G'$ is obtained from $f\colon G \to G$ by one of the safe operations already discussed, then each exponentially growing stratum H_r of G determines a unique exponentially growing stratum of G' that is denoted by H'_r. By Corollary 5.6, $\lambda'_r = \lambda_r$.

We say that $f\colon G \to G$ is a *relative train track map* if the following conditions hold for every exponentially growing stratum H_r:

(RTT-i) Df maps the set of oriented edges in H_r to itself; in particular all mixed turns in (G_r, G_{r-1}) are legal.

(RTT-ii) If $\alpha \subset G_{r-1}$ is a nontrivial path with endpoints in $H_r \cap G_{r-1}$, then $[f(\alpha)]$ is a nontrivial path with endpoints in $H_r \cap G_{r-1}$.

(RTT-iii) For each legal path $\beta \subset H_r$, $f(\beta)$ is a path that does not contain any illegal turns in H_r.

Remark. A topological representative $f\colon G \to G$ is irreducible if and only if there is only one stratum in its associated filtration. In this case (RTT-i) is automatic, (RTT-ii) is vacuous and (RTT-iii) is the train track property. Thus the relative train track property becomes the train track property when restricted to irreducible topological representatives. □

Example 5.7. Let $h: R_3 \to R_3$ be the homotopy equivalence of the rose R_3 defined by $A \mapsto BA$, $B \mapsto BBA$ and $C \mapsto C\bar{A}BA\bar{B}$. In the associated filtration, G_1 is the subgraph generated by A and B, and H_2 equals C. It is easy to check that $h: R_3 \to R_3$ is a relative train track map. We showed in Example 1.9 that the outer automorphism determined by $h: R_3 \to R_3$ cannot be represented by a train track map. □

The following lemma illustrates the usefulness of the relative train track property. A path with no illegal turns in H_r is said to be *r-legal.*

LEMMA 5.8. *Suppose that $f: G \to G$ is a relative train track map and that $\sigma = a_1 b_1 a_2 \dots b_l$ is the decomposition of an r-legal path σ into subpaths $a_j \subset H_r$ and $b_j \subset G_{r-1}$. (Allow for the possibility that a_1 or b_l is trivial, but assume the other subpaths are nontrivial.) Then $[f(\sigma)] = f(a_1)[f(b_1)]f(a_2)\dots[f(b_l)]$ and is r-legal.*

Proof. Since a_j is legal, $f(a_j)$ is already a path and cannot be tightened. By (RTT-iii), $f(a_j)$ is r-legal. By (RTT-ii) each $f(b_j)$ tightens to a nontrivial path $[f(b_j)]$ in G_{r-1}. By (RTT-i) the turns between $f(a_j)$ and $[f(b_j)]$ are nondegenerate. □

Part of the algorithm of Section 1 was used in the proof of Lemma 5.5. The rest of the algorithm is used in the following lemma:

LEMMA 5.9. *If $f: G \to G$ satisfies* (RTT-i), *and if $\Lambda = \Lambda_{\min}$, then $f: G \to G$ also satisfies* (RTT-iii).

Proof. Choose an exponentially growing stratum H_r. Suppose that e is an edge in H_r and that $f(e)$ contains a turn T_1 that is illegal, i.e., that for some $l \geq 1$, $T_l = (Tf)^{l-1}(T_1)$ is degenerate. By (RTT-i) each T_j is entirely in either H_r or G_{r-1}. We must show that $T_1 \subset G_{r-1}$.

Suppose that $T_1 \subset H_r$. Then (RTT-i) implies that $T_j \subset H_r$ for all j. We will apply part of the algorithm of Theorem 1.7: The first step is to subdivide e at the point P that maps to the vertex of T_1. Since $T_1 \subset H_r$, the edges of H_r determine only one stratum (which we continue to call H_r) for the resulting map; in particular P is a valence-two vertex with both incident edges in H_r. The next step is to perform a sequence of subdivisions. The exact definition of these subdivisions need not be repeated here. The key point (see the proof of Lemma 5.1) is that, by (RTT-i), the turns T_j lie in the one irreducible stratum (still called H_r) that is determined by the edges of H_r.

There are no more subtleties to consider. The next step in the algorithm is to use the safe moves discussed in Lemmas 5.2 and 5.3 to reduce λ_r and thereby reduce Λ. This contradicts the definition of $\Lambda_{\min}$ and Corollary 5.6. □

Lemma 5.10. *Suppose that $f: G \to G$ is a relative train track map and H_r is an exponentially growing stratum. Then there is a length function $L_r(\sigma)$ for paths $\sigma \subset G_r$ with the feature $L_r(f(\sigma)) = \lambda_r L_r(\sigma)$ for any r-legal σ; if σ contains an initial or terminal segment of some edge in H_r, then $L_r(\sigma) > 0$.*

Proof. We define a measure μ with support in the maximal invariant set $I_r = \{x \in H_r: f^k(x) \in H_r \text{ for all } k > 0\}$ as follows: Choose a positive vector $\vec{v}$ satisfying $\vec{v}M_r = \lambda_r \vec{v}$. If e_i is the i^{th} edge of H_r, define $\mu(e_i) = v_i$; if e is any edge in G_{r-1}, define $\mu(e) = 0$. Let $\mathscr{V}^m = \{x \in G_r: f^m(x) \in \mathscr{V}\}$. Subdividing at $\mathscr{V}^m$ decomposes G_r into intervals that map onto edges of G_r by f^m. For each such interval a define $\mu(a) = \mu(f^m(a))/\lambda_r^m$. If an edge $e \subset G_r$ is subdivided by $\mathscr{V}^m$ into $e = a_1 \cdot \cdots \cdot a_s$, then $\mu(e) = \sum_{i=1}^s \mu(a_i)$, because $\vec{v}M_r^m = \lambda_r^m \vec{v}$. This defines the desired measure μ.

