

THE COMPUTATIONAL COMPLEXITY OF KNOT GENUS AND SPANNING AREA

IAN AGOL, JOEL HASS, AND WILLIAM THURSTON

ABSTRACT. We show that the problem of deciding whether a polygonal knot in a closed three-dimensional manifold bounds a surface of genus at most g is **NP**-complete. We also show that the problem of deciding whether a curve in a PL manifold bounds a surface of area less than a given constant C is **NP**-hard.

1. INTRODUCTION

In this paper we investigate the computational complexity of some problems in three-dimensional topology and geometry. We show that the problem of determining a bound on the genus of a knot in a 3-manifold is **NP**-complete. Using similar ideas, we show that deciding whether a curve in a metrized PL 3-manifold bounds a surface of area less than a given constant C is **NP**-hard.

Determining whether a given knot is trivial or not is one of the historically central questions in topology. The problem of finding an algorithm to determine knot triviality was posed by Dehn [1]. Dehn's investigations into this area led to the formulation of the word and isomorphism problems, which played an important role in the development of the theory of algorithms. The first algorithm for the unknotting problem was given by Haken [3]. Haken's procedure is based on normal surface theory, a method of representing surfaces introduced by Kneser [14]. Analysis of the computational complexity of this algorithm is more recent. Hass, Lagarias and Pippenger showed that Haken's unknotting algorithm runs in time at most c^t , where the knot K is embedded in the 1-skeleton of a triangulated manifold M with t tetrahedra, and c is a constant independent of M or K [5]. It was also shown in [5] that the unknotting problem is in **NP**.

The notion of genus was defined by Seifert [23] in 1935 for knots in the 3-sphere, and extends directly to knots in an arbitrary 3-manifold M . Given a knot K , consider the class $\mathcal{S}(K)$ of all orientable spanning surfaces for K . These are surfaces embedded in M with a single boundary component that coincides with K . Seifert showed that this class is non-empty for any knot K in the 3-sphere. For knots in a general manifold, $\mathcal{S}(K)$ is non-empty when K represents a trivial element in the

Received by the editors July 17, 2002 and, in revised form, May 28, 2004.

2000 *Mathematics Subject Classification.* Primary 11Y16, 57M50; Secondary 57M25.

Key words and phrases. Computational topology, complexity, knot, 3-manifold, NP-complete, normal surface, genus.

The first author was partially supported by ARC grant 420998. This work was carried out while the second author was visiting the Institute for Advanced Study, and was partially supported by NSF grant DMS-0072348, and by a grant to the Institute for Advanced Study by AMIAS. The third author was partially supported by NSF grant DMS-9704286.

©2005 Ian Agol, Joel Hass, and William Thurston

first integer homology group of M . The genus $g(K)$ of a knot K is the minimum genus of a surface in $\mathcal{S}(K)$, or ∞ if $\mathcal{S}(K) = \emptyset$. The genus measures one aspect of the degree of “knottedness” of a curve.

The unknotting problem is a special case of the more general problem of determining the genus of a knot in a 3-manifold. Given a knot K in an orientable 3-manifold and a positive integer g , this problem asks for a procedure to determine whether the *knot genus* of K , the minimal genus of an orientable spanning surface for K in a 3-dimensional manifold, is at most g . A knot is trivial, or unknotted, precisely when its genus is zero. We will show that the problem of determining the genus of a knot in a 3-manifold is **NP**-complete. Previous results on this problem were given in [5], where it was shown to lie in **PSPACE**, roughly the class of problems that run in polynomial space. No lower bounds on the running time were previously known.

We work with 3-manifolds that are triangulated and orientable, and with orientable embedded surfaces. This is not a significant restriction, since all compact 3-dimensional manifolds admit unique PL structures [15]. A *knot* in a triangulated 3-manifold M is a connected simple (non-self-intersecting) closed curve in the 1-skeleton of M . Any smooth knot in a smooth manifold, or more generally any tame knot, is equivalent to a knot that lies in the 1-skeleton of some triangulation.

We formulate the problem of computing the genus as a language-recognition problem in the usual way; see [2]. In 1961 Schubert [21], in an extension of Haken’s work, showed the decidability of the problem:

Problem. 3-MANIFOLD KNOT GENUS

INSTANCE: A triangulated 3-dimensional manifold M , a knot K in the 1-skeleton of M , and a natural number g .

QUESTION: Does the knot K have $g(K) \leq g$?

The size of an instance is measured by the sum of the number of tetrahedra in M . In Section 3 we establish

Theorem 1. 3-MANIFOLD KNOT GENUS *is NP-hard*.

It was established in [5] that 3-MANIFOLD KNOT GENUS is in **PSPACE**. We improve this bound in Section 5 .

Theorem 2. 3-MANIFOLD KNOT GENUS *is NP*.

In combination these two results give:

Theorem 3. 3-MANIFOLD KNOT GENUS *is NP-complete*.

Theorem 1 is proved through a connection to ONE-IN-THREE SAT, a known **NP**-complete problem that will be reviewed in Section 3. The theorem carries out a construction that transforms an instance of ONE-IN-THREE SAT to an instance of 3-MANIFOLD KNOT GENUS. By “transform” we mean that an instance of one problem is changed to an instance of the second by a procedure that requires time polynomial in the size of the instance. To a boolean expression representing an instance of ONE-IN-THREE SAT we associate a positive integer g and a certain knot in a triangulated, compact 3-manifold. This knot bounds a surface of genus at most g exactly when there is a truth assignment to the boolean expression satisfying the requirements of ONE-IN-THREE SAT. Since ONE-IN-THREE SAT is **NP**-hard, this establishes that 3-MANIFOLD KNOT GENUS is also **NP**-hard.

In Section 5 we prove Theorem 2, giving a certificate which demonstrates in polynomial time that a genus g knot K bounds a surface of genus at most g . The argument in [5] established that the unknotting problem is **NP** using the existence of a normal disk that lies along an extremal ray in the space of normal solutions, called a vertex surface in Jaco-Tollefson [11]. The existence of such an extremal normal surface of minimal genus spanning a knot is not known, so a new technique is needed (See [10] for known results here). This is provided in Theorem 12, which gives an algorithm to count the number of orbits of a type of pseudogroup action on a set. Theorem 12 seems likely to have more general applicability. In Section 4 we describe this algorithm, and in Section 5 we apply it to a pseudogroup action that arises in the theory of normal surfaces. This allows us to determine in polynomial time the number of components in a normal surface described by an integer vector in \mathbb{Z}_+^{7t} . In particular we are able to certify that a normal surface is connected, orientable and has connected boundary. Since calculating the Euler characteristic of a normal surface can be done efficiently, establishing orientability and connectedness are the key steps in constructing a certificate of its genus.

In Section 6 we extend the orbit-counting algorithm to allow the counting of additional integer weight sums associated to each orbit. This allows for the polynomial time calculation of the genus of all the components of a fundamental normal surface, as well as a count of the number of components.

The genus and the area of a surface are closely connected. In Section 7 we extend the methods developed in studying genus to study the problem of determining the smallest area of a spanning surface for a curve in a 3-manifold. We show that computing an upper bound on the area of a smallest area spanning surface is **NP**-hard.

We refer to [16] for a discussion of complexity classes such as **NP** and **PSPACE** and to [25] for a discussion of complexity problems in low-dimensional topology.

Remarks. (1) Knots are often studied in \mathbb{R}^3 or S^3 rather than in a general manifold. Our methods show that determining knot genus in \mathbb{R}^3 or S^3 , or any fixed manifold, is **NP**. It is not clear whether the corresponding problem remains **NP**-hard if one restricts consideration to knots in \mathbb{R}^3 or S^3 .

(2) Casson has shown that a procedure to determine whether a 3-manifold is homeomorphic to the 3-sphere, following the 3-sphere recognition algorithm described in [18] and [24], runs in time less than $3^t p(t)$, where $p(t)$ is a polynomial. In the direction of lower bounds, it was shown in [13] that determining certain values of the Jones polynomial of alternating links is **#P**-hard.

We are grateful to the referee for numerous suggestions on the exposition of this paper.

2. NORMAL SURFACES

General surfaces in 3-manifolds can wind and twist around the manifold in complicated ways. Kneser described a procedure in which surfaces can be “pulled taut”, until they take a simple and rigid position [14]. In this *normal* position, they have very succinct algebraic descriptions. We use an approach to normal surfaces in triangulated 3-manifolds based on work of Jaco-Rubinstein [12] and Jaco-Tollefson [11]. A *normal surface* S in a triangulated compact 3-manifold M is a *PL*-surface whose intersection with each tetrahedron in M consists of a finite number of disjoint

elementary disks. These are properly embedded disks that are isotopic to either triangles or quadrilaterals as shown in Figure 1, by an isotopy preserving each face of the tetrahedron.

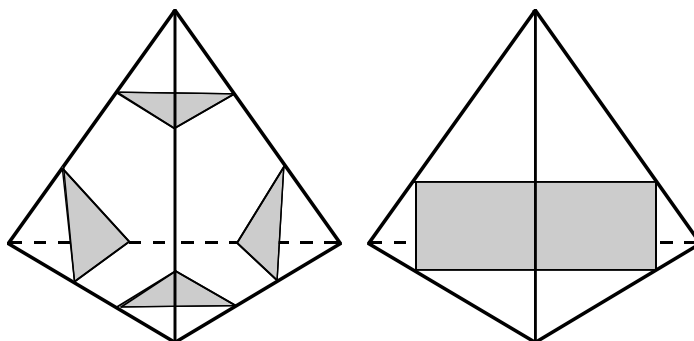


FIGURE 1. Elementary disks in a normal surface.

Within each tetrahedron of M there are four possible triangles and three possible quadrilaterals, up to a *normal isotopy of M* , an isotopy which leaves each cell of the triangulation of M invariant. W. Haken observed that a normal surface is determined up to such isotopies by the number of pieces of each of the seven kinds of elementary disks that occur in each tetrahedron, or a vector in \mathbb{Z}_+^{7t} . A normal surface S is described by a non-negative integer vector $\mathbf{v} = \mathbf{v}(S) \in \mathbb{Z}^{7t}$ that gives the *normal coordinates* of S . There is a homogeneous rational cone \mathcal{C}_M in \mathbb{R}^{7t} , called the *Haken normal cone*, that contains the vectors $\mathbf{v}(S)$ for all normal surfaces S in M .

If $\mathbf{v} = (v_1, v_2, \dots, v_{7t}) \in \mathbb{R}^{7t}$, then the Haken normal cone is specified by linear equations and inequalities of the form

$$\begin{aligned} v_{i_1} + v_{i_2} &= v_{i_3} + v_{i_4} \quad (\text{up to } 6t \text{ equations}), \\ v_i &\geq 0, \quad \text{for } 1 \leq i \leq 7t. \end{aligned}$$

The first set of equations expresses *matching conditions*, which say that the number of edges on a common triangular face of two adjacent tetrahedra, coming from a collection of elementary disks in each of the tetrahedra, must match. For each triangular face there are three types of edges (specified by a pair of edges on the triangle), which yield 3 matching conditions per face. Triangular faces in the boundary ∂M give no matching equations. The second set are called the *positivity conditions*. The cone \mathcal{C}_M is *rational* because the above equations have integer coefficients. We let $\mathcal{C}_M(\mathbb{Z}) = \mathcal{C}_M \cap \mathbb{Z}^{7t}$ denote the set of integral vectors in the cone \mathcal{C}_M . An additional set of conditions, the *quadrilateral conditions*, is required for an integral vector in the cone to correspond to the normal coordinates of an embedded surface. This condition states that of the three types of quadrilateral found in each tetrahedron, only one can occur in the vector with non-zero coefficient. The quadrilateral conditions are required because two distinct types of quadrilateral in a single tetrahedron necessarily intersect, and we are interested in embedded, non-self-intersecting surfaces. A vector in the Haken normal cone that satisfies the quadrilateral conditions corresponds to an embedded normal surface. This surface is unique up to a normal isotopy. However it is important to note that the surface corresponding to a given

vector in the cone \mathcal{C}_M may not be connected. A normal surface meets each edge of a triangulation in a finite number of points. The sum of these intersection numbers over all the edges of a triangulation is called the *weight* of the normal surface.

