Invent. math. 90, 189–215 (1987)

Inventiones mathematicae

Mutation and volumes of knots in S^3

Daniel Ruberman
Brandeis University, Department of Mathematics, Waltham, MA 02254, USA

Knots in S^3 can be decomposed into simpler pieces in several different ways. The most basic is by connected sum into prime pieces; such a decomposition is unique. Further, it is easy to understand the contribution of each summand to any knot invariant one might wish to compute. Another splitting of knots occurs when a 2-sphere (called a Conway sphere) in S^3 hits the knot transversally in four points. The resulting splitting into so-called tangles has proved quite fruitful in various investigations [7, 8, 22, 6] of symmetries and other properties of knots. Given a Conway sphere, there is an operation called mutation which yields a new knot. Roughly speaking, one takes out the tangle, flips it over and glues it back in. The resulting knot tends to differ from the original one, unless the tangle on one side was symmetric [6]. However, many invariants of a knot are preserved by mutation [7], e.g. the signature and Alexander polynomial, as well as the new two-variable knot polynomial [12].

In this paper we show that the Gromov norm (see below) of a knot and its mutant coincide. In particular if S^3-K is a hyperbolic manifold then S^3-mutant of K is as well, and their volumes are the same. These results are instances of a more general theorem which shows that Gromov's norm is preserved by certain kinds of cutting and pasting along surfaces. By using the torus decomposition of a general 3-manifold, one reduces the problem to understanding what happens for hyperbolic manifolds. For the case of a hyperbolic manifold we show that for certain surfaces $F \subset M$ and symmetries τ of F, cutting M along F and regluing via τ results in a new hyperbolic manifold M' with $\mathrm{vol}(M)=\mathrm{vol}(M')$. Colin Adams [1] has proved a similar result for the special case of a thrice-punctured sphere in M. Our basic method can be extended to show that for Dehn surgeries on a knot which produce hyperbolic manifolds, the volume is the same for the corresponding surgeries on the mutant knot. We also give a similar result about the volumes of the branched cyclic covers of the knot and its mutant. In a future paper, we will consider the effect of this sort of cutting and pasting on the Chern-Simons invariant and η-invariant of hyperbolic manifolds.

In the course of the proof of the theorem concerning hyperbolic manifolds, we need to use the results of [11] on embeddedness and intersections of least area surfaces. The theorems of [11] deal only with the compact case, and we need their analogues in the case of finite-volume hyperbolic 3-manifolds. These results are established in section 3 by methods similar to those of [11] with proper attention paid to the behavior of least area surfaces in the cusps of hyperbolic 3-manifolds.

We thank Joel Hass and Peter Scott for help with the minimal surfaces, and Bill Thurston and John Morgan for encouraging conversations.

1. Definitions and statements of theorems

Definition 1.1. *Let $K \subset S^3$ be a knot or link. A Conway sphere for K is an embedded 2-sphere meeting K transversally in four points.*

The surface $S^2 - 4$ *points* admits several orientation preserving involutions as drawn in Fig. 1.

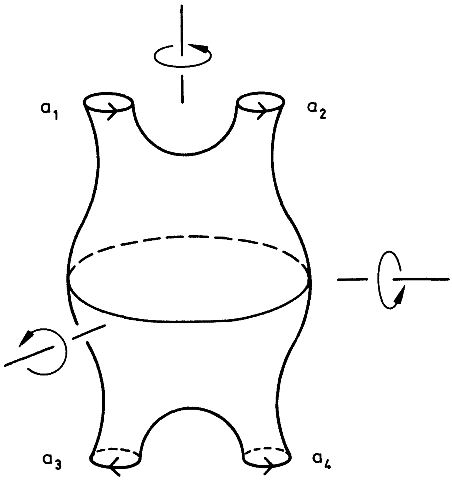

Fig. 1.

These evidently extend over S^2. If S is a Conway sphere, write $(S^3, K) = (B^3_+, K_+) \cup (B^3_-, K_-)$ where $K_\pm = B^3_\pm \cap K$.

Definition 1.2. *Let μ be any of the involutions in Fig.* 1. *Then the mutation of K via μ is $(S^3, K^\mu) = (B^3_+, K_+) \cup_\mu (B^3_-, K_-)$.*

The property of a four-punctured sphere and the involution μ that is relevant is that μ is an isometry of any hyperbolic structure on S^2-four points. There are a few other surface symmetries with this property, which are pictured in Fig. 2 below.

Convention. For the rest of this paper, (F, τ) will denote one of the following surfaces with the indicated involution, or S^2-four points with one of the involutions from Fig. 1.

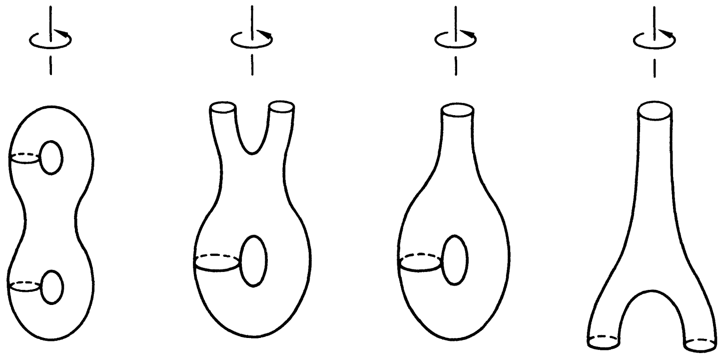

Fig. 2. Symmetric surfaces

For convenience, we will refer to a pair (F, τ) as a *symmetric surface*. Suppose (F, τ) is such a symmetric surface, and that $(F, \partial F)$ is embedded in $(M^3, \partial M)$. Cutting M along F and regluing via τ will be denoted M^τ.

Theorem 1.3. *Let $(F, \partial F) \subset (M, \partial M)$ be incompressible and ∂-incompressible. If $\operatorname{int}(M)$ has a hyperbolic structure of finite volume, then so does M^τ. Furthermore,* $\operatorname{vol}(M^\tau) = \operatorname{vol}(M)$.

Since the complement of the mutant of a knot K is obtained by cutting the complement of K along the Conway sphere, and regluing via the symmetry μ, we obtain:

Corollary 1.4. *Let μ be any mutation of the hyperbolic knot K. Then K^μ is also hyperbolic, and the volumes of their complements are the same.*

M. Gromov [13, 30] has defined an invariant of 3-manifolds with boundary a union of tori, denoted $|M|$, which reduces in some sense to the volume of hyperbolic manifolds. In fact, if M is hyperbolic, then $|M| = \frac{\operatorname{vol}(M)}{v_3}$ where v_3 is the volume of an ideal regular tetrahedron in $\mathbf{H}^3$. In general, $|\cdot|$ adds for connected sums, and if M is decomposed along tori into hyperbolic and Seifert-fibered pieces M_i, then $|M| = \sum_{M_i \text{ hyperbolic}} |M_i|$. Using these facts, we show:

Theorem 1.5. *If $(F, \partial) \subset (M, \partial)$ is incompressible and ∂-incompressible, then* $|M^\tau| = |M|$.

The technique used to prove theorem 1.3. stems from the observation that having a hyperbolic metric is, after all, a local property. Hence we look for a surface F in M with the property that in the induced metric on F, τ can be taken to be an isometry. Once given such a surface (and some additional information on the behavior of τ in a neighborhood of F), M^τ evidently has a metric of constant curvature -1 made of the old metric on either side of F. One also has to argue that this metric is complete; this follows by understanding what happens in the cusps. The surfaces F we consider all have the property that for any hyperbolic metric on F, τ is an isometry. Hence if F could be taken to be totally geodesic, theorem 1.4 would have a more straightforward proof. This happens, for instance if F is a three-punctured sphere [1], and thus we recover Adams' result. However, for more complicated surfaces, F is unlikely to be totally geodesic, so a different approach is needed.

Instead of cutting and pasting along a totally geodesic surface, we will use a least area surface. We demonstrate existence, embeddedness, and uniqueness for such surfaces, analogous to the theorems in [11]; such theorems were not previously known for the case of non-compact hyperbolic manifolds. Let M be a complete hyperbolic manifold, and let $f: F \to M$ be a proper incompressible and ∂-incompressible embedding of a surface, taking parabolics to parabolics. Consider $\mathbf{F} = \{g: F \to M \mid g \simeq_p f\}$ ($\simeq_p$ means properly homotopic to); if F double covers some embedded surface K, let $\mathbf{K}$ be the set of surfaces properly homotopic to the embedding of K.

Theorem 1.6. $\operatorname{Inf}\{area(F) \mid F \in \mathbf{F}\}$ *is realized by a least area map. Further a least area map is embedded, or double covers a least area element of* $\mathbf{K}$; *any two least area elements of* $\mathbf{F}$ *or* $\mathbf{K}$ *coincide or are disjoint.*

This will be proved in section 3; the proof follows the outline of [11], but some of the arguments are complicated by the non-compact nature of the problem. In particular, the known existence theorems for least area maps do not apply, as the basic results of Morrey [23] require that the injectivity radius be bounded below. These difficulties are avoided by using the topological approach of [15] plus the explicit understanding one has of the geometry of a cusp of a hyperbolic 3-manifold.

2. Symmetric representation spaces

The first step in proving theorem 1.3 is to strengthen the sense in which τ is a symmetry of F. If (F, τ) is a symmetric surface, then, as mentioned above, τ is an isometry of any hyperbolic metric on F. That fact can be interpreted in terms of representations of $\pi_1(F)$ in $\mathrm{PSL}_2(\mathbf{R})$; we will show that the same holds in the larger group $\mathrm{PSL}_2(\mathbf{C})$, i.e. that discrete and faithful representations of $\pi_1(F)$ in $\mathrm{PSL}_2(\mathbf{C})$ are symmetric (up to conjugacy) with respect to τ. In showing this, it is technically easier to work in $\mathrm{SL}_2(\mathbf{C})$. Any discrete subgroup of $\mathrm{PSL}_2(\mathbf{C})$ with no 2-torsion lifts to $\mathrm{SL}_2(\mathbf{C})$ ([9, 19]) so this is no loss of generality. We need to see that a given representation lifts to $\mathrm{SL}_2(\mathbf{C})$ in a way that is nice when restricted to the boundary of F. Recall that a parabolic element of $\mathrm{PSL}_2(\mathbf{C})$ is one that fixes a single point on the sphere at infinity; such an element if lifted to $\mathrm{SL}_2(\mathbf{C})$ will have trace ± 2.