Given a path $\sigma \subset G_r$, write $\sigma = \sigma_1 \cdot \cdots \cdot \sigma_k$ as a concatenation, where each σ_i, $(i \neq 1, k)$, is an edge and σ_1 and σ_k are contained in edges. Define $L_r(\sigma) = \sum_{i=1}^k \mu(\sigma_i)$. The desired properties of L_r are an immediate consequence of Lemma 5.8 and the definition of μ. □

For each path $\sigma \subset G_r$ write $\sigma = \sigma_1 \cdot \cdots \cdot \sigma_k$ as a concatenation, where each σ_i, $i \neq 1, k$, is an edge and σ_1 and σ_k are contained in edges. We define $\sigma \cap H_r \subset G_r$ to be the ordered subsequence of the σ_i's that are contained in H_r.

Let $f: G \to G$ be a relative train track map and let H_r be an exponentially growing stratum. If $f': G' \to G'$ is another relative train track map, we say that the *r stratum of $f: G \to G$ is projectively equivalent to the r' stratum of $f': G' \to G'$* if there is bijection h between the edges of H_r and $H'_{r'}$ such that $f'(h(E)) \cap H'_{r'} = h(f(E) \cap H_r)$ for each edge $E \subset H_r$.

Lemma 5.11. *Suppose that $f: G \to G$ is a relative train track map and H_r is an exponentially growing stratum. Then*

(1) *Each indivisible Nielsen path $\rho \subset G_r$ that intersects* $\text{int}(H_r)$ *contains exactly one illegal turn in H_r; thus $\rho = \alpha\beta$, where α and β are r-legal and $(\bar{\alpha}, \beta)$ is an illegal turn in H_r.*

(2) *There are only finitely many indivisible Nielsen paths $\rho \subset G_r$ that intersect* $\text{int}(H_r)$.

(3) *There is a finite set $R \subset (0, \infty)$ such that if the r' stratum of $f': G' \to G'$ is projectively equivalent to the r stratum of $f: G \to G$, and if $\rho' \subset G'_{r'}$ is an indivisible Nielsen path that intersects* $\text{int}(H'_{r'})$, *then $(L'_{r'}(\rho')/L'_{r'}(G')) \in R$.*

Proof. Let $\rho \subset G_r$ be an indivisible Nielsen path that intersects $\text{int}(H_r)$. Lemma 5.10 implies that ρ is not r-legal. Decompose $\rho = \alpha\beta \cdots$ into a concatenation of maximal r-legal subpaths.

There is a smallest initial segment a_m of α that satisfies $[f^m(a_m)] = \alpha$. Let b_m be the complementary subinterval of α, $\alpha = a_m b_m$. There is a largest initial subinterval c_m of β such that $[f^m(b_m c_m)]$ is the trivial path; let d_m be the complementary interval $\beta = c_m d_m$. Then $[f(a_{m+1})] = a_m$ and $[f(b_{m+1}c_{m+1})] = b_m c_m$. Thus $[f(\bigcap_{m=1}^{\infty} a_m)] = \bigcap_{m=1}^{\infty} a_m$ and $[f(\bigcup_{m=1}^{\infty} b_m c_m)] = \bigcup_{m=1}^{\infty} b_m c_m$. The irreducibility of ρ now implies that $\bigcap_{m=1}^{\infty} a_m$ is a point and $\rho = \alpha\beta$ is the closure of $\bigcup_{m=1}^{\infty} b_m c_m$. This establishes (1).

Fix $m > 0$. We will show that $a_m \cap H_r$, $b_m \cap H_r$, $c_m \cap H_r$ and $d_m \cap H_r$ are determined by $[f^m(\alpha)] \cap H_r$ and $[f^m(\beta)] \cap H_r$. Let $[f^m(\alpha)] \cap H_r = E_1 \cdot \cdots \cdot E_l$ and let $[f^m(\bar{\beta})] \cap H_r = E'_1 \cdot \cdots \cdot E'_m$. These sequences agree on some maximal terminal segment $E_{s+1} \cdot \cdots \cdot E_l = E'_{t+1} \cdot \cdots \cdot E'_m$. Since the terminal edges of $[f^m(a_m)] = \alpha$ and $[f^m(\bar{d}_m)] = \bar{\beta}$ are contained in H_r, these edges must be E_s and E'_t, respectively. Subdivide α along $\mathscr{V}^m = \{x \in G_r\colon f^m(x) \in \mathscr{V}\}$ to get $\alpha = u_1 \cdot \cdots \cdot u_r$, where each $f^m(u_l)$ is an edge in G. Then a_m is the initial segment of α terminating at the end of the edge u_j that maps onto E_s. This verifies our claim for a_m; the argument for b_m, c_m and d_m is similar.

Suppose that $\rho' \subset G_r$ is another indivisible Nielsen path that intersects $\operatorname{int}(H_r)$ and suppose that $\alpha \cap H_r$ and $\beta \cap H_r$ are contained in $\alpha' \cap H_r$ and $\beta' \cap H_r$, respectively. By Lemma 5.8 each $[f^m(\alpha)] \cap H_r$ is contained in $[f^m(\alpha')] \cap H_r$ and each $[f^m(\beta)] \cap H_r$ is contained in $[f^m(\beta')] \cap H_r$. This easily implies that each $b_m \cap H_r = b'_m \cap H_r$ and each $c_m \cap H_r = c'_m \cap H_r$. But $\rho = \bigcup_{m=i}^{\infty} b_m c_m$ and $\rho' = \bigcup_{m=i}^{\infty} b'_m c'_m$. We conclude that $\rho \cap H_r = \rho' \cap H_r$.