A *fundamental normal surface* is a normal surface S such that

$$\mathbf{v}(S) \neq \mathbf{v}_1 + \mathbf{v}_2, \quad \text{with } \mathbf{v}_1, \mathbf{v}_2 \in \mathcal{C}_M(\mathbb{Z}) \setminus \{\mathbf{0}\}.$$

In the terminology of integer programming, such a vector $\mathbf{v}(S)$ is an element of the *minimal Hilbert basis* $\mathbb{H}(\mathcal{C}_M)$ of \mathcal{C}_M ; see Schrijver [20, Theorem 16.4]. A fundamental normal surface is always connected, but connected normal surfaces need not be fundamental. A *vertex minimal solution* is a special kind of fundamental surface, one that corresponds to a solution of the normal surface equations that lies along an extremal ray of the cone of solutions and is not a multiple of another such extremal solution. Hass, Lagarias and Pippenger [5, Lemma 6.1] gave a bound for the size of the vectors corresponding to any fundamental surface.

Theorem 4. *Let M be a triangulated compact 3-manifold, possibly with boundary, that contains t tetrahedra.*

- *Any vertex minimal solution $\mathbf{v} \in \mathbb{Z}^{7t}$ of the Haken normal cone \mathcal{C}_M in \mathbb{R}^{7t} has $\max_{1 \leq i \leq 7t} (v_i) \leq 2^{7t-1}$.*
- *Any minimal Hilbert basis element $\mathbf{v} \in \mathbb{Z}^{7t}$ of the Haken fundamental cone \mathcal{C}_M has $\max_{1 \leq i \leq 7t} (v_i) < t \cdot 2^{7t+2}$.*

Schubert [21] showed that a surface of smallest genus spanning K can be found among the fundamental surfaces.

Theorem 5. *There is a minimal genus spanning surface for K which is a fundamental normal surface.*

Similar to the theory of normal surfaces, though somewhat easier, is the theory of normal curves. These are curves on a surface that intersect each triangle in a collection of *normal arcs*, arcs that have endpoints on distinct edges. Normal curves arise as the boundaries of normal surfaces in a manifold with boundary. Since there are three such arcs in each triangle, normal isotopy classes of normal curves in a surface that contains t triangles are described by integer vectors in \mathbb{Z}_+^{3t} .

3. 3-MANIFOLD KNOT GENUS IS NP-HARD

In this section we show how to reduce an instance of ONE-IN-THREE SAT to an instance of 3-MANIFOLD KNOT GENUS. Since ONE-IN-THREE SAT is known to be NP-hard, this establishes that 3-MANIFOLD KNOT GENUS is also NP-hard. The problem ONE-IN-THREE SAT concerns logical expressions involving collections of literals (boolean variables or their negations) gathered in clauses consisting of three literals connected with \vee 's. The logical expression contains a collection of clauses connected with \wedge 's.

Problem. ONE-IN-THREE SAT

INSTANCE: A set U of variables and a collection C of clauses over U such that each clause $c \in C$ contains 3 literals.

QUESTION: Is there a truth assignment for U such that each clause in C has exactly one true literal?

Schaefer [19] established that ONE-IN-THREE SAT is **NP**-complete. To prove Theorem 1, establishing that 3-MANIFOLD KNOT GENUS is **NP**-hard, we show that an arbitrary problem in ONE-IN-THREE SAT can be reduced in polynomial time to a problem in 3-MANIFOLD KNOT GENUS. See Garey and Johnson [2] for a discussion and many examples of such reductions.

Let $U = \{u_1, u_2, \dots, u_n\}$ be a set of variables and $C = \{c_1, c_2, \dots, c_m\}$ be a set of clauses in an arbitrary instance of ONE-IN-THREE SAT. We will describe a knot K in a compact 3-dimensional manifold (with no boundary) and an integer g such that K bounds a surface of genus smaller than or equal to g if and only if C is satisfiable so that each clause in C contains exactly one true literal. We construct the 3-manifold M in stages. First we construct a 2-dimensional simplicial complex, then we thicken this complex, replacing triangles with subdivided triangular prisms, getting a triangulated, 3-dimensional manifold with boundary, as indicated in Figure 2. Finally we use a doubling construction, taking two copies of the manifold with boundary and gluing their boundaries together, to obtain a closed 3-manifold.

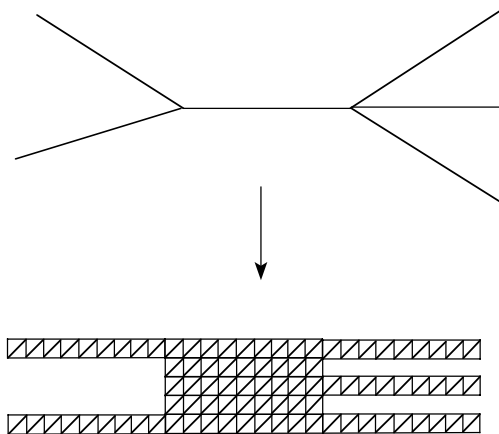


FIGURE 2. A branching surface B , shown in cross-section, is “thickened” to produce a triangulated 3-manifold with boundary.

To begin, we form a type of singular surface B , that we call a *branching surface*, by identifying boundary curves, of a collection of $2n + 1$ surfaces with boundary, each forming what we refer to as a *piece* of the branching surface. We construct this collection of surfaces as follows. Let k_i be the number of times that the variable u_i appears in the collection of clauses, and \bar{k}_i be the number of times that the negation \bar{u}_i of u_i appears. For $i = 1, \dots, n$, let F_{u_i} and $F_{\bar{u}_i}$ be genus one surfaces with $k_i + 1$ and $\bar{k}_i + 1$ boundary curves, respectively. Also set F_0 to be a planar surface with $n + m + 1$ boundary curves. One of these boundary curves will later become the knot K . The branching surface B is constructed by identifying these surfaces along appropriate boundary components as indicated in Figure 3.

Branching occurs when more than two boundary curves are identified along a single curve. We identify pairs of boundary curves by giving a homeomorphism between them. Up to isotopy, this is determined by specifying an orientation on the curves and setting the homeomorphism to be orientation reversing. We first fix an orientation on each of F_0, F_{u_i} and $F_{\bar{u}_i}$. This induces an orientation on

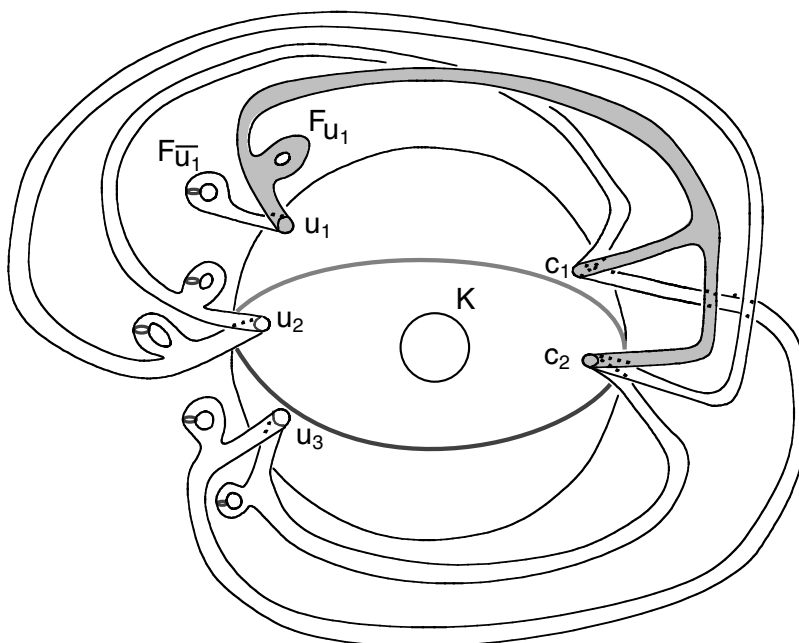


FIGURE 3. A branching surface with boundary curve K corresponding to the boolean expression $(u_1 \vee u_2 \vee u_3) \wedge (u_1 \vee \bar{u}_2 \vee \bar{u}_3)$. The shaded surface F_{u_1} indicates the occurrence of u_1 in each of the clauses c_1 and c_2 .

each boundary curve. All identifications will involve gluing a boundary component of F_{u_i} or $F_{\bar{u}_i}$ to a boundary component of F_0 , and we require this gluing to be an orientation-reversing homeomorphism. Label by $K, u_1, \dots, u_n, c_1, \dots, c_m$ the $1 + n + m$ boundary components of F_0 . The boundary component K of F_0 , which will become our knot, has nothing identified to it. For each $i, 1 \leq i \leq n$, one boundary curve from the surface F_{u_i} is identified to u_i . The remaining k_i boundary components of F_{u_i} are identified with k_i of the curves c_1, \dots, c_m on ∂F_0 , with one component of ∂F_{u_i} identified with c_j for each occurrence of the literal u_i in the j^{th} clause in C . Similarly, one curve of $\partial F_{\bar{u}_i}$ is identified to the component of F_0 labeled u_i , and the remaining \bar{k}_i boundary components are glued to c_1, \dots, c_m , with a component glued to c_j for each occurrence of the literal \bar{u}_i in the j^{th} clause of C . A total of three surface boundaries are identified along each of u_1, \dots, u_n , and exactly four surface boundaries are identified along each of c_1, \dots, c_m , as in Figure 3.

Lemma 6. *There is a truth assignment for U such that each clause in C has exactly one true literal if and only if there is a surface S with connected boundary and genus at most $m + n$ and a continuous map $f : S \rightarrow B$ such that $f|_{\partial S}$ is a homeomorphism onto K .*

Proof. Suppose there is a truth assignment for U such that each clause in C contains exactly one true literal. Form a surface S inside B by taking the union of F_0 and either F_{u_i} if u_i is true, or $F_{\bar{u}_i}$ if u_i is false. Then exactly two boundary components will be identified along each of the boundary components of F_0 other than K itself,

and K becomes the boundary of the resulting embedded surface S . There is a contribution of one to the genus from each of the literals, since F_{u_i} and $F_{\bar{u}_i}$ each have genus one, and a contribution of one to the genus from each handle formed when a boundary component of a surface F_{u_i} or $F_{\bar{u}_i}$ is glued to F_0 along a curve c_j . The genus of S is therefore equal to $m + n$.

Now suppose there is a surface S_1 of genus $\leq n + m$ mapped continuously into B that has a single boundary component mapped homeomorphically to the boundary curve K . We will show in this case that there is a surface S_4 with the same boundary, consisting of certain pieces of B identified along their boundaries, having genus precisely $n + m$, and containing, along each curve c_j , exactly one of the three pieces of surface joining F_0 .

The map f_1 of S_1 into B may be quite complicated, winding back and forth across B , but by standard transversality arguments we can homotop f_1 so that it is a union of homeomorphisms of subsurfaces of S_1 mapped homeomorphically to one of the pieces $F_0, F_{u_i}, F_{\bar{u}_i}$ forming B . More precisely, we can perturb S_1 by a small homotopy, so that its intersection with the boundary components of F_0 is transverse and pulls back to a collection of simple closed curves on S_1 . If any of these curves bounds a disk in S_1 , then the disk is mapped into some subsurface X of B while the boundary curve is mapped to ∂X . Since X is not itself a disk, the disk can be homotoped into ∂X , and we can therefore homotop S_1 in a neighborhood of this disk to remove a component of $S_1 \cap f^{-1}(\partial F_0)$. After repeating finitely many times, each component of the complement of $S_1 \cap f^{-1}(\partial F_0)$ in S_1 has non-positive Euler characteristic. The image of S_1 in each piece $F_0, F_{u_i}, F_{\bar{u}_i}$ has an algebraic degree, which is either even or odd. This degree equals the number of pre-images in S_1 of a generic point in the piece. The degree of the map from S_1 on F_0 is odd, since K is the boundary of S_1 and therefore S_1 maps an odd number of times to points near K . The sum of the degrees along each of the pieces meeting a curve u_i or c_j is even, since S_1 has no boundary along these curves. In particular, for each $1 \leq i \leq n$, exactly one of $F_{u_i}, F_{\bar{u}_i}$ has odd multiplicity in S_1 . Form a new surface S_2 by taking the union of F_0 and each of the pieces $F_{u_i}, F_{\bar{u}_i}$ which have odd multiplicity in S_1 . We will show that $\chi(S_2) > \chi(S_1)$. The surface S_2 is obtained from S_1 by a series of operations that either discard a subsurface with non-positive Euler characteristic or replace a subsurface of S_1 that maps with odd degree to some piece F_{u_j} with a subsurface mapping homeomorphically to F_{u_j} . The Euler characteristic of a discarded subsurface of S_1 is smaller than or equal to that of the subsurface that replaces it, so $\chi(S_2) \geq \chi(S_1)$ in either case.