Proposition 2.1. *Suppose* φ *is a discrete, faithful representation of* $\pi_1(F)$ *in* $\mathrm{PSL}_2(\mathbf{C})$ *such that* φ *takes boundary elements of* $\pi_1(F)$ *to parabolics. Let* $\hat{\varphi}$ *be a lifting of* $\varphi \mid \partial F$ to $\mathrm{SL}_2(\mathbf{C})$. *Then* $\hat{\varphi}$ *extends to all of* $\pi_1(F)$ *if and only if the number of boundary curves with* $trace(\hat{\varphi}) = +2$ *is even.*

Proof. Let $\tilde{F}$ denote the universal cover of F, and let $M_F = \mathbf{H}^3/\varphi(\pi_1(F))$. A standard 3-manifolds argument shows that there is an embedded copy of F

in M_F, with the ends of F embedded in a standard fashion in the cusps of M_F. The representation φ defines a (flat) $\mathbf{H}^3$ bundle $E_\varphi = \tilde{F} \times_\varphi \mathbf{H}^3$; note that topologically, this bundle is just the tangent bundle of F plus its normal bundle in M_F. Let c be any curve of F homotopic to an end of F; then E_φ restricted to c can be trivialized in the following manner: The bundle over c is by definition $\mathbf{R} \times \mathbf{H}^3$ divided by the action $(t, x) \to (t+1, \varphi_0(x))$, where φ_0 is the parabolic element $\varphi(c)$. Choosing a path of parabolics (all fixing the same point at infinity) back to the identity provides a trivialization of $E_\varphi | c$, and hence a frame field near c. One can check that this frame field has the property that one vector is tangent to c, one is normal to c (in F) and the third is normal to F (in M_F); in fact this same procedure provides such a frame field on the whole end of F cut off by c.

A framing near the ends of F provides a spin structure near the ends and hence a lift of the representation on the ends into $\mathrm{SL}_2(\mathbf{C})$. A direct geometric construction shows that this spin structure (and hence the lift) extends over F precisely when the number of boundary components of F is even. If the number of boundary components is odd, then change the framing on one of them by rotation of 2π (tangent to F) as you go around c; the resulting framing will now extend to a spin structure on E_φ.

Note that for the framing we have defined on the ends of F, the corresponding lift into $\mathrm{SL}_2(\mathbf{C})$ has trace $+2$ where it is untwisted, and trace -2 where we rotated as in the previous paragraph. One way to see this is to restrict attention to one cusp; then the whole set-up can be deformed so that φ_0 is real, i.e. is in $\mathrm{PSL}_2(\mathbf{R})$. Then the framing defines a lift of φ_0 into $\widetilde{\mathrm{PSL}_2(\mathbf{R})}$, the universal cover of $\mathrm{PSL}_2(\mathbf{R})$. Using the identification of $\mathrm{PSL}_2(\mathbf{R})$ with the unit tangent bundle of $\mathbf{H}^2$, one can see that the lift which corresponds to a framing which is tangent to the curve c will project to an element of $\mathrm{SL}_2(\mathbf{R})$ of trace $+2$, and that the rotated framing will project to an element of trace -2.

Finally, recall that there is an action of $H^1(F; \mathbf{Z}_2)$ on lifts of a given $\mathrm{PSL}_2(\mathbf{C})$-representation given by regarding the cohomology group as homomorphisms into $\mathbf{Z}_2$. Using this action, one can change the sign of the traces of any pair of boundary curves, and hence the lift we have constructed can be changed to extend any lift on the boundary with an even number of elements of trace $+2$. Conversely, if there are an odd number of such curves, then the representation cannot be lifted, because the corresponding spin structure on the boundary will not extend over F.

Remark. I thank Bill Goldman for help with this argument; Marc Culler has shown the same result for the case that φ is quasi-Fuchsian (or a limit of such representations.) By hard results of Thurston (see below), this suffices for representations arising from F being an incompressible surface in a hyperbolic 3-manifold.

Theorem 2.2. *Let (F, τ) be a symmetric surface as defined above, and let $\rho: \pi_1(F) \to \mathrm{PSL}_2(\mathbf{C})$ be a discrete and faithful representation taking the cusps of F to parabolics. Then $\rho \circ \tau_*$ is conjugate (in $\mathrm{PSL}_2(\mathbf{C})$) to ρ.*

Proof. By the preceding proposition, ρ lifts to a representation $\hat{\rho}$ into $\mathrm{SL}_2(\mathbf{C})$, with the property that there are an even number of cusps of F with trace 2.

By using the action of $H^1(F; \mathbf{Z}_2)$ as in the preceding proof, we may assume that all pairs of cusps which are interchanged by τ are represented by elements of $SL_2(\mathbf{C})$ which have trace equal to $+2$. Since $\hat{\rho}$ is discrete and faithful, it is an irreducible representation; the same is evidently true for $\hat{\rho}\circ\tau_*$.

By [10, 1.5.2], to show that two irreducible representations are conjugate in $SL_2(\mathbf{C})$, we need only show that their characters are the same, i.e. that

$$tr(\hat{\rho}(\alpha)) = tr(\hat{\rho}\circ\tau_*(\alpha))\,\forall\alpha\in\pi_1(F) \tag{1}$$

Let $g_1, \ldots, g_n$ generate $\pi_1(F)$; according to the proof of Proposition 1.4.1 of [10] we must verify that the equation (1) holds for elements of the form $\alpha = g_{i_1}, \ldots, g_{i_r}$ where the i_k are distinct positive integers $\leqq n$ and $i_1 < i_2 <, \ldots, i_r$. The verification is the essentially the same in all cases, and so we give it for $F = S^2 - 4$ points.

Write ρ' for $\hat{\rho}\circ\tau_*$. Let a_1, a_2, a_3, a_4 be generators of $\pi_1(F)$ going around the four punctures and oriented as indicated in Fig. 1, so that $a_1 a_2 a_3 = a_4$. The assumption on $\hat{\rho}|_{\partial F}$ means $\chi_{\hat{\rho}} = \chi_{\rho'}$ for the fundamental group elements: $a_1, a_2, a_3, \alpha_1 a_2 a_3$. Further, since trace is an invariant of conjugacy classes, we don't have to watch base points. Recall that the trace equality in $SL_2(\mathbf{C})$ implies that $tr(A) = tr(A^{-1})$, and note that τ sends each of the a_i to some $a_j^{\pm 1}$, so that $\rho'(a_i)$ has the same trace as $\hat{\rho}(a_i)$. By explicitly drawing curves, one sees easily that $\tau(a_1 a_2) = a_1 a_2$ and $\tau(a_2 a_3) = a_2 a_3$ up to conjugacy, so that the trace is preserved. Likewise $\tau(a_1 a_3) = (a_1 a_3)^{-1}$ so that its trace is preserved. Finally, since $a_1 a_2 a_3 = a_4$, $tr(\rho'(a_1 a_2 a_3)) = \pm 2 = tr(\hat{\rho}(a_1 a_2 a_3))$. Hence $\hat{\rho}$ and ρ' are conjugate in $SL_2(\mathbf{C})$; projecting down into $PSL_2(\mathbf{C})$ gives a conjugacy between ρ and $\rho\circ\tau_*$.

If F is an incompressible surface in M, it determines a covering space of M which will be denoted M_F, and which is of course homotopy equivalent to F. If M is hyperbolic, then M_F is hyperbolic as well, and the hyperbolic structure on M_F is determined by the restriction of the representation ρ of $\pi_1(M)$ in $PSL_2(\mathbf{C})$ to $\pi_1(F)$.

Lemma 2.3. *Suppose $\Gamma \subset PSL_2(\mathbf{C})$ is a discrete, non-elementary group, and that $A \in PSL_2(\mathbf{C})$ commutes with every element of Γ. Then A is the identity.*

Proof. Using the classification of isometries of $\mathbf{H}^3$, it is easy to show that two elements of $PSL_2(\mathbf{C})$ commute if and only if they have the same fixed point set on the sphere at infinity. So if A commutes with every element of Γ, the fixed point set of Γ is one or two points, so that Γ is elementary.

Corollary 2.4. *Let F be a hyperbolic surface (i.e. $\chi(F) < 0$). If two discrete faithful representations of the group $\pi_1(F)$ are conjugate, then there is a unique $A \in PSL_2(\mathbf{C})$ which conjugates one to the other.*

Proof. Since F is a hyperbolic surface, $\rho(\pi_1(F))$ is non-elementary. If A and B both conjugate ρ to ρ', then for all $g \in \pi_1(F)$, $A\rho(g)A^{-1} = B\rho(g)B^{-1}$. Hence $B^{-1}A$ commutes with all of $\rho(\pi_1(F))$, so by the above lemma must be the identity.

Lemma 2.5. *Suppose φ is an automorphism of $\pi_1(F)$ such that $\rho\circ\varphi$ is conjugate to ρ in $\mathrm{PSL}_2(\mathbf{C})$. Then there is an isometry φ of M_F whose induced map on π_1 is φ. If φ is an involution, then φ is one as well.*

Proof. Let $A\in\mathrm{PSL}_2(\mathbf{C})$ be the matrix which conjugates $\rho\circ\varphi$ to ρ. Since A takes the subgroup $\rho(\pi_1(F))$ to itself, there is a unique isometry of M_F covered by A. If $\varphi^2 = identity$, then A^2 will conjugate $\rho\circ\varphi^2$ to ρ, hence A^2 will be the identity. Putting together lemmas 3, 4 and 5 with theorem 2.2, we obtain the main result of this section.

Theorem 2.6. *Let F be a properly embedded incompressible surface in the hyperbolic 3-manifold M, and M_F the covering of M corresponding to the subgroup $\pi_1(F)$. If τ is a symmetry of F (in the sense of this paper), then there is an isometry $\hat{\tau}$ of M_F which is an involution and which induces the isomorphism τ_* on $\pi_1(F)$.*

Remark 2.7. There is a less elementary proof of theorem 2.6, which relies on deep results of Ahlfors, Bers, and Thurston on the deformations of hyperbolic structures on geometrically finite manifolds. For the reader's convenience, we sketch the argument. The first point is that $\pi_1(F)$, as a subgroup of $\mathrm{PSL}_2(\mathbf{C})$, is either quasi-Fuchsian, or a limit of quasi-Fuchsian groups; the second case happens exactly when F is (virtually) a fiber in a fibration of M over the circle. That these are the only alternatives is a consequence of [30, chapter 9], and [5]. If $\pi_1(F)$ is quasi-Fuchsian, then, according to the theory of Ahlfors and Bers [4], it is determined by hyperbolic structures on copies of F 'at infinity in M_F'. An elementary computation of the Teichmuller space of any of the symmetric surfaces F shows that τ is an isometry of any hyperbolic metric on F—this is of course a consequence of 2.2, but can be done directly. From the isometries "at infinity", we get the desired isometry of M_F. In the case where F is a fiber, one has to argue further that the isometry persists in the limit.