By Lemma 5.10, $L_r(\alpha) = \lambda_r L_r(\alpha) - L_r([f(b_1)])$ and $L_r(\beta) = \lambda_r L_r(\beta) - L_r([f(c_1)])$. Since $[f(b_1)] = [f(\bar{c}_1)]$, $L_r(\alpha) = L_r(\beta)$. The "bounded cancellation lemma" ([Co] or [T1]) and Lemma 5.10 imply that there is a uniform bound for $L_r(\alpha) = L_r(\beta)$, and hence a uniform bound for the length of the sequence of edges $\rho \cap H_r$. Combined with the preceding argument, this implies that $\rho \cap H_r$ takes on only finitely many values. In particular the first and last edge segments of ρ take on only finitely many values. Since these expand over α and β, ρ takes on only finitely many values. This completes the proof of (2).

These arguments only depend on $[f^m(\rho) \cap H_r]$ and therefore work equally well when the r' stratum of $f'\colon G' \to G'$ is projectively equivalent to the r stratum of $f\colon G \to G$ and when $\rho' \subset G'_r$ is an indivisible Nielsen path that intersects $\operatorname{int}(H'_{r'})$. This completes the proof of (3). □

For any topological representative $f\colon G \to G$ and exponentially growing stratum H_r let $N(f, r)$ be the number of indivisible Nielsen paths $\rho \subset G_r$ that intersect $\operatorname{int}(H_r)$. Let $N(f) = \sum N(f, r)$. Finally let $N_{\min}$ be the minimum value of $N(f)$ that occurs among topological representatives with $\Lambda = \Lambda_{\min}$. We say that a topological representative of $\mathcal{O}$ is *stable* if $\Lambda = \Lambda_{\min}$ and $N(f) = N_{\min}$. We will justify our choice of notation in Lemma 5.17.

THEOREM 5.12. *For every outer automorphism $\mathcal{O}$ of F_n there exists a stable relative train track map $f\colon G \to G$ representing $\mathcal{O}$.*

To prove Theorem 5.12 we will need two more operations.

Assume that a topological representative $f\colon G \to G$ acts linearly on each edge with respect to some metric on G. For each exponentially growing H_r let I_r be the maximal invariant subset $I_r = \{y \in H_r\colon f^l(y) \in H_r \text{ for all } l > 0\}$. If $f(H_r) \subset H_r$, then $I_r = H_r$. On the other hand, since M_r is irreducible and $\lambda_r > 1$, if $f(H_r) \not\subset H_r$, then I_r is a closed, infinite, nowhere dense set. (The reader can easily verify that I_r is a Cantor set.) For each edge $e_l \subset H_r$ we define the *core* of e_l to be the smallest closed subinterval of e_l that contains $I_r \cap \operatorname{int}(e_l)$. Since $f(\mathcal{V}) \subset \mathcal{V}$ and $f(I_r) = I_r$, it follows that f maps a sufficiently small initial or terminal subinterval of a core to an initial or terminal subinterval of a core. In particular the endpoints of cores comprise a finite set that is mapped into itself by f. It therefore makes sense to "subdivide" $f\colon G \to G$ by declaring the endpoints of cores to be vertices. We refer to this as a *core subdivision* of H_r.

LEMMA 5.13. *If $f'\colon G' \to G'$ is obtained from $f\colon G \to G$ by a core subdivision of H_r, then $\Lambda' = \Lambda$ and Df' maps the set of oriented edges in H'_r to itself. If $f\colon G \to G$ is stable, then $f'\colon G' \to G'$ is stable. If $j \neq r$ and H_j satisfies* (RTT-i) *or* (RTT-ii), *then H'_j satisfies* (RTT-i) *or* (RTT-ii), *respectively.*

Proof. During a core subdivision of an edge, H_r is divided into a (possibly trivial) initial segment, a nontrivial core and a (possibly trivial) terminal segment. Let S be the union of the nontrivial initial and terminal segments that are created, as is shown in Figure 12. We orient the segments σ_i of S so that the endpoint it shares with the core of its edge is its terminal endpoint. Since $f(\sigma_i)$ does not intersect the interior of any core, $f(\sigma_i)$ is either a path in G_{r-1} or a

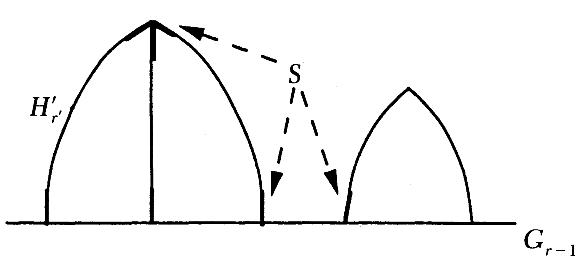

FIGURE 12

path in G_{r-1} followed by some σ_k. The filtration $\emptyset = G'_0 \subset \cdots \subset G'_{m'} = G'$ is obtained from $\emptyset = G_0 \subset \cdots \subset G_m = G$ in two steps. First a nonmaximal filtration is defined by inserting $G_{r-1} \cup S$ between G_{r-1} and G_r. Then the filtration is made maximal, as in our previous construction of a maximal filtration, by dividing the elements of S into equivalence classes and reordering the elements of S. The transition submatrix determined by the cores is exactly the same as M_r, and the transition submatrices for the strata determined by the elements of S are either zero or permutations. Thus $\Lambda' = \Lambda$.

As we noted above, initial or terminal subintervals of cores are sent to initial or terminal subintervals of cores; this immediately implies that Df' maps the set of oriented edges in H'_r to itself. This operation has no effect on H_j for $j \neq r$, so properties (RTT-i) and (RTT-ii) for H_j are preserved. Similarly this operation has no effect on $N(f)$ and so stability is preserved. $\square$

Our final operation is described in the statement and proof of the next lemma.