The collection of pieces among $\{F_{u_i}, F_{\bar{u}_i}\}$ that are in the image of S_2 can be attached to F_0 along common boundary curves among $\{u_1, \dots, u_n\}$, forming a connected surface S_3 of genus n . Finally, a connected surface S_4 with boundary K is obtained by identifying pairs of curves in ∂S_3 that are mapped to the boundary components c_1, \dots, c_m . Each c_i then has either two or four curves in ∂S_3 mapped to it, so there are either one or two identifications made along each c_i . The choice of which pairs to identify, in case there are four curves mapped to c_i , is not important. The Euler characteristic of S_4 is the same as that of S_3 and S_2 , and therefore greater than or equal to that of S_1 . So the genus of S_4 is at most that of S_1 , $\text{genus}(S_4) \leq n + m$. Each identification of a pair of surface boundaries along c_1, \dots, c_m contributes one to the genus of S_4 . There is at least one such identification along each of the m curves $\{c_j\}$, though there may be two if each of the

three surfaces meeting F_0 along c_j has multiplicity one in S_2 . So identification of curves along c_1, \dots, c_m adds at least m to the genus of S_4 , and it follows that $\text{genus}(S_4) \geq n + m$. Since we have seen that $\text{genus}(S_4) \leq n + m$, equality must hold. Equality holds when exactly one of the three surface pieces meeting F_0 along c_j has odd multiplicity for each $1 \leq j \leq m$. We then assign the value “TRUE” to a literal u_i if F_{u_i} is used in S_4 , and the value “FALSE” to u_i if $F_{\bar{u}_i}$ is used in S_4 . This gives a truth assignment to U in which each clause in C has exactly one true literal. \square

To show that 3-MANIFOLD KNOT GENUS is **NP**-hard, we reduce in polynomial time an instance of the **NP**-hard problem ONE-IN-THREE SAT to an instance of 3-MANIFOLD KNOT GENUS.

Proof of Theorem 1. Given an instance of ONE-IN-THREE SAT, form B as in Lemma 6. Then there is a truth assignment for U such that each clause in C has exactly one true literal if and only if K is the boundary of a surface of genus $m + n$ mapped continuously into B . Form a 3-manifold N by thickening B so that it is embedded inside a triangulated 3-dimensional manifold with boundary. The thickening process replaces each subsurface forming B by a product of a surface with an interval, and then glues these surfaces together along portions of their boundaries, as indicated in Figure 2. The curve K remains on the boundary of N , and there is a projection map $p : N \rightarrow B$ which fixes K . Form a closed manifold M by doubling N along its boundary, namely by taking two copies of N and identifying them along their boundaries by the identity map. Then M admits an involution τ that fixes K , and has quotient N .

To find a triangulation of N , we first describe an explicit triangulation of the branching surface B . An orientable surface with boundary has a triangulation with one vertex on each boundary component and no vertices in the interior. The number of triangles is $4g + 5c - 4$, where g is the genus and c is the number of boundary components. We choose such triangulations for each of the $2n + 1$ subsurfaces in the branching surface, and we match them together along boundary components to get a triangulation of B . So B has a triangulation in which the number of triangles is linearly bounded in $n + m$. We thicken the surfaces F_{u_i} and $F_{\bar{u}_i}$ by taking their product with an interval. We form a cell structure of the thickened surface by dividing the product into prisms, products of a triangle and an interval. For the thickening of F_0 , we start by doing the same and then go on to divide each interval into five subintervals. The top, middle and bottom subintervals are identified with the intervals from the three thickened surfaces meeting each thickened boundary component of F_0 corresponding to a curve c_i . Only the top and bottom intervals are identified with thickened surfaces from the other boundary components. We can therefore now glue these thickened surfaces together to get a 3-cell structure on N . Each cell in this structure is a prism. We form a closed 3-manifold M by doubling N along its boundary, gluing two copies of N together along their boundaries to obtain a 3-manifold with no boundary.

Finally, we stellar subdivide the cell structure to get a triangulation. Each prism is divided into 14 tetrahedra, by dividing each rectangular face into four triangles by coning to a vertex in the center of each such face, and then coning the 14 triangles of the boundary of the prism to a vertex added to its center. The number of resulting simplices in M is linearly bounded by $n + m$.

We now check that if K bounds a surface of genus $g \leq m + n$ with interior in $M \setminus K$, then it bounds a surface of the same genus in B . Suppose that F is an embedded surface in M with boundary K . If F does not already lie in N , then perturb it slightly so that the interior of F meets ∂N transversely in a finite number of simple closed curves and arcs. Using the involution τ , reflect the portion of F not in N into N , forming an immersed surface F' lying in N and with the same boundary as F . The interior of F' remains disjoint from K since K is fixed by the involution. The projection $p(F')$ is a surface of genus g mapped into B with boundary K . So if K bounds an embedded surface of genus at most $m + n$ in M , then it bounds a surface of genus at most $m + n$ mapped into B . The converse was shown in the proof of Lemma 6. It then follows from Lemma 6 that K bounds an embedded surface of genus at most $m + n$ in M if and only if C is satisfiable with each clause containing exactly one true literal.

The construction of B, N and M described above each requires only a linear number of steps in the size of the instance of ONE-IN-THREE SAT with which we started, so that the reduction requires polynomial time. \square

4. ORBITS OF INTERVAL ISOMETRIES

In this section we develop a combinatorial procedure that will allow us to count the components of a normal surface. The procedure computes the number of orbits of a collection of k isometries between subintervals of an interval $[1, N] \subset \mathbb{Z}$ in time polynomial in $k \log N$. By lining up the intersections of a normal surface with the edges of a triangulation, we obtain such subintervals. Arcs of the normal surface on the faces of the triangulation give rise to correspondences of these intersection points which are subinterval isometries. We can then apply the algorithm developed here to count the number of components of a normal surface.

Assume that we have a set of integers $\{1, 2, \dots, N\}$ and a collection of bijections, $g_i : [a_i, b_i] \rightarrow [c_i, d_i]$, $1 \leq i \leq k$, either increasing or decreasing, that are called *pairings*. If a pairing identifies two intervals $[a, b]$ and $[c, d]$ by sending a to c and b to d , we call it an *orientation-preserving* pairing, and if it sends a to d and b to c , we call it *orientation-reversing*. If $a < c$ we refer to $[a, b]$ as the *domain* and $[c, d]$ as the *range* of the pairing. We work only with integers, and use the term “connected interval” to refer to the integers in a connected real interval. The *width* of an interval $[a, b]$ with integer endpoints is $b - a + 1$, the number of integers it contains. The *width* of a pairing is the width of its domain or range, $w = b - a + 1 = d - c + 1$. If the pairing preserves orientation, its *translation distance* t measures how far it moves points, so $t = c - a = d - b$. We can compose two pairings if the range of the first lies in the domain or range of the second. The collection of pairings generates a pseudogroup under the operations of composition where defined, inverses, and restriction to subintervals.

The interval $[1, \dots, N]$ is divided into equivalence classes by the action of the pairings, which are called *orbits*. We are interested in the orbit structure of the collection of pairings, since with appropriate interpretation an orbit corresponds to a connected component of a normal surface. We introduce several simplification processes on the set of pairings in order to analyze the structure of the set of orbits.

We introduce some terminology to describe the behavior of pairings. An interval is called *static* if it is in neither the domain nor the range of any pairing, so that its points are identified to no other points by pairings. Given a collection of pairings

acting on the integers $[1, \dots, N]$, a pairing is said to be *maximal* if its range contains both N and the range of any other pairing containing N . More precisely, define a linear order on pairings using the lexicographical order $(d_i, -c_i, -a_i, -\text{orientation})$, so that the maximal pairing has the highest upper endpoint, and among those with that endpoint the widest range, and among those with that range the biggest translation distance (if orientable). Finally, we say that a pairing $g : [a, b] \rightarrow [c, d]$ is *periodic* with period t if it is orientation preserving with translation distance t and $a < c = a + t \leq b + 1$, so there is no gap between the domain and range. The combined interval $[a, d]$ is then called a *periodic interval* of period t .

The following lemma describes the orbits of a periodic pairing.

Lemma 7. *A periodic pairing has $t = c - a$ orbits on $[a, d]$.*

Proof. Each point in $[a, b]$ greater than or equal to c lies in the range of g and can be mapped to a smaller point in $[a, b]$ by a power of g^{-1} . Therefore, each orbit on $[a, d]$ has a representative in $[a, c - 1]$. Since the congruence class modulo t of a point is preserved by g , each of the t integers in $[a, c - 1]$ lies in a distinct orbit. These points uniquely represent the t orbits. \square

We now show how to merge two pairings with sufficient overlap into a single pairing with the same orbits.

Lemma 8. *Let R_1 be a periodic interval with pairing g_1 of period t_1 . Suppose that there exists an orientation-preserving pairing g_2 with translation distance t_2 and an interval $J_1 \subset R_1$ such that J_1 has width t_1 and $g_2(J_1) \subset R_1$. Then the orbits of $g_1 \cup g_2$ on R_1 are the same as those of a single periodic action on R_1 of period $\text{GCD}(t_1, t_2)$.*

Proof. Let J_1 be an interval in R_1 of width t_1 that is in the domain of g_2 and which is paired by g_2 to an interval in R_1 . Each point in R_1 has a unique orbit representative in J_1 under the action of g_1 on R_1 . The interval $g_2(J_1)$ lies in R_1 by assumption. For $x \in J_1$ let $f(x)$ be the unique point in J_1 obtained by carrying $g_2(x)$ back to a point in J_1 by a power of g_1 or g_1^{-1} . The effect of $f(x)$ on J_1 is a shift of $t_2 \pmod{t_1}$. The orbits of f on J_1 divide the points of J_1 into congruence classes modulo $\text{GCD}(t_1, t_2)$. Neither g_1 nor g_2 change the congruence class of a point mod $(\text{GCD}(t_1, t_2))$, so a subinterval of width $\text{GCD}(t_1, t_2)$ in R_1 contains exactly one representative of each orbit of R_1 under the action of $g_1 \cup g_2$. The same orbits arise from a periodic action of period $\text{GCD}(t_1, t_2)$. \square

The following is a special case of Lemma 8 that applies when both g_1 and g_2 are periodic pairings.

Lemma 9. *Let R_1 and R_2 be overlapping periodic intervals associated with pairings g_1, g_2 having periods t_1 and t_2 . Suppose that $\text{width}(R_1 \cap R_2) \geq t_1 + t_2$. Then the orbits of $g_1 \cup g_2$ on $R_1 \cup R_2$ are the same as those of a single periodic pairing on $R_1 \cup R_2$ of period $\text{GCD}(t_1, t_2)$.*

Proof. The leftmost interval J_1 of width t_1 in $R_1 \cap R_2$ is translated by g_2 a distance of t_2 to the right, to an interval J_2 which lies in $R_1 \cap R_2$. Lemma 8 then states that the action of $g_1 \cup g_2$ on R_1 is the same as those of a single periodic pairing on R_1 of period $\text{GCD}(t_1, t_2)$. By symmetry the same result holds for the action of $g_1 \cup g_2$ on R_2 . Thus the orbits of $g_1 \cup g_2$ on $R_1 \cup R_2$ are the same as those of a single periodic pairing on $R_1 \cup R_2$ of period $\text{GCD}(t_1, t_2)$. \square

We state a consequence in a form which will be convenient for our applications.

Lemma 10. *Let g_1, g_2 be periodic pairings with periods t_1, t_2 and let J_1, J, J_2 be intervals with $g_1(J_1) = J$ and $g_2(J) = J_2$. Suppose that J is contained in the union $J_1 \cup J_2$. Then the hypothesis of Lemma 9 is satisfied, and the orbits of $g_1 \cup g_2$ are the same as those of a single periodic pairing of period $\text{GCD}(t_1, t_2)$, acting on the union of the periodic intervals of g_1, g_2 .*

Proof. Let $J_1 = [a_1, b_1]$, $J = [a, b]$ and $J_2 = [a_2, b_2]$. We have $a_2 \geq a_1$ and $J \subset J_1 \cup J_2$. Since $J_1 \cup J_2$ is a connected interval, we have that $a_2 \leq b_1 + 1$. Note that $t_1 = b - b_1 = a - a_1$ and $t_2 = b_2 - b = a_2 - a$. Then $t_1 + t_2 = (b - b_1) + (a_2 - a) = (b - a + 1) + (a_2 - b_1 - 1)$. The width of J is $(b - a + 1)$, and we have seen above that $(a_2 - b_1 - 1) \leq 0$. Therefore, the width of J is at least $t_1 + t_2$, and Lemma 9 implies the conclusion of the lemma. \square

The orbit-counting algorithm applies a series of modifications to a collection of pairings. We now describe these modifications.