The above theorem will be used in showing that cutting and pasting via τ preserves volume; in order to show that Dehn surgery on mutant knots produces manifolds of the same volumes, we need a slight generalization. Suppose that F is one of our symmetric surfaces, and that some of the cusps of F are replaced by points labeled with integers, to get an 'orbifold' F in the sense of [30]; the local structure at such a point is that of $\mathbf{H}^2$ modulo an elliptic element. We will refer to the order of the elliptic element as the order of the point. If elliptic points which are interchanged by τ have the same order, then τ induces a symmetry of the orbifold. The orbifold group Γ is abstractly the obvious Fuchsian group, and has an involution induced by τ. If F sits as an incompressible suborbifold in some 3-dimensional hyperbolic orbifold M, then we get the following generalization of theorem 2.6.

Theorem 2.8. *Suppose the orbifold group of M has no 2-torsion, and that any two elliptic points on F which are interchanged by τ have the same order and lie on a manifold component of the singular set of M. Let M_F be the covering of M with $\pi_1 = \Gamma$. Then τ induces an isometry $\hat{\tau}$ of M_F which is an involution and induces the isomorphism τ, on Γ.*

Proof. The proof is the same sort of calculation with traces which gives theorem 2.7 above. The requirement that the orbifold group of M have no 2-torsion implies, by [9], that the representation of Γ lifts to $SL_2(\mathbf{C})$. Note that the singular set of M is a manifold, so that the elliptic points of F which lie on a given component of M yield generators which are conjugate in $\pi_1(M)$ (and therefore in $SL_2(\mathbf{C})$.) Hence the elliptic generators of Γ which are interchanged by τ have the same trace. The rest of the argument goes through as in 2.7.

3. Minimal surface arguments

In this section we prove the embeddedness and uniqueness theorems for least area surfaces which will be used in the following section to prove theorem 1.3. The results, and indeed the proofs, are similar to those in the paper of Freedman, Hass, and Scott [11], but the problem is complicated by the non-compactness of the manifolds we study. Instead of repeating all of [11], we will discuss the points at which their proofs have to be modified, and supply new ingredients as necessary. At various points in the argument, there will be immersions that are not necessarily in general position. The proofs given here will work when the immersions are in general position; the modifications necessary to deal with the general case are exactly those in [11], so we will not discuss them here.

What makes the theorems work, ultimately, is the fact that the cusps of a hyperbolic manifold are well understood. A cusp, by definition, is the quotient of a horoball region of $\mathbf{H}^3$ by a subgroup consisting of parabolic motions with a single fixed point at ∞; we will refer to a $\mathbf{Z}$ or $\mathbf{Z}\oplus\mathbf{Z}$ cusp depending on the subgroup. In a slight misuse of terminology, an end of a surface which is homeomorphic to a half-open annulus will be called a cusp of the surface, even if there is no specified hyperbolic metric on the surface. Unless it is specified to the contrary, a surface will have finitely many cusps, and a proper map of a surface to a hyperbolic 3-manifold will have the property that the cusps of the surface are sent to cusps of the 3-manifold. The usual definition of incompressibility makes sense for F a non-compact surface; however we need to say what we mean by boundary-incompressible in this context.

Definition 3.1. *A proper map of F to the hyperbolic manifold M is boundary-incompressible if every proper map of $\mathbf{R}$ to F which is properly homotopic in M to a map to a cusp of M is properly homotopic in F to a map into a cusp of F, and it is boundary-incompressible in the usual sense.*

Suppose f is a proper map of the surface F to the hyperbolic manifold M. The reader can verify that if f extends to a map of a compact surface with boundary into a bounded manifold, then f is boundary-incompressible if and only if its extension is in the usual sense. We will also assume that our surfaces are not homotopic to a surface lying in a cusp; this is the same as not being boundary-parallel in a bounded manifold. Look at the image of f in one cusp C of M, which may be assumed to be the image of the region $t \geqq t_0$ in the upper half-space model of $\mathbf{H}^3$. For $t \geqq t_0$, let C_t be the portion of the cusp above height t, so that ∂C_t is the torus or annulus in the cusp at height t. Let $A(t)$ be the area of f in C_t and $l(t)$ be the length of the (possibly singular) curves $f(F) \cap \partial C_t$. With this notation we have the following easily verified lemma.

Lemma 3.2. *For* $t \geqq t_0$, $A(t) \geqq \int_t^\infty \frac{l(s)}{s} ds$.

It follows that if f is a map with finite area, then $l(t)$ can be made arbitrarily small by choosing an appropriate (large) t.

Fixing the point at infinity of a cusp gives it a product structure as $\bigcup_{t \geqq t_0} \partial C_t \times t$. We say that f is a product in a cusp, if for $t \geqq$ some t_0, f can be parametrized as $\gamma \times t$ for a curve $\gamma \subset \partial C_{t_0}$. Note that the area of a product is given exactly by the length of the base curve γ.

Lemma 3.3. *If f has finite area, and $\delta > 0$ is given then there is a map g properly homotopic to f, with g a product in the cusps and area$(g) \leqq$ area$(f) + \delta$.*

Proof. Since F has finitely many cusps, it suffices to prove the lemma for just one of them. Let c be a curve in F freely homotopic to the cusp under consideration. By a small perturbation of f, which changes area by less than $\delta/2$, we can assume that f is transverse to almost all the ∂C_t. Thus $f^{-1}(\partial C_t)$ is a finite union of circles which are either homotopic to c or bound discs in F. Let α_t be the outermost among the curves homotopic to c, and γ_t be its image under f. For some t, which can be chosen arbitrarily large, the length of γ_t will be less than $\delta/2$, so that the area of the product annulus based at γ_t will have area less than $\delta/2$. Putting this annulus together with f restricted to the rest of F (inside of α_t) produces the desired map g. Since the only change has been in the cusp, it is easily verified that g is properly homotopic to f; g will be a product past a compact region of the cusp of M.

This technique of straightening F out in the cusps leads to the following lemma which will play a key role in what follows.

Lemma 3.4. *Let f be a proper general position immersion of F in a hyperbolic manifold such that f is a homotopy equivalence relative to the cusps. If f is not an embedding, then there is an embedding f' properly homotopic to f, with area $(f') <$ area (f).*

Proof. If f is not an embedding, then according to lemma 3.3 we can change it to a map f' which is a product in the cusps, increasing the area by an aribtrarily small amount (to be specified in a moment.) Once f is a product in the cusps, we can apply the mostly combinatorial arguments of section 2 of [11] to show that there is an embedding f'' with area less than area(f'). (The hypothesis that f' is a product in the cusps is used to say that the double point set has finite type so that the tower of covering spaces constructed in the proof of theorem 2.1 of [11] has only a finite number of stages.) But there is an *a priori* lower bound on how much the area is reduced in the course of the argument: Let x be a double point of f in M and δ be the area saved by rounding a corner at x after an area swap which switches the sheets of F passing through x. Since δ is independent of where we started making f a product in the cusps, if we make f a product far enough out so that $area(f') \leq area(f) + \delta$, then the map f'' resulting from the area swap argument will have area strictly less than area(f).

We have not yet established the existence of least area maps; once we do (see below), the preceding argument will yield:

Corollary 3.5. *A least area map f of F to a hyperbolic 3-manifold which is a homotopy equivalence relative to the cusps is an embedding.*

Proof. According to work of Gulliver [14], a least area map must be an immersion. If it is in general position, the argument of the preceding lemma applies to find an embedding of smaller area, contradicting the least area property of f. If f is not in general position, we use the local picture of a non-transverse self-intersection exactly as in lemma 2.5 of [11] and repeat the above argument.

The proof of our main theorem uses various properties of least area surfaces; hence we need to verify that such surfaces actually exist in finite-volume hyperbolic manifolds. If the ambient manifold is not compact, then it is not *homogeneously regular* in the sense used by Morrey [23], because the injectivity radius goes to zero in the cusps; hence the proofs of existence of least area maps given, for example in [26, 25], do not work. (We remark that M. Anderson [2] has shown the existence of stable minimal surfaces in the finite volume case; his surfaces are not necessarily of least area, however.) To get around this difficulty, we use recent work of Hass and Scott [15] which first combines a topological argument with the local existence theorem of Morrey [23] to establish the existence of an embedding which has least area *among embeddings* of a surface in a homogeneously regular manifold and then uses this to prove the existence of least area maps. This method adapts to the context of finite volume hyperbolic manifolds. In order to carry out the program of [15] we need some preliminary lemmas.

Lemma 3.6. *Suppose M is $\mathbf{H}^3$ modulo $\mathbf{Z} \oplus \mathbf{Z}$ or $\mathbf{Z}$. A minimal surface in M cannot have a local minimum (with respect to height function in the cusp.)*

Proof. This is a standard consequence of the maximum principle for minimal surfaces; the point is that the torus or annulus at a given height is a surface of constant mean curvature, with mean curvature vector pointing into the cusp, so a minimal surface cannot lie (even locally) all to the side of the torus or annulus closer to the cusp.

Note that this argument also applies to minimal immersions as well. Of course there may be local maxima; the picture to keep in mind is a horizontal plane (in the upper half-space model of $\mathbf{H}^3$) tangent to a hemisphere. The hemisphere is totally geodesic (and hence minimal) but clearly has a local maximum. The other fact we will need is a property of least area disks in a cusp.

Lemma 3.7. *Let M be a horoball, and let γ be an embedded curve in ∂M. Then γ bounds an embedded disk D in M which is least area among all disks bounded by γ. If γ and γ' are disjoint, then D and D' will be disjoint. Moreover, if F is a non-compact planar surface with boundary γ which hits the horospheres ∂M_t for all sufficiently large t, then $area(D) < area(F)$.*

Proof. The existence and embeddedness of such a disk is a theorem of Meeks and Yau [21], as is the disjointness property. For the other part, suppose that F is non-compact, and that $area(D) - area(F) = \varepsilon > 0$. As in lemma 3.2, the length of the intersection Γ_t of F with ∂M_t can be made small by taking t sufficiently large. The isoperimetric inequality in the Euclidean plane ∂M_t implies that the collection of disks bounded by Γ_t in ∂M_t has area bounded by a constant times $(length(\Gamma_t))^2$. Hence by choosing t sufficiently large, we can insure that the components of Γ_t bound a family of disks with total area less than ε. Surgery on F, using this family of disks, will result in a disk (not necessarily embedded) with smaller area than D. This contradiction shows that for F as in the hypothesis, $area(F) \leqq area(D)$.