LEMMA 5.14. *Suppose that f: $G \to G$ is a stable topological representative and $\alpha \subset G_{r-1}$ is a path with endpoints in $H_r \cap G_{r-1}$ such that $[f(\alpha)]$ is trivial. Then there is a stable topological representative f': $G' \to G'$ such that $H'_r \cap G'_{r-1}$ has fewer points than $H_r \cap G_{r-1}$. If $j \geq r$ and H_j satisfies* (RTT-i), *then H'_j satisfies* (RTT-i). *If $j > r$ and H_j satisfies* (RTT-i) *and* (RTT-ii), *then H'_j satisfies* (RTT-i) *and* (RTT-ii).

Proof. Subdivide G at each point in $\alpha \cap f^{-1}(\mathscr{V})$ to obtain f_1: $G(1) \to G(1)$ and an identifying homeomorphism p: $G \to G(1)$. Define $h_1 = f \circ p^{-1}$: $G(1) \to G$ and $\alpha_1 = p \circ \alpha \subset G_1$. Note that $p \circ h_1 = f_1$. For each edge $a \subset \alpha_1$, $h_1(a)$ is an edge in G. Write α_1 as a sequence of edges $\alpha_1 = a_1 \cdots a_k$ and let $W_1 = h_1(a_1) \cdots h_1(a_s)$. Since $[W_1]$ is the trivial path, there exists $1 \leq l \leq s-1$ such that $h_1(a_{l+1}) = h_1(\bar{a}_l)$. This implies that $f_1(a_{l+1}) = f_1(\bar{a}_l)$, so we can fold $\bar{a}_l$ and a_{l+1}. Denote the resulting homotopy equivalence by f_2: $G(2) \to G(2)$ and the resulting quotient map by p_1: $G(1) \to G(2)$. There is a map h_2: $G(2) \to G$ such that $p_1 p_2 h_2 = f_2$. For each edge b in $[p_1(\alpha_1)] \subset G(2)$, $h_2(b)$ is a single edge in G. Decompose $\alpha_2 = b_1 \cdot \cdots \cdot b_l$ into a concatenation of edges and note that $W_2 = h_2(b_1) \cdot \cdots \cdot h_2(b_l)$ is obtained from W_1 by cancelling certain adjacent trivial pairs. In particular $l < s$ and $[W_2]$ is the trivial path. Repeat this procedure $(1 \leq i \leq k)$ to define f_i: $G(i) \to G(i)$ and p_{i-1}: $G(i-1) \to G(i)$ such that $\alpha_k = [p_{k-1} \cdots p_1 p(\alpha)]$ is the trivial path. Define f': $G' \to G'$ from f_k: $G(k) \to G(k)$ by tightening and collapsing the maximal pretrivial forest.

Corollary 5.6 implies that $\Lambda' = \Lambda = \Lambda_{\min}$. Since the moves described in Lemmas 5.1 and 5.3 do not change $N(f)$, $f'\colon G' \to G'$ is stable. The subdivision and folding are supported in G_{r-1}, and so no new points in $H_r \cap G_{r-1}$ are created. The previously distinct endpoints of α have now been identified; thus $G'_{r-1} \cap H'_r$ has fewer points than $H_r \cap G_{r-1}$.

Suppose that $j \geq r$ and H_j satisfies (RTT-i). As seen in our analysis of subdivision and folding, if e is an edge of H_j, then the only effect of our operation on $f(e)$ (thought of as a concatenation of edges) is to erase or change the names of some of the edges that are contained in G_{r-1}. This preserves condition (RTT-i).

Now suppose that $j > r$ and H_j satisfies (RTT-i) and (RTT-ii). It suffices to show that a single fold in G_{r-1} followed by tightening and the collapse of the maximal pretrivial forest preserves property (RTT-ii) for H_j. We say that a path α is *precollapsing* if $[f^l(\alpha)]$ is the trivial path for some $l \geq 1$. Since $[f([f^{l-1}(\alpha)])] = [f^l(\alpha)]$, property (RTT-ii) for H_j is equivalent to the statement that no path in G_{j-1} with endpoints in $H_j \cap G_{j-1}$ is precollapsing. Since folding does not create precollapsing paths, property (RTT-ii) for H_j is preserved by our construction. □

We will refer to the operation defined in the proof of Lemma 5.14 as *collapsing the inessential connecting path* α.

Proof of Theorem 5.12. Start with any stable topological representative of $\mathcal{O}$ and work down through the strata, performing core subdivisions and collapsing inessential connecting paths until each exponentially growing stratum satisfies (RTT-i) and (RTT-ii). By Lemma 5.9, (RTT-iii) is also satisfied. □

The second main result of this section is an analogue of Lemma 3.9.

THEOREM 5.15. *If $f\colon G \to G$ is a stable relative train track map and H_r is an exponentially growing stratum, then there is at most one indivisible Nielsen path ρ in G_r that intersects the interior of H_r. If ρ exists, then the illegal turn of ρ in H_r is the only illegal turn in H_r, ρ crosses every edge in H_r and $L_r(\rho) = 2\sum L_r(e)$, where the sum is taken over the edges of H_r.*

The following example has an indivisible Nielsen path that does not lie in single stratum.

Example 5.16. The following occurs as a pseudo-Anosov homeomorphism of a four-times punctured sphere. Three of the punctures determine loops in $G(1)$

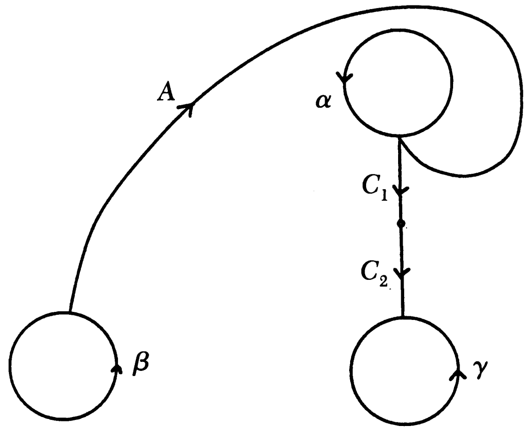

FIGURE 13

and the fourth determines an indivisible Nielsen path that intersects the interior of $H(2)$.