- Periodic merger

The *periodic merger* operation replaces g_1 and g_2 by a single periodic action on $R_1 \cup R_2$ of period $\text{GCD}(t_1, t_2)$, as in Lemma 9.

- Contraction

The operation of *contraction* is performed on a static interval $[r, s]$. We eliminate this interval, replace $[1, N]$ by $[1, N - (s - r + 1)]$, and alter each g_j by replacing any point x in a domain or range which lies entirely to the right of s by $x - (s - r + 1)$. (This operation will lead to a decrease of $s - r + 1$ in the number of orbits, since the eliminated points are each unique representatives of an orbit.)

- Trimming

The *trimming* operation simplifies an orientation-reversing pairing whose domain and range overlap. Suppose that $g : [a, b] \rightarrow [c, d]$ is a pairing with $g(a) = d, g(b) = c$ and $b \geq c$. Define a new pairing $g' : [a, (a + d)/2] \rightarrow ((a + d)/2, d]$ by restricting the domain and range of g , and say that g' is obtained from g by *trimming*. The domain and range of a trimmed pairing are disjoint.

- Truncation

If an interval lies in the domain and range of exactly one pairing, then the interval can be “peeled off” without changing the orbit structure, in a way we now describe. The algorithm applies this operation to strip off points from the right of the interval $[1, N]$.

When there is a pairing $g : [a, b] \rightarrow [c, N]$ and a value N' with $c \leq N' + 1 \leq N$, such that all points in the interval $[N' + 1, N]$ are in the range of no pairing other than g , then we can perform an operation called *truncation* of g . Truncation shortens the interval $[1, N]$ to the interval $[1, N']$, and similarly shortens the domain and range of g . If g is orientation preserving, pairings other than g are unchanged, while g is eliminated entirely if $c = N' + 1$, or replaced by a shortened pairing $g' : [a, b - (N - N')] \rightarrow [c, N']$ if $c \leq N'$. We can perform this operation even if the interval $[N' + 1, N]$ intersects both the range and the domain of g .

If g is orientation reversing, truncation is applied only when g has disjoint domain and range (i.e. after trimming), and $[N' + 1, N]$ is contained in the range. Suppose $g : [a, b] \rightarrow [c, N]$ is a pairing with $g(a) = N, g(b) = c$ and $b < c$, that $[N' + 1, N]$ is disjoint from the domains and ranges of all pairings other than g and that $c \leq N' + 1$. If $N' + 1 = c$, then we eliminate g . Otherwise replace g by a shortened orientation-reversing pairing, formed by restricting its domain to $[a + N - N', b]$ and its range to $[c, N']$.

- **Transmission**

Transmission is an operation in which two pairings are composed. A pairing g_1 is used to shift down the domain and range of a second pairing g_2 as much as possible. This operation will allow us to shift pairings leftwards from the right end of the interval $[1, N']$ and subsequently apply truncation.

If g_1 is orientation reversing and has overlapping domain and range, then as a first step in transmission we trim g_1 . Now consider a pairing g_1 , either orientation preserving or orientation reversing, and a second pairing g_2 whose range is contained in the range of g_1 . If the domain of g_2 is not contained in the range of g_1 , form the composite map $g'_2 = g_1^{-r} \circ g_2$, where $r = 1$ if g_1 is orientation reversing and otherwise $r \geq 1$ is the largest integer such that $g_1^{-r+1}([c_2, d_2])$ is contained in the range of g_1 . The domain of g'_2 is the same as that of g_2 in this case. If the domain of g_2 is also contained in the range of g_1 , then form the composite map $g'_2 = g_1^{-r} \circ g_2 \circ g_1^s : g_1^{-s}([a_2, b_2]) \rightarrow g_1^{-r}([c_2, d_2])$, where r is as above, $s = 1$ if g_1 is orientation reversing, and otherwise $s \geq 1$ is the largest integer such that $g_1^{-s+1}([a_2, b_2])$ is contained in the range of g_1 . The domain of g'_2 is then that of g_2 shifted left by g_1^{-s} . The process of replacing g_2 by $g_1^{-r} \circ g_2 \circ g_1^s$ is called a *transmission of g_2 by g_1* .

We now construct a sequence of pseudogroups of pairings, terminating with the trivial pseudogroup acting on the empty set. At each stage, each orbit is associated to a unique orbit in the previous set of pairings, though the number of orbits may decrease. A counter is kept at each stage that records the total decrease in the number of orbits. Since the final pseudogroup has no orbits, the final value of this counter gives the initial number of orbits.

Lemma 11. *A contraction decreases the number of orbits by the width of the contracted interval. Altering a collection of pairings by any of the operations of periodic merger, trimming, truncation and transmission preserves the number of orbits.*

Proof. We refer to the collection of pairings g_1, g_2, \dots, g_k as G , and to the new collection of pairings produced by one of the operations as G' .

The effect of a contraction is to shorten the interval $[1, N]$ by removing points which are fixed by the entire collection of pairings. The number of orbits removed equals the width of the contracted interval.

A periodic merger joins two periodic intervals R_1 and R_2 and their pairings into one. Lemma 9 shows that the orbits in $R_1 \cup R_2$ of $g_1 \cup g_2$ are the same as those produced by a single periodic pairing g' . Suppose that x and y are points in $[1, N]$ in the same orbit of the action of $G = \{g_1, g_2, \dots, g_k\}$. Then $h \cdot x = y$, where h is some finite word in the elements of G and their inverses. Wherever a g_1 or g_2 occurs in this word we can replace it by a power of g' , since g' translates by an amount

that divides the translation distance of g_1 and g_2 . So x and y are in the same orbit of $G' = \{g', g'_3, \dots, g'_k\}$. Conversely, suppose that x and y are in the same orbit under G' . Then $h' \cdot x = y$, where h' is some finite word in $g', g'_3, g'_4, \dots, g'_k$. Lemma 9 implies that wherever a g' occurs in h' we can replace it by some word in g_1 and g_2 , since the orbits of g' and $g_1 \cup g_2$ coincide on the periodic interval of g' . So periodic mergers preserve orbits.

Suppose next that $g_i : [a_i, b_i] \rightarrow [c_i, d_i]$ is an orientation-reversing pairing with $g_i(a_i) = d_i, g_i(b_i) = c_i \leq b_i$, and that $g'_i : [a_i, (a_i + d_i)/2] \rightarrow ((a_i + d_i)/2, d_i]$ is obtained by trimming g_i . If x and y are in the same orbit of $G = \{g_1, g_2, \dots, g_k\}$, then there is a sequence of points $x = x_1, x_2, \dots, x_r = y$, where each x_j is the image of x_{j-1} under a pairing in G . We can replace an occurrence of g_i or g_i^{-1} by g' if the point x_{j-1} is smaller than $(a_i + d_i)/2$ and by g'^{-1} otherwise. So x and y remain in the same orbit under the action of G' . Conversely, if x and y are in the same orbit under G' , then replacing each occurrence of g' by g and of g'^{-1} by g^{-1} gives a word in G taking x to y . We conclude that trimming preserves orbits.

Next consider the effect of a truncation. Suppose that all points in $[N' + 1, N]$ are in the range of exactly one pairing $g : [a, b] \rightarrow [c, N]$, and we truncate, shortening to an interval $[1, N']$ and either eliminating g or shortening it to $g' : [a, b - (N - N')] \rightarrow [c, N']$ in the orientation-preserving case, or to $g' : [a + N - N', b] \rightarrow [c, N']$ in the orientation-reversing case. Suppose that $y = h \cdot x$, where h is a reduced product of pairings (a product that does not contain a subproduct of the form $g_j \cdot g_j^{-1}$). We can assume that $y > x$, as otherwise we can replace h by its inverse. Let h' be obtained from h by replacing all occurrences of g with g' . The successive images of x under the subwords of h can never enter and leave the interval $[N' + 1, N]$, since they can only enter it under the action of some positive power of g , and only leave it under a negative power of g . These do not occur in succession in a reduced product. The image of a point z under g_j is unchanged unless $g_j = g$, and $g(z) \geq N' + 1$. So the image of x under h is the same as its image under h' unless $y > N'$. In that case h is of the form $h = g_i^r h_1$, where $h_1 \cdot x \leq N'$, $r \geq 1$, and $y = g_i^r h_1 \cdot x > N'$. So two points in $[1, N']$ are in the same orbit under G' if and only if they are in the same orbit under G . But each point in $[N' + 1, N]$ is in the same G orbit as a point in $[1, N']$. It follows that the number of orbits is unchanged by truncation.

Finally suppose that a transmission of g_j by g_i replaces the pairing g_j in G by $g'_j = g_i^{-r} \circ g_j \circ g_i^s$. If $y = h \cdot x$, where h is a word in G , then $y = h' \cdot x$, where h' is obtained by replacing every occurrence of g_j with $g_i^k \circ g_j \circ g_i^{-s}$. Similarly if $y = h' \cdot x$, where h' is a word in G' , then $y = h \cdot x$, where h is obtained by replacing every occurrence of g'_j with $g_i^{-k} \circ g_j \circ g_i^s$. So transmissions preserve the collection of orbits and do not change their number. \square

We now describe an algorithm which uses these operations to count the number of orbits of a pseudogroup of pairings acting on $[1, N]$. We initially set a counter for the number of orbits to zero. A pairing g_i is said to be *maximal* if $d_i = N$ and $[c_i, N]$ contains the range of any other pairing with an endpoint at N . Let $S = \{1, 2, \dots, N\}$ and let $g_i : [a_i, b_i] \rightarrow [c_i, d_i]$, $1 \leq i \leq k$, be a collection of pairings between subintervals of S . The algorithm will contract S and reduce the number of pairings. It keeps a running count of the number of orbits detected in an integer referred to as the orbit counter. Denote by N' the current size of the interval as we

proceed. The algorithm repeats the following steps, reducing N' and k , until there are no points remaining.

Orbit counting algorithm.

- (1) Delete any pairings that are restrictions of the identity.
- (2) Make any possible contractions and, if any exist, increment the orbit counter by the sum of the number of points deleted by the contractions. If the number of pairings remaining is zero, output the number of orbits and stop.
- (3) Trim all orientation-reversing pairings whose domain and range overlap.
- (4) Search for pairs of periodic pairings g_i and g_j whose domains and ranges satisfy the condition of Lemma 9. If any such pair exists, then perform a merger as in Lemma 9, replacing g_i and g_j by a single periodic pairing, with translation distance $\text{GCD}(t_i, t_j)$. The new pairing acts on the union of the domains and ranges of g_i and g_j . Repeat until no mergers can be performed.
- (5) Find a maximal g_i . For each $g_j \neq g_i$ whose range is contained in $[c_i, N']$, transmit g_j by g_i .
- (6) Find the smallest value of c such that the interval $[c, N']$ intersects the range of exactly one pairing. Truncate the pairing whose range contains the interval $[c, N']$.

Theorem 12. *The orbit-counting algorithm gives the number of orbits of the action of the pairings $\{g_i\}_{i=1}^k$ on $[1, \dots, N]$ in time bounded by a polynomial in $k \log N$.*

Proof. We first check that the orbit-counting algorithm correctly counts the number of orbits. In step (1), deleting a pairing which is the identity on its domain does not change the number of orbits. In step (2), contracting a static interval removes a number of points that are unique orbit representatives, and the count of these is added to the running total kept in the orbit counter. If all the points have been removed, then there are no more orbits to count, and the orbit count is complete. Lemma 11 shows that the operations of transmission, trimming, truncation and merger occurring in steps (3), (5) and (6) do not change the number of orbits of the collection of pairings.

Mergers carried out in step (4) occur when the conditions of Lemma 9 are satisfied, and Lemma 9 shows that they preserve the orbit structure of the collection of pairings. A merger reduces the number of pairings by one, replacing two pairings g_i and g_j by a single periodic pairing acting on the union of their periodic intervals.