Suppose that $area(F) = area(D)$; this implies that F is a minimal surface. Take a horosphere H just above ∂M, and transverse to F, so that $H \cap F = \gamma'$ is a single curve. As above, γ' bounds a least area disk D', and surgery on F using D' yields a disk of area less or equal to that of D. If the area is less than D, we are done. If $area(D') = area(D)$, then D' will be a least area surface, and hence smooth. But D' coincides with F on an open set (below H), contradicting the unique continuation principle [3, 11].

Remark 3.8. Because of the disjointness property for disks in M, we get the existence of embedded least area disks with given null-homotopic boundaries in a $\mathbf{Z} \oplus \mathbf{Z}$ or $\mathbf{Z}$ cusp C. Also, by lifting to the horoball $\tilde{C}$, we get that such a disk has strictly less area than an inessential planar surface in C with the same boundary.

The existence proof we provide uses a convergence argument for a sequence of surfaces with area decreasing to a minimum value. There are two obvious ways in which such sequence might fail to converge. The first is that some inessential curve on the surface might bound a sequence of disks which 'bubble off' to infinity, leaving a puncture in the limit. This will be prevented with the use of lemma 3.7. The other worry is that some curve on F which is homotopic to a cusp in M could be pushed out into the cusp, removing that circle in the limit. (See Fig. 3.)

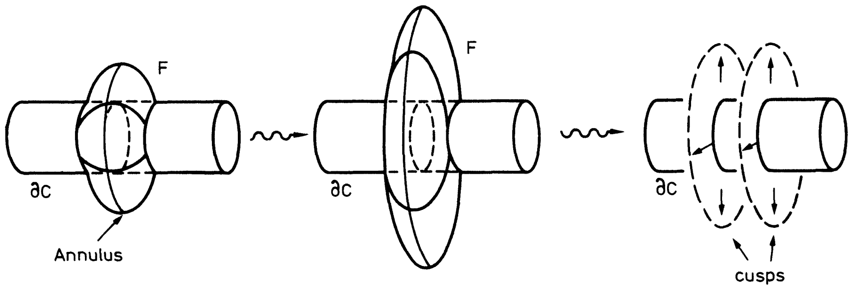

Fig. 3. Annuli converging to a pair of cusps

The following lemma will take care of this possibility.

Lemma 3.9. *Let C be a $\mathbf{Z} \oplus \mathbf{Z}$ cusp in a hyerbolic manifold M. Then there exist numbers t_0 and ε_0 with the following property. If $(A, \partial A) \subset (C_{t_0}, \partial C_{t_0})$ is an essential* *annulus such that $A \cap \partial C_{2t_0}$ contains an essential curve, there is an annulus $A' \subset \partial C_{t_0}$ with the same boundary, and with area$(A') \leqq$ area$(A) - \varepsilon_0$.*

Proof. Let γ be the shortest essential curve on some torus ∂C_t. The area of an essential annulus running from ∂C_t to ∂C_{2t} in any other homology class will be greater than or equal to the area of the product annulus from ∂C_t to ∂C_{2t} in the homology class of γ. By lemma 3.2, this is proportional to $1/t$. On the other hand, the area of ∂C_t is proportional to $1/t^2$. If A is an annulus as in the hypotheses, there must be some sub-annulus of A with boundary on ∂C_t and ∂C_{2t}, so the lemma follows by taking t large enough.

Theorem 3.10. *Let f be a proper embedding of a surface F of finite area in a finite volume hyperbolic manifold, and let $\mathbf{F} = \{$embeddings of F isotopic to $f\}$. Then $\mathbf{I} = \operatorname{Inf}\{area(g) | g \in \mathbf{F}\}$ is realized, in the following sense: Either there is an embedding of least area among embeddings isotopic to f, or there is a 1-sided embedded surface K with area$(K) = \frac{1}{2}\mathbf{I}$ with F isotopic to the boundary of a tubular neighborhood of K.*

Proof. Write M as the union of a compact manifold M_{thick} homotopy equivalent to M, and the cusps of M. Let B_i be a locally finite collection of balls covering M; note that each B_i is isometric to a closed ball in hyperbolic space. Using the fact that $\mathbf{H}^3$ is homogeneously regular [23], B_i has the property that the least area disc in $\mathbf{H}^3$ spanning an embedded circle on the boundary of B_i must in fact be embedded and lie in B_i. By lifting a disc in M to $\mathbf{H}^3$, we see that an embedded circle on ∂B_i bounds an embedded least area disc in B_i. Note that by compactness of M_{thick}, only a finite number of the B_i hit M_{thick}; also we may assume that only finitely many B_j hit a given B_i.

Let F_i be a sequence of embedded surfaces in $\mathbf{F}$ with the property that area(F_i) converges to $\mathbf{I}$. Given the existence of least area discs in B_i spanning curves in ∂B_i, sections 4 and 5 of [15] describe a procedure for replacing the sequence $\{F_i\}$ with sequences $\{F_{i,j}\}$ of surfaces in $\mathbf{F}$ with the properties that:

(i) For all j, $\lim_{i \to \infty} area(F_{i,j}) = \mathbf{I}$

(ii) For each n and each $j \leqq n$, the intersection of $\{F_{i,n}\}$ with $\operatorname{int}(B_j)$ converges to a (possibly empty) union of least area discs.

(iii) If $j < n$ then $F_{i,n} \cap B_j = F_{i,j} \cap B_i$ for all i.

Convergence in (ii) means that the limit consists of all the limit points of the sequence, and that any limit point has a disc neighborhood which is the limit (in the C^∞-topology) of maps of discs into $F_{i,n}$.

Now take a sequence $F'_j = F_{i(j),j}$ where $i(j)$ is large enough so that area $((F_{i(j)j}) - \mathbf{I} < 1/j$. By properties (ii) and (iii), any limit point of $\{F'_j\}$ must be contained in some B_i, and hence is on a least area disc. Therefore the limit is empty or is a surface. But if the limit is empty, then for sufficiently large j, $F'_j \cap M_{thick}$ will be empty, and so must lie in a cusp. But this is impossible since F was incompressible. Let G be the limit surface; then clearly G is a closed subset of M and is hence properly embedded. In the case of compact manifolds, the argument of [15] is that the F'_j are close to G and transverse to the fibers of a tubular neighborhood of G, so that each F'_j is either isotopic to G or double covers G. However, our construction insures only that the F_j' are close to G on compact sets, so there is still some work to do. From now on we will just write F_i for F_i'.

Consider first the case of a properly embedded essential half-open annulus A in a cusp C, where we are looking for the least area such half-annulus with a fixed boundary $\gamma = \partial A$. Section 6 of [15] shows how to obtain half-annuli A_i all with the same boundary converging to a surface E in the same sense as (i)–(iii) above. Suppose K is a compact set in C. Then we have convergence in K, so that for i sufficiently large $E \cap K$ is isotopic to or doubly covered by $A_i \cap K$. By construction E is a minimal surface, so by lemma 3.6, the height function in C restricted to E has no minima, except of course at γ. This implies that for any t, the portion of E below height t is connected. Since the A_i all have boundary equal to γ, it follows that any compact subsurface of E will in fact be isotopic to a subsurface of the A_i for sufficiently large i. Hence any compact subsurface of E is planar, so E is itself a planar surface.

Now let t be any level such that E is transverse to ∂C_t.

Claim 1. There is at most one essential curve of intersection of E with ∂C_t.

Proof of claim 1. Suppose that the claim is false, and that there is more than one such curve at some level t. For i sufficiently large, the A_i will approximate the portion of E below height t; in particular we can assume that A_i intersected with the portion of C below height t is connected. But a connected subsurface of A with γ part of its boundary will be a sub-annulus of A minus some disks. In particular, there can only be one other essential boundary component other than γ. Since ∂C_t is incompressible, all the other curves must be inessential in ∂C_t, proving claim 1.

Now choose a sequence of levels $\{t\}$ going to infinity so that E is transverse to the ∂C_t. By claim 1, the intersection with each ∂C_t consists of one essential curve and some inessential curves.

Claim 2. Each inessential curve bounds a disk on E.

Proof of claim 2. Suppose to the contrary that some inessential curve α doesn't bound a disk on E. Note first that since since E is planar and the portion of E below height t is connected, α must be the boundary of the component of $E \cap C_t$ on which it lies. By the convergence on compact sets, α is approximated by circles on the A_i which bound disks on the A_i. If some sequence of these disks stays below level t, then it converges to a least area disk lying in E. This easily contradicts lemma 3.6. Since the component bounded by α is approximated on compact sets by subdisks of the A_i, it follows that this component must be an inessential planar surface P. By lemma 3.7, there is a disk in C_t with strictly smaller area than P, and with boundary α. It is easy to see that this leads to a contradiction.

These two claims now imply that the surface E is a least area annulus which is isotopic to A. We remark that an argument similar to those in lemma 3.7 shows that E has area strictly smaller than that of any planar surface P with the same boundary such that image $(\pi_1(P))$ in $\pi_1(C)$ is the same as that of A.

We now return to proving the existence for the whole surface F. We will consider the case when there is only one cusp of F in a given cusp C; the general case is similar, with a little more bookkeeping. For any level t such that G is transverse to ∂C_t, the intersection of G with ∂C_t consists of finitely many essential and inessential curves. Note that by incompressibility of F, its intersection with C will consist of subsurfaces of F which are punctured annuli and disks, where the punctures all bound disks on F and are inessential in ∂C as well. The same remark applies to the F_i as well. Let t_0 and ε_0 be the constants from lemma 3.9. Consider the portion of G that lies between ∂C_{t_0} and ∂C_{2t_0}. Every component is near to and hence covered by subsurfaces of the F_i for sufficiently large i. It can be shown that the covering must be orientable, and hence that each such component is isotopic to a connected subsurface of F_i.

Claim 3. There is only one essential component of intersection of G with ∂C_{2t_0}.