The filtration is $\emptyset = G(0) \subset G(1) \subset G(2) = G$, where $G(1)$ consists of three disjoint loops α, β and γ and where $H(2)$ consists of arcs A, connecting β to α, and C, connecting α to γ. The automorphism is given by $\alpha \mapsto \alpha$, $\beta \mapsto \beta$, $\gamma \mapsto \gamma$, $A \mapsto A\alpha C\gamma\bar{C}$ and $C \mapsto C\bar{\gamma}\bar{C}\bar{\alpha}\bar{A}\beta A\alpha C\gamma\bar{C}\alpha C$. Since Df respects Greek and Roman letters, it is easy to see that f is a train track map. The only illegal turn is $(C, \bar{A})$. The unique indivisible Nielsen path ρ that intersects the interior of $H(2)$ is a loop based at the fixed point x in C that corresponds to the occurrence of the letter C in the interior of the word $f(C)$. If we subdivide C at x into $C = C_1C_2$, then $A \mapsto A\alpha C_1C_2\gamma\bar{C}_2\bar{C}_1$, $C_1 \mapsto C_1C_2\bar{\gamma}\bar{C}_2\bar{C}_1\bar{\alpha}\bar{A}\beta A\alpha C_1$, $C_2 \mapsto C_2\gamma\bar{C}_2\bar{C}_1\alpha C_1C_2$ and $\rho = C_2\gamma\bar{C}_2\bar{C}_1\bar{A}\beta A\alpha C_1$, as can be seen in Figure 13. □

The following lemma is an analogue of Lemma 3.8.

LEMMA 5.17. *Suppose that $f\colon G \to G$ is a stable relative train track map, H_r is an exponentially growing stratum and $\rho \subset G_r$ is an indivisible Nielsen path that intersects* $\operatorname{int}(H_r)$. *Then the fold at the illegal turn of ρ in H_r is a full fold.*

Proof. The proof of Lemma 3.8 is entirely local and so carries over to this context with just one change. Instead of constructing $f'\colon G' \to G'$ from $f_3\colon G_3 \to G_3$, we simply stop at $f_3\colon G_3 \to G_3$. The edges α_2 and β_2 determine a new nonexponentially growing stratum. The indivisible Nielsen path ρ has moved from an exponentially growing stratum to a nonexponentially growing stratum in contradiction to the assumption that $N(f)$ is minimal. □

Proof of Theorem 5.15. We follow the proof of Lemma 3.9, which depends on iteratively folding the illegal turn in an indivisible Nielsen path. We must address the fact that folding does not always preserve (RTT-i) and (RTT-ii). For example, folding the indivisible Nielsen path in Example 5.16 does not preserve either (RTT-i) or (RTT-ii).

Suppose that $\rho \subset G_r$ is an indivisible Nielsen path that intersects $\operatorname{int}(H_r)$ and that α and β are as in Lemma 5.11. Let $e_1 \subset H_r$ and $e_2 \subset H_r$ be the initial edges of $\bar{\alpha}$ and β, respectively. We may assume that $L_r(e_1) \geq L_r(e_2)$. A new stable relative train track map $f'\colon G' \to G'$ is defined below; we say that $f'\colon G' \to G'$ is obtained from $f\colon G \to G$ by *folding* ρ.

As a first case, suppose that $L_r(e_1) = L_r(e_2)$. Perform the elementary fold of e_1 with e_2. If the resulting topological representative is a relative train track map, then this completes the construction of $f'\colon G' \to G'$. If it is not a relative train track map, then we restore the relative train track property by collapsing inessential connecting paths and by performing core subdivisions in G_{r-1}, as in the proof of Theorem 5.12. The result is a stable relative train track map $f'\colon G' \to G'$. In either case, H'_r has fewer edges than H_r.

Suppose now that $L_r(e_1) > L_r(e_2)$. Let $a \subset e_1$ be the initial segment that satisfies $f(a) = f(e_2)$. For the second case, suppose that the edge e_3, which follows e_2 in β, is contained in H_r. Subdivide e_1 at the terminal endpoint of a and perform an elementary fold of a with e_2. The resulting topological representative $f'\colon G' \to G'$ is stable. Let e'_1 be the edge other than a that is created when e_1 is subdivided. Then e'_1 and the edges of H_r, other than e_1, determine a stratum H'_r of G'. Since e_2 and e_3 are contained in H_r, $Df'(e'_1)$ and $Df'(\bar{e}'_1)$ are oriented edges in H'_r. Thus H'_r satisfies (RTT-i). The remaining relative train track conditions are easy to check.

For the third and final case, suppose that e_3 is contained in G_{r-1}. Let s be the maximal subpath of β that follows e_2 and is contained in G_{r-1}. There is an initial interval $b \subset e_1$ satisfying $f(b) = [f(e_2 s)]$. Subdivide e_1 at the terminal endpoint of b and fold all of b with all of $e_2 s$. The resulting topological representative $f'\colon G' \to G'$ is stable. The proof that $f'\colon G' \to G'$ is a relative train track map is exactly as in the previous case.

If the fold at the indivisible Nielsen path is of the first type, then the number of edges in the r^{th} stratum decreases. Since the number of edges in the r^{th} stratum never increases, folds of the first type occur only finitely many times. Ignoring these first folds, we may assume that folding ρ requires only a subdivision and a fold.