In each cycle through these steps the interval width decreases by at least one, either in step (2) or in step (6). It follows that the algorithm terminates, yielding a count of the number of orbits after a number of steps bounded by the width N of the interval. We will obtain a much better bound. To do so, we define a complexity which decreases as we iterate the above steps. Recall that $w_i = b_i - a_i + 1$, $1 \leq i \leq k$, and k is the number of pairings. The complexity X is defined to be

$$X = 4^k \prod_{i=1}^k w_i.$$

The process of executing steps (1)-(7) in turn is called a *cycle*. We will show that when we run through $2k$ cycles, X is reduced by a factor of at least two. See the remark at the end of this section for a geometric interpretation of this complexity.

Call an interval $[x, y]$ a Z -close interval if it is the domain or range of a pairing of subintervals of $[1, Z]$ and if $y - x + 1 \geq (Z - x + 1)/2$, or equivalently, if $y \geq Z/2 + x/2 - 1/2$. Being Z -close corresponds to being relatively close to Z . It means that there is no room for another interval of the same size to the right of y . The value of Z is initially set to N , and as the algorithm proceeds it is reset to the current interval width N' each time the number of pairings k is decreased. The number of pairings decreases when a merger occurs, or when a pairing is truncated to zero width, or when a pairing is transmitted to become trivial (a restriction of the identity) and then eliminated. \square

Claim 1. *The union of two Z -close intervals of equal width is a connected interval.*

Proof. Suppose that $[a, b]$ and $[c, d]$ are equal width Z -close intervals with $c \geq a$. Then $b - a = d - c$, $b - a + 1 \geq (Z - a + 1)/2$ and $d - c + 1 \geq (Z - c + 1)/2$, so $b \geq Z/2 + a/2 - 1/2 \geq d/2 + a/2 - 1/2 = b/2 + c/2 - 1/2$, implying that $b \geq c - 1$ and $[a, b] \cup [c, d]$ is connected. \square

Claim 2. *Suppose that the domain $[a, b] \subset [1, Z]$ of a pairing g is not Z -close. Then the image $[x', y']$ of an interval $[x, y]$ under g^{-1} is not Z -close.*

Proof. Suppose that $[x', y']$ is the image of $[x, y]$ under g^{-1} , where $g : [a, b] \rightarrow [c, d]$. By assumption we have that $b < Z/2 + a/2 - 1/2$. The interval $[x', y']$ is contained in $[a, b]$, so $a \leq x' \leq y' \leq b$. Then $y' < Z/2 + a/2 - 1/2 \leq Z/2 + x'/2 - 1/2$, and $[x', y']$ is not Z -close. \square

Claim 3. *After a series of five cycles either the number of Z -close intervals decreases or the number of pairings decreases.*

Proof. In each cycle at least one maximal pairing is truncated. Suppose that during a series of five cycles, five successive maximal pairings g_1, g_2, g_3, g_4, g_5 occur in turn in the orbit-counting algorithm. The initial maximal pairing g_1 is truncated in the first cycle. Eventually g_1 stops being maximal, and it is then transmitted by the new maximal pairing g_2 in the next cycle. The cycle g_2 is then itself truncated until it is no longer maximal and it in turn is then transmitted by g_3 and so on. When g_{i+1} transmits g_i , the pairing of $[a_i, b_i]$ with $[c_i, d_i]$ is replaced by a pairing of $[a'_i, b'_i]$ with $[c'_i, d'_i]$, and one of the following three cases occurs:

- (1) The range $[c_i, d_i]$ of g_i is transmitted to a non- Z -close interval by g_{i+1} .
- (2) The range $[c_i, d_i]$ is transmitted to a Z -close interval $[c'_i, d'_i]$ by g_{i+1} , and the domain $[a'_i, b'_i]$ of the transmitted pairing is Z -close.
- (3) The range $[c_i, d_i]$ is transmitted to a Z -close interval $[c'_i, d'_i]$ by g_{i+1} , and the domain $[a'_i, b'_i]$ of the transmitted pairing is not Z -close.

We first consider the case where each of g_1, g_2 , and g_3 is orientation preserving. Setting Z to the initial interval width N' and noting that g_1 is initially maximal, we see that the range of g_1 is initially Z -close, with $Z = N' = d_1$. Truncation reduces the range of g_1 to $[c_1, d_2]$, at which point the pairing g_2 becomes maximal. We can assume the range $[c_1, d_2]$ of g_1 is still Z -close, or we are done. In the next cycle the interval $[c_1, d_2]$ is transmitted by g_2 to an interval $[c'_1, d'_1]$. If case (1) applies, this new interval is not Z -close and the number of Z -close intervals has decreased. If case (2) applies, the three intervals $[c_1, d_2]$, $[c'_1, d'_1]$ and $[a'_1, b'_1]$ are all Z -close and of equal width, and therefore satisfy the hypothesis of Claim 1. Lemma 10 then implies that two pairings can be merged, and the number of pairings decreases

during the execution of step (4) in the next cycle. If case (3) occurs, the maximal pairing g_2 is truncated in the second cycle. We can assume its domain remains Z -close, or we are done. When the next pairing g_3 becomes maximal and transmits g_2 in the third cycle, we must fall into one of the first two cases. So the number of Z -close intervals is decreased in the third cycle, or the number of pairings decreases in a merger during a subsequent application of step (4) in the fourth cycle.

Now we consider the possibility that one of g_1, g_2 , and g_3 is orientation reversing. The domain and range of a trimmed orientation-reversing pairing are disjoint, and hence any interval transmitted by an orientation-reversing pairing is not Z -close. If any of g_2, g_3 or g_4 are orientation reversing, then it transmits the previously maximal pairing's range to an interval that is not Z -close, and the number of Z -close pairings decreases. If none is orientation reversing, then there is a sequence of three successive orientation-preserving pairings, and the previous argument applies. \square

Claim 4. *Suppose that Z is set to the current interval size N' and that after a series of truncations in which the maximal pairings are successively g_1, g_2, \dots, g_r , no Z -close intervals remain. Then the complexity X is reduced by a factor of at least two.*

Proof. Truncation of a maximal pairing g_i results in its range $[c_i, d_i]$ of width w_i being reduced to a shorter interval $[c_i, d'_i]$ of width w'_i . This reduces the complexity X by a factor of

$$\frac{w'_i}{w_i} = \frac{d'_i - c_i + 1}{d_i - c_i + 1}.$$

The maximal pairing changes after a truncation if the range of g'_i has truncated sufficiently so that it is contained in the range of g_{i+1} , where $c_{i+1} \leq c'_i$ and $d_{i+1} = d'_i$. Define d''_i by

$$d''_i = c_i - 1 + (w_{i+1}) \frac{Z - c_i + 1}{Z - c_{i+1} + 1},$$

and define

$$w''_i = d''_i - c_i + 1.$$

Differentiation shows that the function

$$f(x) = \frac{x - c_i + 1}{x - c_{i+1} + 1}$$

is increasing with x for $x \geq c_i$, since $c_i \geq c_{i+1}$. Therefore,

$$\begin{aligned} w'_i &= d'_i - c_i + 1 = d_{i+1} - c_i + 1 = (w_{i+1}) \frac{(d_{i+1} - c_i + 1)}{w_{i+1}} \\ &= (w_{i+1}) \frac{(d_{i+1} - c_i + 1)}{(d_{i+1} - c_{i+1} + 1)} \leq (w_{i+1}) \frac{Z - c_i + 1}{Z - c_{i+1} + 1} \\ &= d''_i - c_i + 1 = w''_i. \end{aligned}$$

After a series of truncations of the pairings g_1, g_2, \dots, g_r , the complexity X is multiplied by a factor of

$$\begin{aligned} & \left(\frac{w'_1}{w_1}\right) \left(\frac{w'_2}{w_2}\right) \dots \left(\frac{w'_r}{w_r}\right) \\ \leq & \left(\frac{w''_1}{w_1}\right) \left(\frac{w''_2}{w_2}\right) \dots \left(\frac{w''_{r-1}}{w_{r-1}}\right) \left(\frac{w'_r}{w_r}\right) \\ = & \left(\frac{d''_1 - c_1 + 1}{w_1}\right) \left(\frac{d''_2 - c_2 + 1}{w_2}\right) \dots \left(\frac{d''_{r-1} - c_{r-1} + 1}{w_{r-1}}\right) \left(\frac{w'_r}{w_r}\right) \\ = & \left(\frac{w_2}{w_1}\right) \left(\frac{Z - c_1 + 1}{Z - c_2 + 1}\right) \dots \left(\frac{w_r}{w_{r-1}}\right) \left(\frac{Z - c_{r-1} + 1}{Z - c_r + 1}\right) \left(\frac{w'_r}{w_r}\right) \\ = & \left(\frac{Z - c_1 + 1}{Z - c_r + 1}\right) \left(\frac{w'_r}{w_1}\right). \end{aligned}$$

Now $Z = d_1$, since g_1 was the first maximal pairing truncated, so $Z - c_1 + 1 = d_1 - c_1 + 1 = w_1$. By assumption, there are no Z -close intervals remaining after g_r is truncated, so that $[c_r, d'_r]$ is not Z -close. We then have by the definition of Z -close that

$$\left(\frac{Z - c_1 + 1}{Z - c_r + 1}\right) \left(\frac{w'_r}{w_1}\right) = \frac{w'_r}{Z - c_r + 1} = \frac{d'_r - c_r + 1}{Z - c_r + 1} < 1/2.$$

It follows that X is multiplied by a factor smaller than $1/2$. □

Proof of Theorem 12. The operations involved in one cycle consist of comparisons, additions, subtractions and computing greatest common divisors of the k pairings of G . The number of these operations occurring in each cycle is linear in $k^2 \log N$. Pairings are described by pairs of intervals, whose boundary points are integers of size at most N . So the running time of each cycle is polynomial in $k \log N$.

If the number of pairings k decreases during a cycle, then a pairing has been eliminated because its width has truncated to zero, because it has been transmitted and become the identity on its domain, or because two pairings have merged. When two pairings g_1, g_2 merge, the value of k decreases by one. The product $(w_1)(w_2)$ which occurs in X is replaced by the width of the new pairing, which is at most $(w_1 + w_2) \leq 2w_1w_2$. Therefore $X = 4^k \prod_{i=1}^k w_i$ decreases by a factor of at least two after a merger.

Since each pairing has width at most N , the initial complexity is bounded above by $N^k 4^k$. By Claim 3, each time we run through five cycles, either the number of pairings decreases or the number of Z -close intervals decreases. There are at most k Z -close intervals initially, so after $5k$ cycles either there is a reduction in the number of pairings or there are no Z -close intervals remaining. In the second case the complexity X has decreased by a factor of at least two, by Claim 4.

Therefore, X decreases by a factor of at least two after $5k$ cycles, and the complexity reduces to zero after at most $k(2 + \log_2 N)$ successive series of $5k$ cycles, or after at most $5k^2(2 + \log_2 N)$ cycles. Each cycle runs in time polynomial in $k \log N$, so the total running time is also polynomial in $k \log N$. □

We now apply Theorem 12 to count the number of components of a normal curve or normal surface. An obvious algorithm to count components proceeds by marking vertices connected by common edges until all vertices in a component are reached. This procedure takes time linear in the number of edges of the curve. This is equal

to the sum of the normal coordinates (called W below), but we can achieve an exponential improvement. We first look at normal curves.

Corollary 13. *Let F be a surface with a triangulation T containing t triangles and let γ be a normal curve in F with normal coordinates summing to W . There is a procedure for counting the number of components of γ that runs in time polynomial in $t \log W$.*

Proof. The 1-skeleton of T contains e edges, where $e \leq 3t$. Fix once and for all an ordering of these edges. A normal curve γ intersects each edge of T in a finite number of points. Set N to be the weight W of γ , the sum of the number of intersection points of γ with all the edges of T . Label the intersections of γ and the first edge of the 1-skeleton by the integers $1, 2, \dots, i_1$, the intersections of γ and the j^{th} edge by $i_{j-1} + 1, i_{j-1} + 2, \dots, i_j$, and the intersection of γ and the e^{th} edge of T by $i_{e-1} + 1, i_{e-1} + 2, \dots, i_e$. Then $i_e = N$. Each triangular face of T has three sets of arcs pairing points of $[1, N]$, with one set running between each pair of edges of the face. To each set of arcs we associate a pairing between the intervals at either end of the arcs, as in Figure 4. All the pairings are orientation reversing in this example. In general some will be orientation preserving, as the edge orientations on any triangle can be arbitrary.