Proof of claim 3. Suppose that there is more than one essential component of intersection of G with ∂C_{2t_0}; the same will be then true for all of the F_i. It follows that there will be an essential punctured annulus A in one of the F_i with its essential boundary components on ∂C_{t_0} and with an essential curve of intersection with ∂C_{2t_0}. Assume that i is large enough so that $area(F_i) - \mathbf{I} < \varepsilon_0$. If there are inessential curves of $A \cap \partial C_{t_0}$, then by reflecting the parts of A which go below level t_0 across ∂C_{t_0}, we obtain an annulus with smaller area lying in C_{t_0}, and the same non-trivial boundary components. By lemma 3.9, there is an annulus A' (in ∂C_{t_0}) with $area(A') \leqq area(A) - \varepsilon_0$. Replacing A by A' on F_i gives rise to a mapped-in copy of F with area less than $\mathbf{I}$; this may be replaced by an embedded copy of F with no greater area, which is a contradiction. Hence claim 3 is established.

Now consider the portion of G lying above level $2t_0$. As in the proof of claim 2, each inessential circle of $G \cap \partial C_{2t_0}$ bounds an inessential planar surface in C_{2t_0}. Using lemma 3.7, replace each inessential non-compact by the least area disk with the same boundary, and the essential component by the least area half-annulus constructed above. The result (after possibly cutting and pasting to get an embedded surface) is a surface isotopic to F, with area equal to $\mathbf{I}$. If there is more than one cusp of F in a cusp C, it is possible that the portion of G below C is double covered by a subsurface of F. The covering on the boundary is trivial, so one still adds a single copy of the half-annulus to get a surface which is double-covered by F.

Remark 3.11. The preceding argument used the hypothesis of finite volume only to establish that the limit surface G was non-empty. Hence the same argument would hold in any *geometrically finite* [30] manifold, where the 'convex core' has finite volume.

Corollary 3.12. *Let f be a proper map to a geometrically finite hyperbolic 3-manifold which is incompressible and boundary incompressible. Then f is properly homotopic to a least area map.*

Proof. The same corollary is deduced in [15] by an argument involving the covering M_F to which f lifts. The key point is that the lifted map can always be replaced by an embedding of smaller area; we have shown this in lemma 3.4.

The following result is crucial to the the proof of theorem 1.3, and will also be used in the remainder of this section.

Proposition 3.13. *Least area homotopy equivalences which are properly homotopic either coincide or are disjoint.*

Proof. By a previous corollary (3.5) the two maps, say f and g, are both embeddings. Assume that f and g are transverse; their intersection is then a properly embedded 1-manifold. If there were only a finite number of components of $f(F) \cap g(F)$, then the arguments of Waldhausen [31] would provide a product region between $f(F)$ and $g(F)$. The standard area-swap argument then shows how to reduce the area of both f and g, contradicting minimality. Likewise, if there are an infinite number of closed components of $f(F) \cap g(F)$, then either there must be some which are inessential circles, or there are pairs of essential circles in cusps of $f(F) \cap g(F)$. In either case, the area swap argument produces a contradiction.

The trick of replacing f and g by maps which are products in the cusps would complete the argument, just as in lemma 3.4, if we knew that the maps which replace f and g were embeddings. But this can be guaranteed with the use of lemma 3.2: Look at some annulus ∂C_t which is transverse to both f and g, and assume that t is large enough and that $l(t)$ is small for both f and g. (Here $l(t)$=length of the intersection of F with R_t.) Let φ and γ be the essential circles of $f(F) \cap R$ and $g(F) \cap R$, respectively. The circles which are trivial on F bound discs on R which are disjoint from the essential circles φ and γ. By the isoperimetric inequality for the Euclidean plane which covers R, the area of a disc is bounded by the square of the length of its boundary circle, and so the discs on R can be taken to have small area. Now replace f by the surface which is a product over φ, and fill in the inessential circles by the discs they bound on R, pushed up a little if they are nested, and do the same to g. The resulting surfaces are clearly embedded, and by taking t sufficiently large, the area has been increased by an arbitrarily small amount. The argument is finished as in lemma 3.3; choose a point on the original intersection of f and g, and let δ be the area reduction that would result from rounding a corner at that point. Do the construction outlined to make f and g be products in the cusps, arranging that the area of neither goes up by more than $\delta/2$. The area swap argument will result in an area reduction of at least δ, contradicting the minimal area property of f and g.

As in lemma 3.3 of [11], the proposition implies that if f is a least area homotopy equivalence, and $\tilde{f}: \tilde{F} \to \tilde{M}$ a finite cover of f, then $\tilde{F}$ has least area as well. It is also used in proving the following proposition which states that a least area map which is properly homotopic to an embedding is so by a particularly nice proper homotopy.

Proposition 3.14. *Let f be a least area map which is properly homotopic to an incompressible, boundary incompressible embedding. Then there is a bounded function d on F which goes to zero in the cusps of F, and a proper homotopy H to an embedding, such that every point x of F is moved a distance $\leqq d(x)$ by H.*

Proof. In each cusp of M, choose a torus T to which f is transverse, and let γ be the essential curve in F which maps to T. Since f is homotopic to an embedding, γ is homotopic to an embedded curve α which is the image of a horocycle in T. Let A be the annulus in F with $\partial A = \gamma$, and let C be the cusp with boundary T. By applying lemma 3.6 to the lift of A to the covering of M with $\pi_1 = \pi_1(C)$, we see that A must lie entirely in C. Since F minus its cusps is compact, it suffices to show that there is a proper homotopy of (A, γ) in (C, T) to an embedding with every point moving distance less than some number d.

Since f is a least area map, $f|A$ has least area in its proper homotopy class among all maps with the same boundary $f(\gamma)$. Consider the lift of f to C_A, the cover of C corresponding to $\pi_1(A)$, and let t generate the covering translations. Since it is a homotopy equivalence (rel ∂), f is now an embedding. The lift of γ to ∂C_A is a compact curve, and hence lies between β and $t^k(\beta)$ for some horocycle β which projects homeomorphically to α. Now β and $t^k(\beta)$ lie on totally geodesic annuli; these annuli are of least area, and so by lemma 3.13 they must be disjoint from A. There is a level preserving ambient isotopy of the region between β and $t^k(\beta)$ to the region between β and $t(\beta)$; note that the distance this isotopy moves points of F goes to zero as you go into the cusp. Projecting this isotopy restricted to the lift of A back down into C provides the required proper homotopy: A is embedded since it is embedded upstairs and lies between β and $t(\beta)$, and the bound on how far points are moved is the same as upstairs.

Lemma 3.15. *Let C be a $\mathbf{Z}\oplus\mathbf{Z}$ cusp, let $A = S^1 \times [0, \infty)$, and let γ be an immersed curve in ∂C with the property that γ lifts to an embedding in the covering corresponding to A. If f is a proper immersion of A in C (in general position) which is least area among all such maps with boundary equal to γ, then the singular set of f has only a finite number of components, all with non-empty boundary on ∂C. If there are several such annuli, each of least area rel boundary and properly homotopic to one another, then their mutual intersections have the same property.*

Proof. Look up in the covering $\tilde{C}$ to which f lifts; by an obvious extension to the bounded case of corollary 3.5, the lift of f is an embedding as it is least area. If there are any self intersections that have no boundary on ∂C, then there will be an intersection between two lifts of A, say A_1 and A_2 which is either a circle or a properly embedded line. A circle either bounds a disc or an open annulus (the same on A_1 and A_2 by incompressibility) and a line bounds a properly embedded half-plane on both A_1 and A_2. In all three cases, an area swap reduces area without changing the boundary, which is a contradiction.

We are now in a position to prove our embeddedness theorem.

Theorem 3.16. *Let f be a least area map from the surface F to the finite-volume hyperbolic manifold M which is proper, incompressible, and boundary incompressible If f is properly homotopic to an embedding, then f is itself an embedding, or double covers an embedded one-sided surface.*

Proof. According to [14] f can be taken to be an immersion. We assume in addition that f is in general position; if it is not the arguments have to be modified somewhat along the lines of [11]. The argument of [11] is in brief (using the notation of 5.1 of [11]): The preimage of F in $\tilde{M}$ is a union of embedded planes, so to see that f is an embedding (or double covers an embedding), one must show that two of these planes cannot intersect. Fix two intersecting planes, say $\tilde{F}$ and $g\tilde{F}$, and let $G=\pi_1(F)\cap g\pi_1(F)g^{-1}$. If F is compact, standard 3-manifold topology shows that there is a compact product region X between subsurfaces of the images of $\tilde{F}$ and $g\tilde{F}$ in $M_G=\tilde{M}/G$. Finally, the "LERF" property [27] of surface groups is used to show that X embeds in finite coverings of M_F and M_{gF} where an area exchange produces a contradiction; this requires the compactness of X and thereby of F.

Let F_1 and F_2 be the images of $\tilde{F}$ and $g(\tilde{F})$ in M_G, and let A and B be the (closures of) the components of M_G-F. According to proposition 3.14, there is a proper homotopy H_t of F to an embedded surface, and a bounded function $d(x)$ going to zero in the cusps of F, such that no point of F moves more than distance d under F. It follows that either $F_2\cap A$ or $F_2\cap B$ lies within a $2d$-neighborhood of F_1, and that the same is true for the images in M_F; suppose it is $F_2\cap A$. As in [11], the projection of F_2 into M_F factors through at most a 2-fold cover, and is proper. There are two cases to consider: either the image of $F_2\cap A$ in M_F is compact or not. If it is compact, the proof finishes exactly as in [11] and outlined above. If it is not compact, then since it lies within the $2d$-neighborhood of F in M_F, the intersection of the image of F_2 and F must go out to infinity only in the cusps of M_F. Let C be a ($\mathbf{Z}$) cusp of M_F in which some part of the intersection lies. Note that there is a cusp $\hat{C}$ of M_G, intersecting F_2, which maps properly onto C. The preimage of F (viewed as in M) in C consists of a $\mathbf{Z}$'s worth of copies of the intersection of F with C.

Now the intersection of F with C is the union of a finite number of discs, and an annulus P which is least area among all annuli with the same boundary. Suppose F_2 maps to an annulus P'. (We will ignore the discs because they are compact and finite in number.) Since both P and P' are least area with fixed boundary, lemma 3.15 implies that there are finitely many non-compact components of intersection, and that the intersection is eventually a product, at least topologically, over a finite set of points.