The proof of Lemma 3.9 now carries over to this context in a straightforward manner. The details are left to the reader. □

6. The Scott conjecture

In this section we complete the proof of the Scott conjecture.

THEOREM 6.1. *For any automorphism* $\Phi\colon F_n \to F_n$, $\operatorname{Rank}(\operatorname{Fix}(\Phi)) \leq n$.

We need only organize results that have already been proved. Most of this is done in Proposition 6.3. We begin with a definition and a straightforward lemma.

For any graph G let $\hat{G}$ be the subgraph consisting of all of the noncontractible components of G; define $\chi_-(G) = -\chi(\hat{G})$.

LEMMA 6.2. *If G is a subgraph of G', then $\chi_-(G) \leq \chi_-(G')$. If there is a path $\tau \subset G'$ that intersects $(G' \setminus G)$ and has endpoints in noncontractible components of G, then $\chi_-(G) < \chi_-(G')$.*

Proof. We may assume without loss that all of the components of G' are noncontractible. By removing valence-one vertices and the edges incident to them, we may assume that each vertex of G' has a valence of at least two. Since each vertex of $(G' \setminus G)$ is incident to at least two edges of $(G' \setminus G)$, and each edge of $(G' \setminus G)$ is incident to at most two vertices of $(G' \setminus G)$, then $\chi_-(G) \leq \chi_-(G')$. Moreover $\chi_-(G) = \chi_-(G')$ if and only if $(G' \setminus G)$ is a union of circles; this will not be the case if there exists τ as above. □

PROPOSITION 6.3. *Suppose that $f\colon G \to G$ is a stable relative train track map. Let $I(f) \subset \operatorname{Fix}(f)$ be the set of fixed points of f that are either vertices of G or isolated in* $\operatorname{Fix}(f)$. *Then there exists a graph Σ with a vertex set V_Σ and a map $p\colon \Sigma \to G$ such that:*

(1) *$p|V_\Sigma\colon V_\Sigma \to I(f)$ is a bijection.*

(2) *For each nontrivial path $\sigma \subset \Sigma$ with endpoints in V_Σ, $[p(\sigma)]$ is a nontrivial Nielsen path with endpoints in $I(f)$.*

(3) *For each Nielsen path $\rho \subset G$ with endpoints in $I(f)$, there is a path $\sigma \subset \Sigma$ with endpoints in V_Σ such that $\rho = [p(\sigma)]$. We say that σ is a* weak lift *of ρ.*

(4) $\chi - (\Sigma) \leq \chi_-(G)$.

Proof. Let $\emptyset = G_0 \subset G_1 \subset \cdots \subset G_m = G$ be the filtration associated to $f\colon G \to G$. To simplify our argument slightly, we assume that we have subdivided at the isolated fixed points of $\operatorname{Fix}(f)$ that lie in nonexponentially growing strata of the filtration. We argue by induction, verifying Proposition 6.3

for each $f_i = f|G_i$: $G_i \to G_i$. If $i = 0$, then $\Sigma = \varnothing$ satisfies the conclusions of the proposition. We therefore assume that p_{i-1}: $\Sigma_{i-1} \to G_{i-1}$ satisfies $(1)_{i-1}$–$(5)_{i-1}$ (i.e., satisfies (1)–(5) for f_{i-1}: $G_{i-1} \to G_{i-1}$) and construct p: $\Sigma \to G$, which satisfies $(1)_i$–$(5)_i$.

We will define Σ_i from Σ_{i-1} by adding a new vertex for each point in $I(f) \cap \mathrm{int}(H_i)$ and at most one edge. An edge is added if and only if there is a Nielsen path $\sigma_i \subset G_i$ that intersects $\mathrm{int}(H_i)$. (The exact choice of σ_i will be made explicit later.) The map p_{i-1}: $\Sigma_{i-1} \to G_{i-1}$ is extended over the new vertices in the obvious way and defined on the new edge ε_i, if it exists, as an immersion onto the Nielsen path σ_i.

Condition $(1)_i$ is immediate.

By Lemma 6.2, $\chi_-(G_i) \geq \chi_-(G_{i-1})$. Since at most one new edge is added to Σ_{i-1}, either $\chi_-(\Sigma_i) = \chi_-(\Sigma_{i-1})$ or $\chi_-(\Sigma_i) = \chi_-(\Sigma_{i-1}) + 1$. The latter occurs if and only if both endpoints of ε_i lie in noncontractible components of Σ_{i-1}. In this case, property $(2)_{i-1}$ implies that the endpoints of σ_i lie in noncontractible components of G_{i-1}; hence Lemma 6.2 implies that $\chi_-(G_i) \geq \chi_-(G_{i-1}) + 1$. Condition $(4)_i$ follows from $(4)_{i-1}$.

It remains to consider properties $(2)_i$ and $(3)_i$. If every Nielsen path $\rho \subset G_i$ satisfies $\rho \subset G_{i-1}$, then $(2)_i$ and $(3)_i$ follow immediately from $(2)_{i-1}$ and $(3)_{i-1}$. We may therefore assume that there is at least one Nielsen path in G_i that intersects $\mathrm{int}(H_i)$; in particular we may assume that M_i is irreducible.

Each Nielsen path $\rho \subset G_i$ has a unique decomposition $\rho = \rho_1 \cdot \cdots \cdot \rho_r$ into Nielsen paths ρ_i with endpoints in $I(f)$ such that each ρ_j is either indivisible or entirely contained in $\mathrm{Fix}(f)$. To see this let Γ_i be the universal cover of the component of G_i that contains ρ, let $\tilde{x} \in \tilde{\Gamma}_i$ be a lift of the initial point x of ρ and let $\tilde{f}_i$: $\Gamma_i \to \Gamma_i$ be the lift of f_i: $G_i \to G_i$ that fixes $\tilde{x}$. There is a unique lift $\tilde{\rho}$ of ρ originating at $\tilde{x}$. Its intersection with $\mathrm{Fix}(\tilde{f}_i)$ decomposes $\tilde{\rho}$ into a concatenation $\tilde{\rho} = \tilde{\rho}_1 \cdot \cdots \cdot \tilde{\rho}_l$, and the projected images of the $\tilde{\rho}_j$'s are the desired ρ_j's.