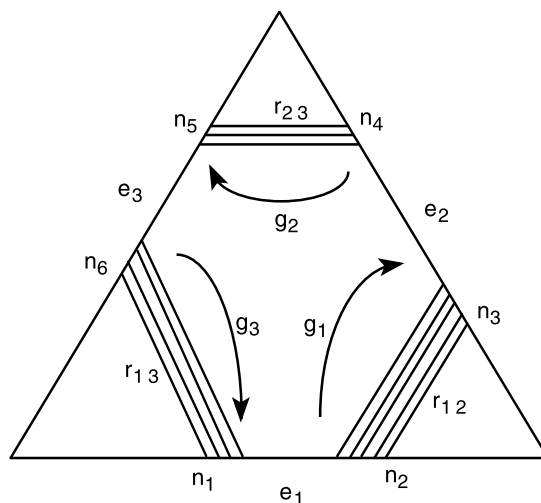


FIGURE 4. Normal arcs on a triangular face give three pairings of ordered intervals: $g_1 : [n_1 + r_{13}, n_2] \rightarrow [n_3, n_3 + r_{12} - 1]$, $g_2 : [n_3 + r_{12}, n_4] \rightarrow [n_5, n_5 + r_{23} - 1]$, and $g_3 : [n_5 + r_{23}, n_6] \rightarrow [n_1, n_1 + r_{13} - 1]$. The normal curve coordinates associated to this face are r_{12}, r_{13} and r_{23} . In a normal surface, r_{12}, r_{13} and r_{23} are each a sum of two normal coordinates.

The number of connected components of γ is the same as the number of orbits of an action of a collection of pairings on $\gamma \cap T^{(1)}$, where two points are paired if they are connected by an edge of γ lying in a triangle. This is precisely the number returned by the orbit counting algorithm. The number of pairings is at most $3t$ and the interval size is W . Applying Theorem 12, we can determine the number of components of the normal curve γ in time polynomial in $t \log W$. \square

A similar argument applies to normal surfaces.

Corollary 14. *Let M be a 3-manifold with a triangulation T containing t tetrahedra and let F be a normal surface in M whose normal coordinates sum to W . There is a procedure for counting the number of components of F which runs in time polynomial in $t \log W$.*

Proof. The 2-skeleton of T contains at most $4t$ faces, and the 1-skeleton contains $e \leq 6t$ edges. Set N to equal the weight W of F , the total number of points in which it intersects the 1-skeleton. Order the edges of $T^{(1)}$ in an arbitrary way and label the intersections of F and the j^{th} edge of $T^{(1)}$ by $i_{j-1} + 1, i_{j-1} + 2, \dots, i_j$. Again we have $i_e = N$. To a pair of edges on a triangular face of the 2-skeleton we associate a pairing between the intervals at either end of the corresponding set of arcs. There are at most three pairings for each face, and the number of faces is at most $4t$, so the number of pairings is bounded above by $12t$. These pairings are determined by the normal coordinates of F .

For a normal surface, the number of connected components of F is the same as the number of components of $F \cap T^{(2)}$, since every component of F intersects the 1-skeleton of T , and if two points on $F \cap T^{(2)} \cap T^{(1)}$ can be joined by a path, then that path can be homotoped into the 2-skeleton of the triangulation. So the number of components equals the number of orbits of $F \cap T^{(1)}$ under pairings that identify two points connected by an edge of $F \cap T^{(2)}$ contained in a face of the 2-skeleton. This number is precisely what is computed by the orbit counting algorithm. Therefore we can determine the number of components of the normal surface F in time polynomial in $t \log W$. \square

Remark. There is a motivating geometrical construction behind the combinatorics of the orbit-counting algorithm which we informally explain. We can associate to each pairing $g_i : [a_i, b_i] \rightarrow [c_i, d_i]$ a pair of “transmission towers” in the upper half-plane, with one tower being the vertical line segment from $(a_i, 0)$ to (a_i, w_i) and the other the vertical segment from $(c_i, 0)$ to (c_i, w_i) . These towers capture the information contained in the pairings. Points in the domain and range of a pairing beam up to the tower leftward at a 45 degree angle, then beam to the paired tower (either straight across, if the pairing preserves orientation, or crossing if not), and then down again. Assign a “cost” to each transmission tower equal to the hyperbolic length in the upper half space model from $y = 1/e$ to the top of the tower. This cost equals $\log(d_i - c_i + 1) + 1$, and the sum of all these costs is essentially the logarithm of the complexity X used above.

The counting algorithm starts with a Euclidean line emerging from the right endpoint of a Z -close interval and going upward at a 45 degree angle. In hyperbolic terms, this is an equidistant curve from the geodesic from the right endpoint to the point at infinity. Initially, we may assume that this sweep line hits the top of at least one transmission tower. We make the highest of these the Z -close tower, and use it to beam all the other towers to its paired interval. If the two domains overlap, we use the highest power of the transmission that can be applied. Eventually the triangle between the current tower and the sweep line is vacant. We sweep leftward with this equidistant curve. As long as the sweep line hits only one tower, we truncate it.

Consider a second equidistant curve P through a point x on \mathbb{R} , with slope $-1/2$. The hyperbolic distance between the slope -1 and slope $-1/2$ lines is a constant

equal to about 0.49. Call the region between these two lines a *zone*. If two paired towers have their tops in a single zone, their domains overlap. If no towers are completely truncated away between the time when the sweep line hits a point x and when there are no towers intersecting the zone above P , each equidistant curve between the two given ones hits a transmission tower based at a point that will be removed by truncation. So in this time either the number of towers has decreased or the sum of all the tower costs has decreased by at least 0.49.

If we merge transmission towers when possible, a complexity based on the cost decreases sufficiently fast to give a polynomial time algorithm. The calculations in Theorem 12 implemented this geometric picture.

5. 3-MANIFOLD KNOT GENUS IS NP

In this section we establish that 3-MANIFOLD KNOT GENUS is NP.

Proof of Theorem 2. We begin with a simplicial complex consisting of t tetrahedra whose faces are identified in pairs, and a collection K of edges in the 1-skeleton of this complex. While there are alternate formats in which a knot and a 3-manifold may be presented, all reasonable ones appear to be transformable to one another in polynomial time.

In time polynomial in t we can check that the link of each vertex is connected and has Euler characteristic two, which means that it is a sphere, and that the link of each edge is a connected curve. These are the necessary conditions to ensure that the underlying space of the complex is a 3-manifold M . Similarly we can check in time polynomial in t that the edges of K form a simple closed curve in M , and that this curve represents a trivial element of the first homology group of M with integer coefficients. We then form the second barycentric subdivision of the triangulation of M , replacing each tetrahedron by 576 tetrahedra. Removing all closed tetrahedra that meet K results in a 3-manifold M_K with a single torus boundary component, the “peripheral torus” that surrounds the knot K . The knot K bounds a surface of genus g in M if and only if there is a surface in M_K of genus g with a single boundary component that is an essential curve on ∂M_K . We restate our problem in the triangulated manifold M_K as the question of whether there exists in M_K an orientable surface of genus g with a single essential boundary component on ∂M_K . By Schubert [21], if such a surface exists, then such a surface exists among the fundamental normal surfaces in M_K . The certificate consists of an integer vector \mathbf{w} in \mathbb{Z}^{7t} giving the normal surface coordinates of this surface.

Recall that not all vectors in \mathbb{Z}^{7t} correspond to normal surfaces. It is necessary that \mathbf{w} satisfies the matching, positivity and quadrilateral conditions. These conditions can be checked in time which is linear in t . Therefore, we can verify that there is a normal surface F with $\mathbf{w} = \mathbf{v}(F)$.

To verify that ∂F is essential on the peripheral torus, we include in the certificate a non-trivial cycle in the 1-skeleton of ∂M_K that intersects ∂F in an odd number of points. Such a cycle can be found in the 1-skeleton of ∂M_K , since curves embedded in the 1-skeleton generate its first homology. Odd intersections with such a cycle imply that ∂F is non-separating on ∂M_K , and in particular does not bound a disk on ∂M_K , ensuring that ∂F corresponds to a longitude curve parallel to K in N_K . Using the orbit counting algorithm, we can count the number of components of the normal surface F and verify that F is connected in time polynomial in $t \log W$, where W is the weight of F , the number of points in which F meets the 1-skeleton

of M_K . We apply Corollary 14 to verify that ∂F is connected in time bounded by a polynomial in $t \log W$. (This last step can be avoided. The number of essential boundary components must be odd, since ∂F represents a non-trivial element in $H_1(\partial M_K; \mathbb{Z}_2)$, and an even number of them can be removed by joining adjacent pairs of curves with annuli on ∂M_K . This gives a surface with one boundary component and the same Euler characteristic. Inessential boundary curves can be capped while increasing Euler characteristic, which gives lower genus.)

To check that F is orientable, we take the vector $2\mathbf{v}(F)$ that doubles each coordinate of the normal surface F , and apply the orbit counting algorithm to determine if the corresponding normal surface \tilde{F} is connected. Since M is orientable, \tilde{F} is connected if F is connected and non-orientable, and has two components if F is connected and orientable. Thus we can verify if F is orientable in time which is linear in $t \log W$.

The Euler Characteristic $\chi(F)$ is determined by the number of vertices, edges and faces of F , which are computable from its normal coordinates in time which are linear in t . Following [11], we let t_i be the number of tetrahedra containing edge e_i and set $\epsilon_{ij} = 1$ if the edge e_i meets the j^{th} normal disk. Then the normal surface F with coordinates $\{v_j\}$ has $\chi(F) = (1/2)f_3 - \sigma(F) + wt(F)$, where $\sigma(F) = \sum v_j$ and $wt(F) = \sum_{i,j} \epsilon_{ij} x_i / t_i$. The values of ϵ_{ij}, t_i are determined by the triangulation and are independent of F . They can be computed in time polynomial in t . Since F is a connected orientable surface with one boundary component, the genus of F is $(1 - \chi)/2$. Thus we can determine the genus of F in time which is linear in t .

Theorem 4 implies that the normal coordinates of this surface are at most $t2^{7t+2}$. There are $7t$ normal coordinates, and each represents a triangle or quadrilateral, so that the total number of intersections with the 1-skeleton satisfies $N \leq 28t^2 2^{7t+2}$. In particular, $\log N$ is bounded above by a polynomial in t . So the fact that F is a spanning surface for K can be verified in time polynomial in t . \square

6. AN EXTENDED COUNTING ALGORITHM

In this section we develop a generalized version of the orbit counting algorithm, that counts not only the number of orbits of a collection of isometries between subintervals of an interval, but also more general quantities which are useful in applications. For example, we can use the extended algorithm to answer the following question: Given a normal surface and a triangulation, how many times does each component of the surface intersect a fixed edge of the triangulation? The extended algorithm allows one to effectively compute the normal coordinates and the Euler characteristic of each connected component of the surface, hence the genus, even when there are exponentially many components. To carry out such computations, we extend the previous analysis to pairings of weighted intervals, in which each point of the interval has associated to it a vector in \mathbb{Z}^d . We are interested in the sum of these vectors over an orbit, the orbit weights.

Consider again a pseudogroup of interval isometries acting on $[1, \dots, N]$. We will assume that there is given as input a non-negative weight function $z : [1, \dots, N] \rightarrow \mathbb{Z}_+^d$, associating to each element of $[1, \dots, N]$ a vector in \mathbb{Z}_+^d satisfying the following condition: The weight at successive points j and $j + 1$ changes at most $4k$ times. In our application, k will be the initial number of pairings, and $4k$ gives an upper bound on how many times an endpoint of a domain or range of a pairing is reached as one moves across $[1, \dots, N]$. The algorithm proceeds as before, while maintaining

data on the orbit weights. We keep track of the orbit weights by maintaining two lists of weighted subintervals. The first

$$L = \{([p_1, q_1], z_1), \dots, ([p_m, q_m], z_m)\}, \quad 1 \leq p_i \leq q_i < p_{i+1},$$

records the current weight values at each point in $[1, N]$, with q_m initially equal to N . This list is updated as the algorithm proceeds. Points in the interval $[p_j, q_j]$ have constant weight $z_j \in \mathbb{Z}^d$, and there are at most $4k$ such intervals. A second list of t subintervals

$$L' = \{([r_1, s_1], v_1), \dots, ([r_t, s_t], v_t)\}, \quad 1 \leq r_i \leq s_i < r_{i+1},$$

consists of a collection of intervals $[r_i, s_i]$ paired with a vector $v_i \in \mathbb{Z}_+^d$. This pair represents $(s_i - r_i + 1)$ orbits, one for each point in the interval $[r_i, s_i]$, and to each of these orbits is assigned the orbit weight v_i . Initially empty, at the algorithm's conclusion L' records the total number of orbits $\sum_i (s_i - r_i + 1)$, along with the orbit weight v_i assigned to each of the $(s_i - r_i + 1)$ orbits in the interval $[r_i, s_i]$.