Since $p|F_2$ is proper, the same is true up in M_G. So the 3-manifold argument of section 3 of [11] applies to give a product region, say X, between F_1 and F_2 up in M_G. If X were compact, the LERF property of surface groups would imply that X injects into finite covering spaces of M_F and M_{gF}, and an area exchange in these covering spaces would finish the argument. But X fails to be compact in a relatively tame way only by going out into cusps, so we may proceed in the following manner: Choose annuli ∂C_j cutting off the cusps C_j of M_F, and let $\hat{C}_j$ be the cusps of M_G sitting above $\{C_j\}$ which intersect F_2. Since $p|F_2$ is proper, $p|\hat{C}_j$ is proper also. Let Y_j be the intersection of F and $p(F_2)$ with ∂C_j; Y_j is compact. Since $p:\hat{C}\to C$ is proper, the inverse image of Y_j is compact, and hence lies in a region R between two circles in $\partial\hat{C}_j$. Now as in the proof of proposition 3.14 the whole intersection of F_1 and F_2 *in* $\hat{C}_j$

must lie between these circles. Finally, let Z be the compact set consisting of the (compact) part of the product region X lying below the C_j, and the finitely many annular regions $\{R\}$ on the ∂C_j discussed above. Using the LERF property of surface groups, Z injects into infinite covers of M_F and M_{gF}. But if the region R maps injectively, then everything lying above it maps injectively, so the whole product region X must map injectively into these covering spaces. Hence the area exchange argument works in this case as well.

In proving the analogue of theorem 1.3 for orbifolds, we will need to use existence and embeddedness results for least area incompressible orbifolds in hyperbolic orbifolds. Since we will be dealing with hyperbolic orbifolds, there will always be a finite (regular) cover which is a manifold, according to the well-known lemma of Selberg [28]. The standard argument in the compact case [20] to get least area embedded orbifolds in an orbifold uses a uniqueness property [11] of least area surfaces in a manifold covering space; in our case we need that the same result holds for hyperbolic orbifolds of finite volume.

Theorem 3.17. *Let F and G be properly embedded incompressible and boundary-incompressible least area surfaces in a finite volume hyperbolic manifold If F and G are properly homotopic to disjoint surfaces, then they are disjoint or coincide.*

Proof. The proof follows the proof of theorem 6.2 of [11], with modifications as in the preceding theorems to take the cusps into account.

As in the compact case [20], an immediate corollary is that the theorems about least area surfaces extend to theorems about orbifolds.

Corollary 3.18. *An incompressible, boundary-incompressible least area orbifold in a hyperbolic orbifold of finite volume which is properly homotopic to an embedding is itself embedded. If the inclusion of F is a homotopy equivalence relative to the cusps, then any other such orbifold either coincides with F or is disjoint from F.*

4. Proof of the main theorem

In this section we use the material of the preceding sections to prove theorems 1.3 and 1.4.

Proof of theorem 1.4. Suppose that (F, τ) is a symmetric surface from Fig. 2, and that F is properly embedded in M which is a complete hyperbolic manifold of finite volume. F can be regarded as the interior of an embedding $\bar{f}:(\bar{F}, \partial \bar{F}) \subset (\bar{M}, \partial \bar{M})$; by assumption, this embedding is incompressible and boundary-incompressible. According to theorem 3.16, F is isotopic to a least area embedded surface, or perhaps double covers a least area embedded surface K. Consider M_F, the covering space of M corresponding to the subgroup $\pi_1(F)$ of $\pi_1(M)$; if F double covers K, there is also a covering space M_K corresponding to $\pi_1(K)$ which is double covered by M_F.

By theorem 2.5, there is an isometry $\hat{\tau}$ of M_F inducing τ_*. We claim that there is an embedding of F in M_F which is invariant under $\hat{\tau}$ and which projects to an embedded copy of F in M. In the case that F is isotopic to a least area surface, let $\hat{F}$ be its lift to M_F. If F double covers a least area surface K, then let $\hat{K}$ be the lift of K to M_K, and let $\hat{F}$ be the preimage of $\hat{K}$ under the 2-fold covering $M_F \to M_K$. In the first case, $\hat{F}$ will be the desired surface; in the second case, we will use the boundary of a tubular neighborhood of $\hat{F}$.

The first step in showing that these surfaces have the property of invariance under $\hat{\tau}$, is to notice that M_F has two ends (relative to the cusps of M_F) and that $\hat{\tau}$ preserves these ends. To see this, note that M_F is homotopy equivalent to F; in fact they are homotopy equivalent relative to their cusps (or 'thin parts' in the terminology of [30].) This easily implies that $\hat{F}$ separates M_F and hits some line from end to end transversally in one point. By assumption on τ, it preserves the orientation of F, and hence acts trivially on $H_2(F, \mathit{cusps})$ and so also on $H_2(M_F, \mathit{cusps})$. Since $\hat{\tau}$ is orientation preserving on M_F as well, it must preserve the orientation of a line going between the ends, i.e. must preserve the ends of M_F.

Now $\hat{F} \subset M_F$ is least area in its proper homotopy class, and hence (by proposition 3.5) is embedded. Since $\hat{\tau}$ is an isometry, $\hat{\tau}(\hat{F})$ is a least area surface which is properly homotopic to $\hat{F}$. So by proposition 3.13, either $\hat{\tau}(\hat{F}) = \hat{F}$, or they are disjoint. But they cannot be disjoint: since $\hat{\tau}$ preserves the ends of M_F, $\hat{\tau}(\hat{F})$ must either be to the right or left of F, which is clearly impossible since $\hat{\tau}$ is an involution. Therefore, in the case when the least area map of F in M is actually an embedding, $\hat{F}$ has the desired properties.

If the least area map of F double covers the least area surface K, the surface $\hat{F}$ is still of least area in M_F, so the argument of the previous paragraph implies that $\hat{F}$ is invariant under $\hat{\tau}$. Since the copy of K in M_K projects diffeomorphically onto K (in M), some tubular neighborhood U does as well. Since $\hat{\tau}$ is an isometry, we may assume that the preimage $\hat{U} \subset M_F$ of this neighborhood is invariant under $\hat{\tau}$. One component of its boundary is a thus a copy of F which is invariant under $\hat{\tau}$ and which projects diffeomorphically to a copy of F in M which is the boundary of the image of U in M.

Given the invariant copy of $\hat{F}$ in M_F, we now show that the hyperbolic metric on $M - F$ can be pieced together across F when we glue up via τ to give M^τ a metric of constant curvature -1. Let x be a point on F and $\hat{x}$ a point on $\hat{F}$ in M_F sitting above x. Let B be a small ball around x, and $\hat{B}$ a ball above it in M_F; F separates B into half-balls $B_\pm$ and likewise for $\hat{B}$. Now $\hat{B}_+$ is taken isometrically onto $\hat{\tau}(\hat{B}_+)$; in other words the metric on $\hat{B}_+$ fits together smoothly across a neighborhood of $\hat{x}$ in $\hat{F}$ to give a metric of curvature -1 on $\hat{B}_+ \cup \hat{\tau}(\hat{B}_-)$. Projecting this local picture down into M says exactly that $B_+(x)$ fits together smoothly with $B_-(\tau(x))$ to give the desired metric on M^τ in a neighborhood of x. Observe that the only property of F that is used is that it is embedded in M in the right isotopy class, and that $\hat{F} \subset M_F$ is invariant under $\hat{\tau}$. Hence we are free to use any such surface and will in fact do so in the next paragraphs.

It only remains to demonstrate that a metric constructed in this manner on M^τ is complete. For this we need some details of how the cusps of F sit in the cusps of M. M is complete and has finite volume, so a neighborhood of each end E can be parameterized as $T^2 \times (0, \lambda]$. Let γ be the (free) homotopy class represented by a cusp of F going into E. A convenient way to parameterize E for this discussion is to let $T^2 \times l$ be the torus in E such that a geodesic (in the induced flat metric on T) homotopic to γ has length exactly l. Of course, going out the end of E is now described by l going to 0. In the covering space M_F, the preimage of any torus is a union of planes and annuli; let A_j be the annulus hitting the j^{th} cusp of $\hat{F} \subset M_F$. Then there are annuli $A_j \times l$, covering $T_j \times l$, whose belt curve has length exactly l.

Consider the intersection of F with a torus $T \times l$. As in the proof of theorem 3.10, the intersection will consist of one curve freely homotopic to γ and a finite set of curves C_k bounding discs D_k (perhaps nested) on $T \times l$. These curves and discs all lift up to some A_j depending on which cusp of F the curves belong to. Suppose τ interchanges the cusps i and j of F (it is possible that $i=j$); then $\hat{\tau}$ interchanges the annuli A_i and A_j and the lifts of the discs sitting in $A_i \times l$ are exchanged with the lifts of the discs sitting in $A_j \times l$. Now replace the original surface with a new one built as follows: Throw away the part of F outside level l, and add a collar onto the essential curves in $F \cap T \times l$ using the product structure on the cusps. The inessential curves can be capped off using the discs D_k, pushed slightly above or below level l if there is nesting of the discs. The new surface (which will still be called F) is still embedded, and still has the property that $\hat{F}$ is invariant under $\hat{\tau}$. This is because it is built of pieces which were invariant under $\hat{\tau}$.

With this description, it is easy to see what the ends of M^τ look like. By the previous paragraph, we can assume for l sufficiently small that $F \cap T \times l$ is a single curve. Under cutting and pasting, the torus $T_i \times l$ gets cut along γ_i and pasted together with another torus $T_j \times l$. The same argument that showed that M^τ has a metric of curvature -1 now shows that the flat metrics on the tori at level l which get pasted together fit together smoothly across the curves $\gamma \times l$ to give flat tori sweeping out all the ends of M^τ. This is not quite enough to imply that M^τ is complete; after all removing a closed neighborhood of complete cusp produces a manifold with the same property. To finish the proof that M^τ is complete, we must show that a path out to infinity in M^τ has infinite length. It clearly suffices to show this for paths contained in any given cusp and starting, say at the torus $T \times 1$.

In each cusp of M, look at the tori $T \times 1/n$ going out towards the end, choosing $n \geqq N$ where N is large enough so that F looks like a product $\gamma \times (0, 1/N]$ in each end. Since F has this form, there is a geodesic in each cusp, missing F and perpendicular to all the tori in the cusp, the segment between the n^{th} and $(n+1)^{st}$ tori has length, say d_n, and since M is complete, it must be that $\sum_n d_n = \infty$. To see that the length of a path going out to ∞ in M^τ is infinite, it suffices to show that the distance between the tori (in M^τ) at levels $1/n$ and $1/(n+1)$ is indeed d_n. For then any path would have length $\geqq \sum_n d_n = \infty$. The region between two successive tori in an end of M^τ is homeomorphic to $T^2 \times \mathbf{I}$; it is easy to see that any such manifold is a subset of a standard $\mathbf{Z} \oplus \mathbf{Z}$ cusp. Hence the distance between the tori is given by the length of *any* geodesic between them and perpendicular to both. But since the geodesics chosen above miss F, they become geodesics in M', perpendicular to all the tori, and of length d_n. Hence the distance between successive tori is as required, and M^τ is complete.