We consider the exponentially growing and nonexponentially growing cases separately. Suppose at first that $\lambda_i > 1$.

By Theorem 5.15 there is at most one indivisible Nielsen path $\sigma_i \subset G_i$ that intersects $\mathrm{int}(H_i)$; no edges of H_i are entirely contained in $\mathrm{Fix}(f)$. Thus (up to a change of orientation) if ρ_j intersects $\mathrm{int}(H_i)$, then $\rho_j = \sigma_i$ and we can lift ρ_j to ε_i. If $\rho_j \subset G_{i-1}$, then $(3)_{i-1}$ guarantees a weak lift of ρ_j. Thus each ρ_j, and hence ρ, can be weakly lifted. This establishes $(3)_i$.

Suppose that $\sigma \subset \Sigma_i$ has endpoints in V_{Σ_i}. By $(2)_{i-1}$, $[p(\sigma)]$ can be written as a concatenation of nontrivial paths $a_k \subset G_{i-1}$ and $b_k = \rho_i$ or $b_k = \bar{\rho}_i$. Since the initial and terminal edges of ρ_i are contained in H_i, there is no cancellation at the junctures of the a_k's and b_k's, and we conclude that $[p(\sigma)]$ is nontrivial. This proves $(2)_i$.

The case where H_i is a single edge that is entirely contained in $\mathrm{Fix}(f)$ is proved in exactly the same way and is left to the reader.

We assume now that $\lambda_i = 1$ and H_i is not a single edge contained in $\mathrm{Fix}(f)$. By assumption there is an indivisible Nielsen path $\rho_j \subset G_i$ that intersects $\mathrm{int}(H_i)$. Since $f(\rho_j)$ has the same number of edges in H_i that ρ_j does, and since $[f(\rho_j)] = \rho_j$, we see that $f(\rho_j)$ and ρ_j cross the same edges of H_i in the same order and that no edges of H_i are cancelled when $f(\rho_j)$ is tightened to ρ_j. It follows that the transition matrix M_i is the identity. By the irreducibility of M_i, H_i consists of a single edge E_i. Replacing E_i by $\overline{E}_i$ if necessary, we find there is a nontrivial path $v_i \subset G_{i-1}$ such that $f(E_i) = E_i v_i$. (This is where our subdivision of isolated fixed points in nonexponentially growing stratum is used.) Since no edges of H_i are cancelled when $f(\rho_j)$ is tightened to ρ_j, the indivisibility of ρ_j implies that E_i may occur as the initial edge in ρ_j and $\overline{E}_i$ may occur as the terminal endpoint of ρ_j, but no other occurrence of E_i or $\overline{E}_i$ is possible. Thus (up to a reversal of orientation) there is a subpath $\beta \subset G_{i-1}$ such that either $\rho_j = E_i\beta$ or $\rho_j = E_i\beta\overline{E}_i$.

There are two subcases to consider. If there is a path $\alpha \subset G_{i-1}$ that originates at the terminal vertex of E_i and satisfies $[f(\alpha)] = [\bar{v}_i\alpha]$, then define $\sigma_i = E_i\alpha$. If $\rho_j = E_i\beta$, then $[f(\beta)] = [\bar{v}_i\beta]$, and so ρ_j can be written as $\rho_j = [\sigma_i\bar{\alpha}\beta]$. If the Nielsen path $[\bar{\alpha}\beta]$ is nontrivial, then it can be weakly lifted by induction; σ_i weakly lifts to ε_i. It follows that ρ_j can be weakly lifted. Similarly if $\rho_j = E_i\beta\overline{E}_i$, then $[f(\beta)] = [\bar{v}_i\beta v_i]$ and $\rho_j = [\sigma_i\bar{\alpha}\beta\alpha\bar{\sigma}_i]$ can be weakly lifted.

If there is no path $\alpha \subset G_{i-1}$, as above, then each ρ_j is of the form $E_i\beta\overline{E}_i$. The path β is a loop based at the terminal endpoint y of E_i and satisfies $[f(\beta)] = [\bar{v}_i\beta v_i]$. In other words β is an element of the fixed subgroup of the automorphism of $\pi_1(G, y)$ determined by f and the path v_i. By Remark 2.3 the subgroup generated by all such β has rank one. Choose β_1 so that it generates this rank-one subgroup and let $\sigma_i = E_i\beta_1\overline{E}_i$. Then ρ_j is a multiple of σ_i and can be weakly lifted to a multiple of ε_i. This proves $(3)_i$.

To verify $(2)_i$ note that $p(\sigma)$ can be written as a concatenation of nontrivial $a_k \subset G_{i-1}$ and $b_k = \sigma_i$ or $b_k = \bar{\sigma}_i$. To see that $[p(\sigma)]$ is nontrivial we will show that no E_i nor $\overline{E}_i$ that occur in a b_k are cancelled when $p(\sigma)$ is tightened to $[p(\sigma)]$. Suppose that $\sigma_i = E_i\alpha$. If $p(\sigma)$ has a subword of the form $E_i w\overline{E}_i$, where $w \subset G_{i-1}$, then $w = \alpha a_k\bar{\alpha}$ is not trivial; if $p(\sigma)$ has a subword of the form $\overline{E}_i w' E_i$, where $w' \subset G_{i-1}$, then $w' = a_k$ is not trivial. This verifies our claim in this subcase. Suppose that $\sigma_i = E_i\beta_1\overline{E}_i$. If $p(\sigma)$ has a subword of the form

$E_1 w \bar{E}_i$, where $w \subset G_{i-1}$, then $w = \beta_1$ or $w = \bar{\beta}_1$ is nontrivial; if $p(\sigma)$ has a subword of the form $\bar{E}_1 w' E_i$, where $w' \subset G_{i-1}$, then $w' = a_k$ is not trivial. This verifies our claim in the remaining subcase and thereby proves $(2)_i$. □

We say that two elements of $\mathrm{Fix}(f)$ are *Nielsen equivalent* if they are connected by a Nielsen path.