We define an additional operation, called *transferring weights by a pairing* g . Suppose that $g : [a, b] \rightarrow [c, d]$ is a pairing and that $[c, d]$ carries n different weights, given by the list

$$\{([c = r_1, s_1], v_1), \dots, ([r_n, s_n = d], v_n)\}, \quad 1 \leq r_i \leq s_i < r_{i+1}.$$

The weight function can be split into constant functions on n subintervals of $[c, d]$, where $1 \leq n \leq d - c + 1$.

The transfer operation sets the weights on $[c, d]$ to zero and keeps the orbit weights the same by translating the weight vectors of $[c, d]$ to smaller orbit representatives, as below:

Case (1): g is orientation preserving and $b < c$. Set the weights on $[c, d]$ to zero and for each $1 \leq j \leq n$, add v_j to $g^{-1}([r_j, s_j])$.

Case (2): g is orientation preserving and $b \geq c$. Then $g : [a, b] \rightarrow [c, d]$ is a periodic pairing of period $t = c - a$. We set the weights of points in $[c, d]$ to zero and adjust weights in $[a, c - 1]$ to preserve the orbit weight. The points in $[c, d]$ have weights given by the intersection of $[c, d]$ with intervals in L . These weights are described by

$$\{([c = r_1, s_1], v_1), \dots, ([r_n, s_n = d], v_n)\}, \quad 1 \leq r_i \leq s_i < r_{i+1}.$$

For each interval $([r_j, s_j], v_j)$, $1 \leq j \leq n$, of width w_j and constant weight v_j , add $[\frac{w_i}{t}]v_j$ to the weight of each point in $[a, c - 1]$, and add an additional v_j to the weight of each point in $[a, c - 1]$ that is congruent mod(t) to a point in $[a + [\frac{w_i}{t}]t, s_j]$, if any such point exists.

Case (3): g is orientation reversing. We first trim g . This does not affect the orbits or the orbit weights. For a trimmed, orientation-reversing pairing $g : [a, b] \rightarrow [c, d]$, set the weights on $[c, d]$ to zero, and for each $1 \leq j \leq n$ add v_j to $g^{-1}([r_j, s_j])$.

Lemma 15. *The operation of transfer sets the weights on $[c, d]$ to zero and preserves the orbit weights of a collection of pairings. The number of distinct weights taken by the weight function on $[1, N]$ increases by at most four following a transfer operation.*

Proof. In Cases (1) and (3), the decrease in the weight function at one point x in an orbit is exactly offset by an equal increase at the point $g^{-1}(x)$ in the same orbit.

In Case (2), each point in $[a, d]$ is in the orbit of a unique point in $[a, t - 1] = [a, c - 1]$ under the iterates of g . The total weight of an orbit within a periodic pairing

is transferred to the orbit representative in this initial subinterval by applying powers of g . Adding weight v_j to an orbit representative in $[a, c - 1]$ of each point in $[r_j, s_j]$ while setting the weight at that point to zero, preserves the orbit weight. The resulting weights on $[a, c - 1]$ are obtained by adding appropriate multiples of v_j , with the factor being the number of orbits of a point in $[a, c - 1]$ that lie in $[r_j, s_j]$. The number of orbits under g of a point in $[a, c - 1]$ that lie in $[r_j, s_j]$ is $\lfloor \frac{w_j}{t} \rfloor$ or one more than this for points whose orbit hits the last $w_j \pmod{t}$ points of $[r_j, s_j]$. It follows that the transfer operation in Case (2) preserves the number of orbits and the orbit weight.

Setting the weights on $[c, d]$ to zero can cause at most two new points where a weight change occurs. The transferred weights from a constant weight interval $[r_j, s_j]$ result in a net increase of at most two pairs of successive points where the weight changes, at the preimages of its two endpoints. In Cases (1) and (3), transferred weights from an interval with non-constant weights results in an increase in the number of weight changes in the domain of g . The increase in the number of weight changes in the domain is exactly cancelled by the decrease in the number of weight changes in the range of g , except possibly for two extra weight changes at the boundary points a, b of the domain. In Case (2) the same holds, but with $[a, c - 1]$ replacing the domain. In each case the number of constant weight intervals m' is increased by at most four during a transfer operation. \square

We now describe the modified algorithm. Again N' represents the current interval length, and we set m' to be the current number of constant weight intervals.

Weighted orbit counting algorithm: Let $\{g_i : [a_i, b_i] \rightarrow [c_i, d_i], 1 \leq i \leq k\}$ be a collection of pairings between subintervals of $\{1, 2, \dots, N\}$, and let

$$L = \{([p_1, q_1], z_1), \dots, ([p_m, q_m], z_m)\}, 1 \leq p_i \leq q_i < p_{i+1} \leq N,$$

be a list representing a collection of weights on $[1, N]$, with the weight on $[p_j, q_j]$ equal to $z_j \in \mathbb{Z}^d$. Initialize a second weight list L' to be empty. The algorithm proceeds as before, reducing the interval size N until it reaches zero, but this time keeping track of orbit weights by maintaining the lists L, L' .

- (1) Search through the pairings and delete any pairings which are restrictions of the identity. Leave the weight lists L, L' unchanged.
- (2) Search for and contract static intervals. If the interval $[r, s]$ is contracted, has constant weight z , and is not contained in a larger contracted interval with the same weight z , add an interval $([N' + 1, N' + s - r + 1], z)$ of width $s - r + 1$ to the end of L' , with associated weight z . Alter L by replacing $[1, N']$ by $[1, N' - (s - r + 1)]$, and altering each g_i by replacing any point x in the domain or range of g_i with $x > s$ by $x - (s - r + 1)$. Replace the weight function by a new weight function w' , which at points $x > s$ satisfies $w'(x) = w(x + (s - r + 1))$ and agrees with w at points $x < r$.
- (3) Trim all orientation-reversing pairings whose domain and range overlap. Leave the weight lists L, L' unchanged.
- (4) Search for pairs of periodic pairings g_i and g_j whose domains and ranges satisfy the condition of Lemma 9. If any such pair exists, then perform a merger as in Lemma 9, replacing g_i and g_j by a single periodic pairing, with translation distance $\text{GCD}(t_i, t_j)$, acting on the union of the domains

and ranges of g_i and g_j . Leave the weight lists L, L' unchanged. Repeat until no more mergers can be performed.

- (5) Find a maximal g_i . For each g_j with $j \neq i$, if the range of g_j is contained in $[c_i, N']$, transmit g_j by g_i . Leave the weight lists L, L' unchanged.
- (6) Find the smallest value of c such that the interval $[c, N']$ intersects the range of at most one pairing g_i , with $g_i : [a_i, b_i] \rightarrow [c_i, N']$. Transfer the weights on $[c_i, N']$ by g_i , and then truncate the pairing g_i .
- (7) If the interval size N' has decreased to zero, output the list L' and stop. Otherwise start again with Step (1).

For a \mathbb{Z}^d -valued function on $[1, 2, \dots, N]$ whose values are given by the list $L = \{([p_1, q_1], z_1), \dots, ([p_m, q_m], z_m)\}$, $1 \leq p_i \leq q_i < p_{i+1}$, define the *total weight* of L to be $\sum_{i=1}^m |z_i|$.

Theorem 16. *Suppose there is a pseudogroup generated by k pairings with \mathbb{Z}^d -valued weights, $\{g_i\}_{i=1}^k$ on $[1, N]$, such that there is a partition of $[1, N]$ into m disjoint subintervals in which the weights are constant, and such that the total weight is at most D . Then the weighted orbit-counting algorithm outputs a list with one point for each orbit and corresponding orbit weights, and runs in time polynomial in $kmd \log D \log N$.*

Proof. We will check that the running time of the algorithm is larger than that of the previous unweighted version by a factor which is a polynomial in $md \log D$.

The proof that the algorithm terminates is the same as that given for Theorem 12. There is some extra overhead involved in keeping track of weights that modifies the calculation of the running time. We indicate these additional calculations below. We now check that at each step in the algorithm the orbit weight is unchanged for any orbit remaining in L , and that eliminated orbits have their orbit weights correctly recorded in L' .

As we run through the steps of the algorithm, steps (1), (3), (4), (5) and (7) preserve the orbit structure, the weight function and the interval $[1, N']$, so neither of the lists L, L' are changed. The number of constant weight sub-intervals is also unchanged.

Step (2), a contraction, does change the orbit structure, and also shortens $[1, N']$. In step (2) the procedure adds the eliminated orbits and their weights to L' . The number of constant weight intervals is not increased, and may be decreased. Maintaining the two lists requires at most $O(md \log D)$ additional steps.

In step (6), a truncation, points are eliminated from the end of the interval $[1, N']$. However since we first transfer the weights of these points, the eliminated points all have weight zero and the weight of an orbit is unaffected.

The number of steps involved in resetting the weights in L for a transfer operation is given by a polynomial in $m'd \log D \log N$. Since m' increases by at most four at each of the polynomially many steps of the algorithm, m' is bounded by a polynomial in $mk \log N$.

Combining the running time of each of the steps, whose number is given by a polynomial in $k \log N$, gives a polynomial in $kmd \log D \log N$ for the total running time. □

Corollary 17. *Let M be a 3-manifold with a triangulation T containing t tetrahedra and let F be a normal surface in M of total weight W . There is a procedure*

for counting the number of components of F and determining the topology of each component which runs in time polynomial in $t \log W$.

Proof. We begin as in Corollary 14 by assigning an integer in $[1, N]$ to each point of intersection between the normal surface and an edge of the triangulation, where N is the total number of intersections of F with the 1-skeleton, and again associate three pairings to each face of the triangulation, one to each pair of edges in the face. The number of pairings that results is bounded above by $12t$.

We next define a weight function $w(x)$ which assigns integer weights $(z_1, z_2, \dots, z_{7t})$ to each point in $[1, N]$. Initially z_i is set to zero for all i at all points $x \in [1, N]$. A tetrahedron can have as many as five distinct elementary disk types with non-zero coefficients, four triangles and one quadrilateral. If the j^{th} elementary disk-type occurs, then fix one of the edges that it meets, and add 1 to the j^{th} component of the weight vector at each of the indices that the j^{th} elementary disk meets on that edge. The orbit weights are then the normal coordinates of the components of the normal surface F . Each point in the output list L' corresponds to a component of the normal surface with normal coordinates given by the corresponding weight in \mathbb{Z}^{7t} . Theorem 16 tells us that the list L' is computed in time polynomial in $kmd \log D \log N$. We now bound these constants in terms of t .

Since each edge of a tetrahedron meets at most three disk types in that tetrahedron, each edge of a tetrahedron can contribute at most six points at which the weight vector changes. Given six edges to each tetrahedron, we have $m \leq 36t$. As before we have a bound for the number of pairings $k \leq 12t$ and the number of normal coordinates is given by $d = 7t$. The total weight bounds the normal coordinates, and with $D = W$ we get a bound on the running time that is polynomial in $t \log W$. \square

Corollary 18. *Let M be a 3-manifold with a triangulation T containing t tetrahedra and let F be a fundamental normal surface in M . There is a procedure for counting the number of components of F and the topology of each component which runs in time polynomial in t .*

Proof. Theorem 4 gives a bound for the normal coordinates of F of $t2^{7t+2}$. Recall that there are $7t$ normal coordinates, each representing a triangle or quadrilateral, so the total number of intersections with the 1-skeleton satisfies $W \leq 28t^2 2^{7t+2}$. In particular, $\log W$ is bounded above by a polynomial in t . Plugging this in for W in Corollary 17 we get a bound for the running time which is a polynomial in t . \square

7. THE COMPLEXITY OF MINIMAL SPANNING AREA

In this section we examine the complexity of the problem of determining the smallest area of a spanning surface for a curve in a 3-dimensional manifold. Such an area calculation problem seems at first to be ill-suited to a complexity analysis, since it has real solutions depending on a choice of Riemannian metric.