Remark 4.1. The use of least area surfaces in the preceding argument is to insure that there is an incompressible surface in M_F which is invariant under τ and which injects into M. Such a surface can be obtained in at least two other ways. One is to use the notion of a *PL* least area surface as in [16]; another is to use the level set of a harmonic function on M_F. It is straightforward to show, that if F is quasi-Fuchsian, then there is a unique harmonic function f from M_F to (0, 1) tending to 0 and 1 at the two ends of M_F. Such a function will be invariant under τ, and so will $f^{-1}(\frac{1}{2})$. One then shows that $f^{-1}(\frac{1}{2})$ injects into M, and that cutting and pasting using it is the same as cutting and pasting using F. (The last statement is not quite obvious, because $f^{-1}(\frac{1}{2})$ may not even be a smooth surface, let alone an incompressible one.)

Theorem 1.4, when F is allowed to be an incompressible surface in a general 3-manifold, is proved essentially by showing that all the interesting changes in a manifold M which is cut and pasted by symmetries such as τ take place on the hyperbolic part of M. This is accomplished using the torus decomposition of a Haken 3-manifold, which we recall in a weak but convenient form.

Theorem 4.2. *([17, 18, 30]). If M is a Haken manifold with boundary consisting of incompressible tori, then there is a family* $\mathbf{T}$ *of embedded incompressible tori such that M* cut along $\mathbf{T}$ *consists of Seifert-fibered spaces, non-orientable line-bundles, and hyperbolic manifolds of finite volume.*

The other result we need is a theorem of C. Adams [1].

Theorem 4.3. *([1]) Let F, F' be 3-punctured spheres properly embedded in hyperbolic manifolds M, M'. Then any cutting and pasting along F and F' results in a hyperbolic manifold whose volume is* $\mathrm{vol}(M)+\mathrm{vol}(M')$.

Proof of theorem 1.4. Let F be an incompressible surface in the compact 3-manifold M whose boundary is a union of tori. It is well-known (and easy to prove) that there is an incompressible surface F' such that F' is in an irreducible summand of M and F is a connected sum (in M) of F' and some 2-spheres embedded in M. It follows directly from this fact that M^τ is the same as the manifold obtained by cutting and pasting along F'. So we may as well assume that M is irreducible. M is evidently sufficiently large as well, because by assumption, it contains the incompressible surface F.

M is now Haken, so that theorem 4.2 applies to it and we get M decomposed as a union of hyperbolic and Seifert-fibered manifolds M_i meeting along incompressible tori. As mentioned in the introduction, $|M|$ is given (up to a universal constant) by the sum of the volumes of the hyperbolic manifolds among the M_i. To prove theorem 1.4, then, we need to see that M^τ has a torus decomposition made from the pieces of M in such a way that the new hyperbolic pieces have volume equal to the sum of the volumes of the old pieces, and that none of the tori involved have become compressible. Saying that the tori are still incompressible is the same as saying that the new Seifert pieces are all non-trivial.

By standard pushing and pulling arguments, we may assume that the intersection of F with each piece M_i is incompressible and boundary-incompressible.

Let F_j be the components into which F is cut by the tori. The fact that τ is in the center of the mapping class group of F implies that the collection F_j may be taken to be equivariant in the sense that a component is either invariant under τ or is exchanged by τ with another component. The possibilities for the surfaces F_j are quite limited; they are either annuli or among the symmetric surfaces from Fig. 2. Moreover, a subsurface which is not invariant is either an annulus or a 3-punctured sphere S which together with $\tau(S)$ fills up F except for annuli. On the other hand, the possibilities for how these surfaces sit in the different M_i are also limited; the annuli must be vertical annuli in the Seifert pieces, and a non-annular component is either in a hyperbolic piece or is in a Seifert piece as the fiber of a bundle over S^1 with periodic monodromy.

Consider first the case of a component F_j of F which is invariant under τ and sits in a torus summand N of M. If N is a hyperbolic piece of M, theorem 1.3 applies, so that N^τ becomes hyperbolic, with its volume the same as that of N. If F_j is a fiber of a bundle over the circle with periodic monodromy, then τ, being in the center of the mapping class group of F_j, commutes with the monodromy. Therefore N^τ is also a bundle over S^1 with periodic monodromy, and is therefore Seifert-fibered with incompressible boundary. The other possibility in this case is that F_i is a vertical annulus in a Seifert piece N. The involution τ restricted to F_i can be taken to be fiber-preserving, so that cutting and pasting of N can be achieved by cutting and pasting the base surface of the fibering of N along the arc to which F_i projects. Since N was a non-trivial Seifert fibered space, it is not hard to see that N^τ will be as well.

The other case, in which components F_i and F_j lying in pieces N_i and N_j are interchanged, is dealt with in a similar manner. If the pieces N_i and N_j are both hyperbolic, then F_i and F_j must both be 3-punctured spheres, so Adams' theorem (4.3) applies and says that the glued up manifold is hyperbolic with volume the sum of the volumes of N_i and N_j. If, say, N_i is hyperbolic but N_j isn't, then cutting N_j produces $F_j \times \mathbf{I}$, so gluing up via τ is the same as cutting N_i along F_i and regluing by an arbitrary automorphism of F_i. But again by [1] this produces a hyperbolic manifold of the same volume as N_i. *If* F_i and F_j are annuli in Seifert-fibered pieces, then as in the previous case, the cutting and pasting can be done by cutting and pasting of arcs in the base surfaces of the fibrations, producing non-trivial Seifert pieces.

The cutting and pasting that is the key to theorem 1.3 can be used in other situations. We give one such generalization:

Theorem 4.4. *Let F be an incompressible, boundary-incompressible surface in the finite volume manifold M, and let ρ be the induced representation of $\pi_1(F)$. Suppose that F is not (virtually) a fiber in a fibering of M over S^1 and that φ is an orientation preserving diffeomorphism of F such that $\rho \circ \varphi$, is conjugate to ρ. Then M^φ is hyperbolic and has the same volume as M.*

Proof. (sketch) The main point is to show that $\hat{\rho}$, the isometry induced on M_F by conjugating ρ to $\rho \circ \varphi_*$, has finite order. For this we must use the work of Thurston, Bonahon and Bers described in remark 2.8. As discussed there, since F is not a fiber in a fibering over S^1, the representation ρ must be quasi-Fuchsian. Also, ρ is uniquely determined by the two hyperbolic structures 'at infinity in M_F'. Since $\hat{\rho}$ is an isometry of M_F, it must be an isometry of these ideal hyperbolic structures, and therefore must have finite order. The rest of the proof follows the proof of 1.3 exactly.

We remark that the above theorem is definitely not true if F is a fiber. For if φ is the monodromy, then cutting and pasting via φ^k multiplies the volume by $k+1$. For instance, cutting and pasting using φ^{-1} produces $F \times S^1$, which isn't even hyperbolic.

5. Dehn surgery and branched covers of mutant knots

The technique that proves theorem 1.3 gives further information on the degree to which mutant knots resemble one another. We show in this section that if the branched cover of a knot is a hyperbolic manifold, then the branched cover of any mutant knot has the same volume. Similarly, it is shown that Dehn surgeries on mutant knots have the same volumes.

Theorems about branched covers can be put in a different context by considering *orbifolds* [30], or spaces whose local structure is that of a manifold modulo a finite group action. In this terminology, the p-fold branched cover of a knot is a p-fold orbifold covering of the orbifold with underlying space S^3 and singular set equal to the knot. The volume of the branched covering is p times the 'orbifold volume' of the associated orbifold. Define a symmetric orbifold to be one of the symmetric surfaces of Fig. 2, with cusps which correspond under τ filled in with the same cone angle, or perhaps not at all. With this terminology, the theorem about branched covers can be stated as:

Theorem 5.1. *Let the symmetric orbifold F be embedded as a suborbifold of the hyperbolic orbifold M and suppose that any two elliptic points of F which are interchanged by τ have the same odd order, and lie on the same component of the singular set of M. Suppose further that there is no 2-torsion in the orbifold fundamental group of M. Then the orbifold M^τ is hyperbolic, and has the same volume as M.*

Proof. (Sketch) The idea is the same as the proof of theorem 1.3. Theorem 2.8 says that the representation of the orbifold fundamental group of F in $\mathrm{PSL}_2(\mathbf{C})$ is conjugate to its composition with τ_*, and hence that there is an involution of the covering space M_F. One needs to find an embedded orbifold isotopic to F with the property that its lift to M_F is invariant under $\hat{\tau}$. This is done in the same way as in the proof of theorem 1.3, using the least area orbifold homotopic to F which is provided by theorem 3.18.

Corollary 5.2. *Let K be a knot S^3, and let $K(p)$ be the p-fold branched cyclic cover of S^3 branched over K. Then $|K(p)|=|K^\tau(p)|$ for any mutation of K. The same is true for a $\mathbf{Z}_p \oplus \mathbf{Z}_q$ cover of a link if the pieces of the link which are glued up by the mutation have the same branch index.*

Proof. The preceding theorem proves this in the case that $K(p)$ or $K(p, q)$ is a hyperbolic manifold and p and q are odd. But a now-standard argument using the equivariant loop-theorem and Dehn's lemma shows that for $p \neq 2$

or $(p, q) \neq (2, 2)$, then the branched cover is atoroidal and irreducible. Since it contains the branched cover of the surface (which is incompressible) and is therefore Haken, Thurston's theorem implies that it is hyperbolic. For $p=2$ or $p=q=2$, it is well-known ([8] or [29]) that the 2-fold branched covers of mutants are actually homeomorphic, so of course they have the same Gromov norm.

For p even, but not equal to 2, we argue as follows (the argument works the same in the case of a link with one of p or q even): In doing the cutting and pasting along F which yields the exterior of the mutant knot, the meridian of K is preserved as a boundary curve of F. The process of taking a p-fold branched cover of K is the same as doing (generalized) Dehn surgery of type $(p, 0)$. With a choice of meridian and longitude, the function $f(m,n) =$ volume of result of (m, n)-Dehn surgery extends to an analytic function on a domain in the (extended) complex plane ([24, 30]). An analytic function on the extended plane is determined by its values on an infinite number of points which accumulate somewhere. Since the points $(p, 0)$ accumulate at infinity (which corresponds to the unsurgered manifold) knowing $f(p, 0)$ for all odd p determines it for all p. Since the meridians of K and K^τ correspond, the corresponding volume functions f and f^τ agree on all points $(p, 0)$ for p odd, and hence for all p.