COROLLARY 6.4. *Suppose that $f\colon G \to G$ is a stable relative train track map, that $N_1, \dots, N_k$ are the Nielsen classes in* $\mathrm{Fix}(f)$ *and that $x_i \in N_i$. Let $(f_i)_\#\colon \pi_1(G, x_i) \to \pi_1(G, x_i)$ be the automorphism determined by f and the constant path at x_i. Then*

$$\sum_{i=1}^{k} \max\big(0, \big(\mathrm{Rank}(\mathrm{Fix}(f_i)_\#) - 1\big)\big) = \chi - (\Sigma) \leq \chi - (G) = n - 1.$$

Proof. Properties (1)–(3) imply that $v_1, v_2 \in V_\Sigma$ belong to the same component of Σ if and only if $p(v_1)$ and $p(v_2)$ are Nielsen equivalent. Let C_i be the component of Σ that contains the vertex v_i that satisfies $p(v_i) = x_i$.

Properties (1)–(3) imply that $p|C_i\colon C_i \to G$ induces an isomorphism between $\pi_1(C_i, v_i)$ and $\mathrm{Fix}((f_i)_\#)$. The corollary now follows from property (4) and the fact that a graph, whose fundamental group has rank n, has Euler characteristic $n - 1$. □

Proof of Theorem 6.1. Corollary 2.2 states that either $\mathrm{Rank}(\mathrm{Fix}(\Phi)) \leq 1$ or Φ is conjugate to one of the automorphisms $(f_i)_\#\colon \pi_1(G, x_i) \to \pi_1(G, x_i)$, referred to in the statement of Corollary 6.4. This corollary completes the proof. □

UCLA, LOS ANGELES, CALIFORNIA
LEHMAN COLLEGE, CUNY, BRONX, NEW YORK

REFERENCES

[BH1] M. BESTVINA and M. HANDEL, Train tracks for surface automorphisms, preprint.

[BH2] ______, Dynamics of the outer automorphism group and a partial Tits alternative, in preparation.

[BC] S. BLEILER and A. CASSON, *Automorphisms of Surfaces after Nielsen and Thurston*, Cambridge University Press, Cambridge, 1988.

[CL] M. M. COHEN and M. LUSTIG, On the dynamics and the fixed subgroup of a free group automorphism, Invent. Math. **96** (1989), 613–638.

[Co] D. COOPER, Automorphisms of free groups have finitely generated fixed point sets, J. Algebra **111** (1987), 453–456.

[CV] M. CULLER and K. VOGTMANN, Moduli of graphs and automorphisms of free groups, Invent. Math. **84** (1986), 91–119.

[DS] J. L. DYER and G. P. SCOTT, Periodic automorphisms of free groups, Comm. Alg. **3** (1975), 195–201.

[FLP] A. Fathi, F. Laudenbach and V. Poenaru, *Travaux de Thurston sur les Surfaces*, Astérisque **66–67**, 1979.

[Fr] D. Fried, Growth rates of surface homeomorphisms and flow equivalence, Erg. Th. and Dynam. Syst. **5** (1985), 535–563.

[Ge1] S. M. Gersten, Fixed points of automorphisms of free groups, Inv. Math. **84** (1986), 91–119.

[Ge2] ______, Dynamics of primitive positive automorphisms of a free group, preprint.

[Ge3] ______, Geometric automorphisms of a free group of rank at least three are rare, Proc. A. M. S. (1983), 27–31.

[GS] S. M. Gersten and J. Stallings, A note on irreducible outer automorphisms of a free group, preprint.

[GT1] R. Z. Goldstein and E. C. Turner, Automorphisms of free groups and their fixed points, Invent. Math. **78** (1984), 1–12.

[GT2] ______, Fixed subgroups of homomorphisms of free groups, Bull. London Math. Soc. **18** (1986), 468–470.

[HT] M. Handel and W. Thurston, New proofs of some results of Nielsen, Adv. in Math. **56** (1985), 173–191.

[Mi] R. T. Miller, Nielsen's viewpoint on geodesic laminations, Adv. in Math. **45** (1982), 189–212.

[ScW] G. P. Scott and C. T. C. Wall, Topological methods in group theory, in *Homological Methods in Group Theory*, C. T. C. Wall, ed., London Math. Soc. Lect. Notes **936**, Cambridge University Press, Cambridge, 1979.

[Se] E. Seneta, *Non-negative Matrices and Markov Chains*, Springer-Verlag, New York, 1981.

[St1] J. R. Stallings, Topology of finite graphs, Inv. Math. **71** (1983), 551–565.

[St2] ______, Topologically unrealizable automorphisms of free groups, Proc. A.M.S. **84** (1982), 21–24.

[St3] ______, Graphical theory of automorphisms of free groups, Ann. of Math. Studies **111** (1987), 79–106.

[T1] W. P. Thurston, Course notes, Princeton University, 1983.

[T2] ______, On the geometry and dynamics of diffeomorphisms of surfaces, Bull. A.M.S. **19** (1988), 417–431.

(Received May 22, 1989)
(Revised September, 1991)