We recast the area calculation problem into a discretized form where its complexity can be analyzed. Given a curve in a suitably discretized Riemannian 3-manifold, we ask whether it bounds a surface of area less than C , where C is an integer. To describe a metric on a 3-manifold with a finite amount of data, we restrict to piecewise flat metrics and manifolds constructed from collections of flat tetrahedra and triangular prisms whose faces are identified by isometries. The curvature of such PL metrics can be defined as a limit of smooth curvatures, and is concentrated

along their edges and vertices. A particular manifold in this class is described by a decomposition into tetrahedra or triangular prisms with a rational (or integer) length assigned to each edge. The metric on this tetrahedron or prism is then taken as the metric on the Euclidean tetrahedron or prism with those edge lengths. In the case of a prism we also set the angles of quadrilateral faces to be right angles. Prisms are allowed in this construction in order to form metrics with rational lengths on spaces that are products. Identified 2-dimensional faces are required to be isometric. We do not require that the total angle around an edge is 2π , nor do we make any metric conditions at a vertex. This type of metric is described by a finite set of data, and can be used to approximate Riemannian metrics on a manifold. Up to scaling, we can take all the edge lengths to be integers. We call these objects *metrized PL 3-manifolds*. A curve is given as a collection of edges in the 1-skeleton of M . We will show that given an integer C , determining whether the smallest spanning surface for a curve in such a 3-manifold has area less than C is **NP**-hard.

Problem. MINIMAL-SPANNING-AREA

INSTANCE: A 3-dimensional metrized PL manifold M , a 1-dimensional curve K in the 1-skeleton of M , and a natural number C .

QUESTION: Does the curve bound a surface of area $A \leq C$?

The size of an instance is given by the number of bits needed to describe all the edge lengths and C .

Theorem 19. *MINIMAL-SPANNING-AREA is NP-hard.*

Proof. We reduce in polynomial time an instance of the **NP**-hard problem ONE-IN-THREE SAT to an instance of MINIMAL-SPANNING-AREA. This shows that MINIMAL-SPANNING-AREA is at least as hard, up to polynomial time reduction, as ONE-IN-THREE SAT.

As a first step, we set up a 2-dimensional version of MINIMAL-SPANNING-AREA. We then construct a 3-manifold by a thickening process, with the property that a minimizing surface must remain within the 2-complex.

Given a boolean expression representing an instance of ONE-IN-THREE SAT, we construct a triangulated metrized 2-complex and an integer C . This complex contains a curve K with the property that the expression admits a satisfying assignment if and only if K bounds a surface of area less than C . This metrized complex is shown in Figure 5 for the expression $(x_1 \vee x_2 \vee x_3) \wedge (x_1 \vee \bar{x}_2 \vee \bar{x}_3)$.

The branching surface is similar to the one used in the proof of Theorem 1, but carries the additional structure of a metrized triangulation, whose triangles have flat metrics of prescribed edge length. The metrized triangles are constructed so that near each of the m boundary components corresponding to clauses of the boolean expression there are three triangulated disks of area close to one, one on each of the three handles coming into the punctured sphere near the boundary component. These disks are shaded in Figure 5. Each of these shaded disks is chosen to have area between 1 and $1 + 1/2m$. The surface is constructed so that the union of all triangles in the rest of the surface has total area less than $1/2$.

We saw in Theorem 1 that a spanning surface which has minimal genus goes over each of the shaded disks at most once and goes over exactly one shaded disk for each of the m clauses. It follows that such a surface has total area $m < A < m + 1$. Furthermore, a satisfying assignment for ONE-IN-THREE SAT leads to

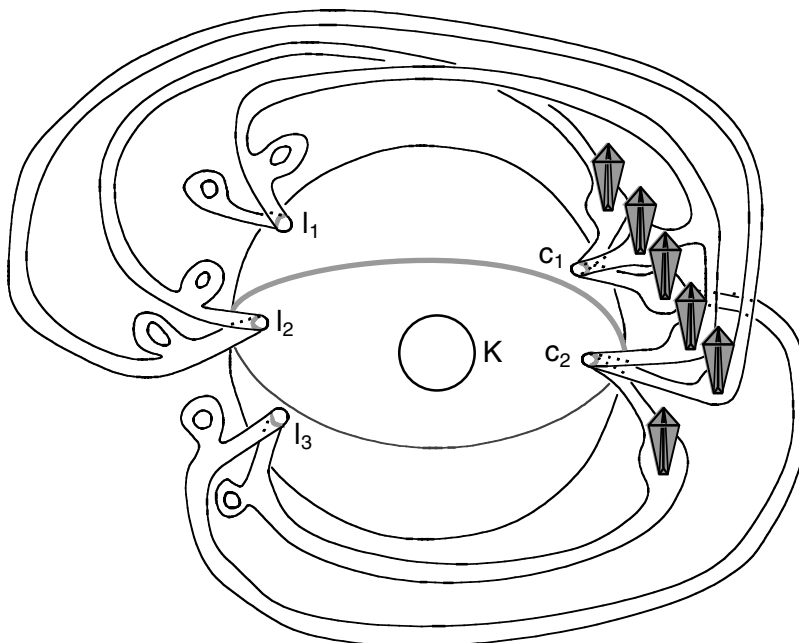


FIGURE 5. A metrized branching surface corresponding to the boolean expression $(x_1 \vee x_2 \vee x_3) \wedge (x_1 \vee \bar{x}_2 \vee \bar{x}_3)$. The picture is not to scale. The shaded prisms are constructed to each have area one, while the rest of the surface has total area less than $1/2$.

an embedded spanning surface, with the satisfying values of the variables selecting branches of the surface, and such a spanning surface has area less than $m + 1$. So an instance of ONE-IN-THREE SAT can be reduced to an instance of MINIMAL-SPANNING-AREA for this 2-complex.

To pass to a 3-manifold, thicken each triangle in the branching surface to a triangular prism, triangulated as in the proof of Theorem 1, and with a product metric. This produces a 3-manifold M which is a thickened-up version of the 2-complex. Projection to the branching surface is area non-increasing, and area decreasing for a surface with boundary on the branching surface but not contained in it. Therefore a least area surface spanning K must lie on the branching surface. A closed manifold DM with a piecewise-smooth metric can be obtained by a doubling construction as in Theorem 1. The doubling involution is an isometry, so that reflecting a surface meeting $DM \setminus M$ into M does not increase area. It follows that the embedded spanning surface on the branching surface is a least area surface in DM . \square

8. OPEN QUESTIONS

Among many unresolved questions are:

1. Is the problem of determining knot genus for knots in the 3-sphere **NP**-hard?
2. Is determining the genus of a knot in a 3-manifold **NP**? This amounts to showing that finding a lower bound to the knot's genus is an **NP** problem, in contrast to the upper bound we have investigated. Recall that the genus of a knot

is the least possible genus of all spanning surfaces. We have shown that certifying that the genus is at most g is **NP**, but have left open the possibility that the genus may be smaller than g . If the answer to this question is yes, then we can certify that a non-trivial knot has positive genus, and it would follow that UNKNOTTING is both **NP** and **coNP**.

REFERENCES

- [1] M. Dehn, "Über die Topologie des dreidimensionalen Raumes", *Math. Annalen*, 69 (1910) 137–168. MR1511580
- [2] M.R. Garey and D.S. Johnson, "Computers and intractability. A guide to the theory of **NP**-completeness", W. H. Freeman and Co., San Francisco, 1979. MR0519066 (80g:68056)
- [3] W. Haken, "Theorie der Normalflächen: Ein Isotopiekriterium für den Kreisknoten", *Acta Math.*, 105 (1961) 245–375. MR0141106 (25:4519a)
- [4] J. Hass, "Algorithms for recognizing knots and 3-manifolds", *Chaos, Solitons and Fractals*, 9 (1998) 569–581. MR1628743 (2000a:57038)
- [5] J. Hass, J. C. Lagarias and N. Pippenger, "The computational complexity of Knot and Link problems", *Journal of the ACM*, 46 (1999) 185–211. MR1693203 (2000g:68056)
- [6] J. Hass and J. C. Lagarias, "The number of Reidemeister moves needed for unknotting", *J. Amer. Math. Soc.* 14 (2001), no. 2, 399–428. MR1815217 (2001m:57012)
- [7] G. Hemion, *The Classification of Knots and 3-Dimensional Spaces*, Oxford University Press, 1992. MR1211184 (94g:57015)
- [8] J. Hempel, *3-Manifolds*, Princeton University Press, Princeton, NJ, 1976. MR0415619 (54:3702)
- [9] W. Jaco, *Lectures on three-manifold topology*, CBMS Regional Conference Series in Mathematics, 43 AMS, Providence, RI, 1980. MR0565450 (81k:57009)
- [10] W. Jaco and U. Oertel, "An Algorithm to Decide If a 3-Manifold Is a Haken Manifold", *Topology*, 23 (1984) 195–209. MR0744850 (85j:57014)
- [11] W. Jaco and J. L. Tollefson, "Algorithms for the Complete Decomposition of a Closed 3-Manifold", *Illinois J. Math.*, 39 (1995) 358–406. MR1339832 (97a:57014)
- [12] W. Jaco and J. H. Rubinstein, "PL Equivariant Surgery and Invariant Decompositions of 3-Manifolds", *Advances in Math.*, 73 (1989) 149–191. MR0987273 (90g:57016)
- [13] F. Jaeger, D. L. Vertigan and D. J. A. Welsh, "On the Computational Complexity of the Jones and Tutte Polynomials", *Math. Proc. Cambridge Phil. Soc.*, 108 (1990) 35–53. MR1049758 (91h:05038)
- [14] H. Kneser, "Geschlossene Flächen in dreidimensionalen Mannigfaltigkeiten", *Jahresbericht Math. Verein.*, 28 (1929) 248–260.
- [15] E. E. Moise, "Affine Structures in 3-Manifolds, V: The Triangulation Theorem and Hauptvermutung", *Ann. Math.*, 56 (1952) 96–114. MR0048805 (14:72d)
- [16] C.H. Papadimitriou, *Computational complexity*. Addison-Wesley Publishing Company, Reading, MA, 1994. MR1251285 (95f:68082)
- [17] M. O. Rabin, "Recursive Unsolvability of Group-Theoretic Problems", *Ann. Math.*, 67 (1958) 172–194. MR0110743 (22:1611)
- [18] J.H. Rubinstein, "An algorithm to recognize the 3-sphere", *Proceedings of the International Congress of Mathematicians*, Vol. 1, 2 (Zurich, 1994), Birkhauser, Basel, 601–611. MR1403961 (97e:57011)
- [19] Schaefer, T.J. "The complexity of satisfiability problems", *Proc 10th Ann. ACM. Symp. on Theory of Computing*, ACM, NY (1978) 216–226. MR0521057 (80d:68058)
- [20] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley and Sons, 1986. MR0874114 (88m:90090)
- [21] H. Schubert, "Bestimmung der Primfaktorzerlegung von Verkettungen", *Math. Zeitschr.*, 76 (1961) 116–148. MR0141107 (25:4519b)
- [22] A. Sebö, "Hilbert Bases, Caratheodory's Theorem and Combinatorial Optimization", in: R. Kannan and W. R. Pulleyblank (Eds.), *Integer Programming and Combinatorial Optimization*, University of Waterloo Press, 1990, 431–455.
- [23] H. Seifert, "Über das Geschlecht von Knoten", *Math. Annalen*, 110 (1935) 571–592. MR1512955

- [24] A. Thompson, Thin position and the recognition problem for S^3 . *Math. Res. Lett.* 1 (1994), 613–630. MR1295555 (95k:57015)
- [25] D. J. A. Welsh, *Complexity: Knots, Colourings and Counting*, Cambridge University Press, 1993. MR1245272 (94m:57027)
- [26] D. J. A. Welsh, “The Complexity of Knots”, *Ann. Discr. Math.*, 55 (1993) 159–173. MR1217989 (94c:57021)
- [27] D. J. A. Welsh, “Knots and Braids”, *Contemp. Math.*, 147 (1993) 109–123. MR1224698 (94g:57014)

DEPARTMENT OF MATHEMATICS, STATISTICS, AND COMPUTER SCIENCE, UNIVERSITY OF ILLINOIS AT CHICAGO, CHICAGO, ILLINOIS 60607

E-mail address: agol@math.uic.edu

SCHOOL OF MATHEMATICS, INSTITUTE FOR ADVANCED STUDY AND DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, DAVIS, CALIFORNIA 95616

E-mail address: hass@math.ucdavis.edu

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, DAVIS, CALIFORNIA 95616

E-mail address: wpt@math.ucdavis.edu

Current address: Department of Mathematics, Cornell University, Ithaca, New York 14853

E-mail address: wpt@math.cornell.edu