The key to the result about Dehn surgery is the remark that cutting and pasting of surfaces that are related in a simple way produces homeomorphic manifolds.

Lemma 5.3. *Let F be a symmetric surface in a 3-manifold M, and let F' result from one of the following operations*:

(i) *A compression of F.*

(ii) *A boundary compression of F where the boundary curves of F which intersect the compressing disc are interchanged by τ.*

(iii) *Joining two boundary components of F which are interchanged by τ by a tube running along the boundary of M.*

Then M^τ is homeomorphic to M'^τ

Proof. These are all straightforward; the point is that any curve or arc along which a compression or boundary-compression takes place will be invariant under τ.

In order to discuss the relation between Dehn surgeries on a knot and on its mutant, one needs to specify a manner of comparing the surgery coefficients, or, equivalently, to specify the meridians and longitudes of the knots. To do this note that a Conway sphere has boundary which is a meridian; hence the mutation preserves boundary. The longitude is defined homologically, and so is specified automatically. Our approach to the question of Dehn surgery is to tube together the boundary components of the 4-punctured sphere so as not to worry about how cutting and pasting affects the structure of the cusp; in order to do this, the sphere must be in a nice position with respect to the knot. Observe that a specific choice of τ gives a pair of S^0's on the knot; each S^0 is preserved by τ.

Definition 5.4. *The Conway sphere S and mutation τ are unlined if these S^0's are unlinked on K.*

Theorem 5.5. *Let K be a knot or link in S^3 admitting a Conway sphere S. If K is link, suppose τ takes the components of K back to themselves. If K is a knot, suppose τ and S are unlinked. Then corresponding Dehn surgeries on K and K^+ give manifolds with the same Gromov norm.*

Proof. The hypothesis on τ and S enables one to tube together the boundary components which are interchanged by τ. The surface which results has genus two, and has an involution which is the same as τ as in Fig. 2. (If S and τ had been linked the resulting involution would be different from τ.) Now do the Dehn surgery; the surgery on the mutated knot is given by cutting and pasting along this closed surface. If the surface remains incompressible, theorem 1.4 applies directly and the volumes are the same. If the surface is compressible it compresses down to either a sphere or an incompressible punctured torus. Cutting and pasting along an incompressible torus doesn't change Gromov's norm [13, 30], and cutting and pasting along a sphere doesn't even change the manifold.

6. Degree-one maps between 3-manifolds

Our original interest in Gromov's norm arose in a study of maps between 3-manifolds with degree equal to one. The connection is the well-known fact that if $f: M \to N$ has degree 1, then $|M| \geqq |N|$. The question has some interest for knot complements, as there are a variety of homological invariants (such as the Alexander polynomial) which can provide obstruction to the existence of such a map. Mutant knots have the same abelian invariants [8], so that these would not help in deciding the existence of a degree-one map from X, the complement of a knot, to X^τ, the complement of a mutant knot. Our theorem 1.3, which says that the complements have the same volume, seems to indicate that volume likewise does not provide an obstruction. Surprisingly, the opposite turns out to be the case; the fact that the volumes are the same leads to an obstruction to the existence of a degree-one map.

Theorem 6.1. *Let K be a hyperbolic knot with a Conway sphere S, and τ a mutation so that S and τ are unlinked. If $f: X \to X^\tau$ has degree one, then $K = K^\tau$ (up to orientation.)*

Proof. By theorem 1.3, X^τ is hyperbolic, and $\mathrm{vol}(X^\tau) = \mathrm{vol}(X)$. Thurston [30] has proved a generalization of the Mostow rigidity theorem which says that a degree-one map between hyperbolic manifolds of the same volume is homotopic to an isometry (which we will continue to denote by f.) The only problem in using the homeomorphism f to show that $K = K^\tau$ is that f may not respect meridians-it automatically respects longitudes, up to sign, for homological reasons.

Let m, l (respectively m^τ, l^τ) be the meridians and longitudes of K (resp. K^τ); then

$$f_*(l) = \pm l^\tau, \quad f_*(m) = \pm m^\tau + a l^\tau$$

It follows that $\mathrm{vol}(X^{\tau}_{p,q}) = \mathrm{vol}(X^{\tau}_{\pm p, a\pm q})$. On the other hand, by theorem 5.5, $\mathrm{vol}(X_{p,q}) = \mathrm{vol}(X^{\tau}_{p,q})$; setting $f(p, q) = \mathrm{vol}(X_{p,q})$, we get the equation $f(p, q) = f(\pm p, a \pm q)$.

But Neumann and Zagier [24] have shown that for sufficiently large and q, $f(p, q)$ increases monotonically with $p^2 + q^2$. It follows easily that $a = 0$ in the above equation; i.e. that f takes meridians to meridians and longitudes to longitudes, up to sign, and hence that $K = K^{\tau}$ up to orientation.

Remark 6.2. The work of Bonahon (see [6], for example) shows that if a knot K and its mutant coincide, then one of the 'tangles' into which K is split must be symmetric in an appropriate sense, so that $K = K^{\tau}$ for 'obvious reasons'. The author and T. Cochran (in preparation) have defined an invariant of tangles which may be used to show that a tangle does not have such a symmetry, thus providing many examples where a knot and its mutant differ.

References

1. Adams, C.: Thrice-punctured spheres in hyperbolic 3-manifolds. Trans. Am. Math. Soc. **287**, 645–656 (1985)
2. Anderson, M.: Complete minimal hypersurfaces in hyperbolic n-manifolds. Comment. Math. Helv. **58**, 264–290 (1983)
3. Aronszajn, N.: A unique continuation theorem for solutions of elliptic partial differential equations or inequalities of the second order. J. Math. Pures Appl. IX. Ser. **36**, 235–249 (1957)
4. Bers, L.: Quasi-conformal mappings and Teichmüller's theorem. In: *Analytic function theory*, pp. 89–119. Princeton: Princeton University Press (1960)
5. Bonahon, F.: Bouts des variétés hyperboliques de dimension 3. Ann. Math. **124**, 71–158 (1986)
6. Bonahon, F., Siebenmann, L.: *Geometric splittings of classical knots, and the algebraic knots of conway.* Lond. Math. Soc. Lecture Notes, Cambridge University Press, Cambridge 1987
7. Conway J.: An enumeration of knots and links and some of their algebraic properties. In: *Computational problems in abstract algebra.* New York: Pergamon Press 1970
8. Cooper, D.: Thesis, Warwick University 1982
9. Culler, M.: Lifting representations to covering groups. Adv. Math **59**, 64–70 (1986)
10. Culler, M., Shalen, P.: Varieties of group representation and splittings of 3-manifolds. Ann. Math. **117**, 109–146 (1983)
11. Freedman, M., Hass, J., Scott, P.: Least area incompressible surfaces in 3-manifolds. Invent. Math. **71**, 609–642 (1983)
12. Freyd, P., Yetter, D., Hoste, J., Lickorish, W.B.R., Millett, K., Ocneanu, A.: A new polynomial invariant of knots and links. Bull. Am. Math. Soc. **12**, 239–246 (1985)
13. Gromov, M.: Volume and bounded cohomology. Publ. Math., Inst. Hautes Etud. Sci. **56**, 5–100 (1982)
14. Gulliver, R.: Regularity of minimizing surfaces of prescribed mean curvature. Ann. Math. **97**, 275–305 (1973)
15. Hass, J., Scott, P.: The existence of least area surfaces. MSRI Preprint, 1985
16. Jaco, W., Rubinstein, H.: Lectures on *PL* minimal surfaces. MSRI (1985)
17. Jaco, W., Shalen, P.: Seifert fibered spaces in 3-manifolds. Mem. Am. Math. Soc. **220**, (1979)
18. Johannson, K.: Homotopy equivalences of 3-manifolds with boundary. Lecture Notes in Math., vol. 761. Berlin-Heidelberg-New York: 1979
19. Kra, I.: On lifting Kleinian groups to $SL_2(\mathbf{C})$. In: *Differential geometry and complex analysis*, pp. 181–194, Chavel, I., Farkas, H. (eds.). Berlin-Heidelberg-New York-Tokyo: Springer 1985
20. Meeks, W., Scott, P.: Finite group actions on 3-manifolds. Invent. Math. **86**, 287–346 (1986)
21. Meeks, W., Scott, P., Yau, S.T.: The classical Plateau problem and the topology of three-dimensional manifolds. Topology **21**, 409–440 (1982)
22. Montesinos, J.M.: Revêtements ramifies de noeuds, espaces de Seifert et scindements de Heegard. Orsay preprint 1980

23. Morrey, C.B.: The problem of Plateau in a Riemannian manifold. Ann. Math. **49**, 807–851 (1948)
24. Neumann, W., Zagier, D.: Volumes of hyperbolic manifolds. Preprint 1983
25. Sacks, J., Uhlenbeck, K.: The existence of minimal immersions of 2-spheres. Ann. Math. **113**, 1–24 (1981)
26. Schoen, R., Yau, S.T.: Existence of incompressible minimal surfaces and the topology of 3-dimensional manifolds with non-negative scalar curvature. Ann. Math. **110**, 127–142 (1979)
27. Scott, P.: Subgroups of surface groups are almost geometric. J. Lond. Math. Soc., II. Ser. **17**, 555–565 (1978)
28. Selberg, A.: On discontinuous groups acting in higher dimensional symmetric space. In: *Contributions to function theory*, pp. 147–164. Tata Institute, Bombay 1960
29. Siebenmann, L.: On the vanishing of the Rochlin invariant and non-finitely amphicheiral homology 3-spheres. In: Lect. Notes Math. vol. 788, pp. 223–244. Berlin-Heidelberg-New York: 1980
30. Thurston, W.: Geometry and topology of 3-manifolds. Princeton: Princeton University Press 1978
31. Waldhausen, F.: On irreducible 3-manifolds which are sufficiently large. Ann. Math. **87**, 56–88 (1968)

Oblatum 16–III–1986 & 18–III–1987

Note added in proof

The author and R. Meyerhoff have recently shown that mutation does not change the Chern-Simons invariant, and determined the effect of mutations on the η-invariant. These results are in our preprint "Cutting and pasting and the η-invariant."