# UNIQUENESS IN HAKEN'S THEOREM

MICHAEL FREEDMAN AND MARTIN SCHARLEMANN

Abstract. Following Haken [Ha] and Casson-Gordon [CG], it was shown in [Sc] that given a reducing sphere or $\partial$-reducing disk $S$ in a Heegaard split manifold $M$ in which every sphere separates, the Heegaard surface $T$ can be isotoped so that it intersects $S$ in a single circle. Here we show that when this is achieved by two different positionings of $T$, one can be moved to the other by a sequence of

- isotopies of $T$ rel $S$
- pushing a stabilizing pair of $T$ through $S$ and
- eyegelass twists of $T$.

This last move is inspired by one of Powell's proposed generators for the Goeritz group [Po].

It is a classic theorem of Haken [Ha] that any Heegaard splitting $M = A \cup_T B$ of a closed orientable reducible 3-manifold $M$ is reducible; that is, there is an essential sphere in the manifold that intersects $T$ in a single circle. Casson-Gordon [CG, Lemma 1.1] refined and generalized the theorem, showing that it applies also to essential disks, when $M$ has boundary and, more specifically, if $E$ is a disjoint union of essential disks and 2-spheres in $M$ then there is a similar family $E^*$, obtained from $E$ by ambient 1-surgery and isotopy, so that each component of $E^*$ intersects $T$ in a single circle. (We say $E^*$ is a $T$-reducing system in $M$.) It is now known [Sc] that in fact we may take $E^* = E$ so long as every sphere in $M$ separates.

Here we consider a naturally related uniqueness question: Suppose $E_0$ and $E_1$ are each $T$-reducing systems in $M$, and the systems $E_0$, $E_1$ are isotopic rel $\partial$ in $M$. Is $E_0$ isotopic to $E_1$ *via $T$-reducing systems*?

Counterexamples spring to mind, even when $M$ is irreducible and each $E_i$ is simply a single disk.

**Example:** Suppose $E_0$ is a $\partial$-reducing disk for $M$ which intersects $T$ in a single circle, so it is a $\partial$-reducing disk for $T$. Suppose $T$ is stabilized, with stabilizing disks $D_A \subset A$, $D_B \subset B$ and both disks are disjoint from $E_0$. A regular neighborhood of $D_A \cup D_B$ is a ball $\beta$ which $T$ intersects in a standard genus 1 summand. We call such a pair

*Date*: April 17, 2020.

$(\beta, T)$ a *standard bubble*. We can imagine $\beta$ as a small ball, the regular neighborhood of a point in the destabilized Heegaard surface $T'$. Now isotope $\beta$ along a path in $T'$ which passes once through $E_0$, and let $E_1$ be the result of pushing $E_0$ by the resulting ambient isotopy of $M$. Then $E_0$ and $E_1$ are isotopic, but they can't be isotopic via $\partial$-reducing disks for $T$, since the circles $T \cap E_i, i = 0, 1$ are not isotopic in $T$.

**Example:** More subtly, suppose $D_A$, $D_B$ are disjoint essential disks in $A$ and $B$ respectively, and $\gamma$ is a path in $T$ connecting their boundaries. The complex $D_A \cup \gamma \cup D_B$ is called an *eyeglass* for $T$ ([FS, Definition 2.1]). Associated to such an eyeglass is an isotopy of $T$ in $M$ back to itself (with support near the eyeglass) called an *eyeglass twist*. It is illustrated in Figure 1. Suppose $E_0$ is a reducing disk for $T$ and the circle $E_0 \cap T$ essentially intersects the bridge $\gamma$ of the eyeglass. Then the disk $E_1$ obtained by pushing $E_0$ along by the resulting ambient isotopy of $M$ cannot be isotopic via $\partial$-reducing disks, again since the circles $T \cap E_i, i = 0, 1$ are not isotopic in $T$.

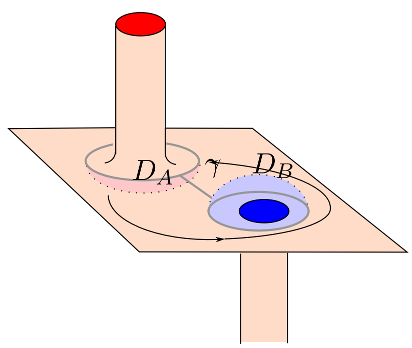


Figure 1. Eyeglass twist

Our goal is to show that the two operations just described are essentially the only two obstacles to uniqueness in general, so long as every 2-spheres separates. The extra condition is needed only to invoke [Sc].

## 1. Background and results

Suppose $M = A \cup_T B$ is a Heegaard splitting of a compact orientable 3-manifold.

**Definition 1.1.** *A disjoint collection of reducing spheres and $\partial$-reducing disks for $M$ is an $M$-reducing system; each element is called an $M$-reducer.*

*For $M = A \cup_T B$ a Heegaard splitting, say an $M$-reducing system (resp $M$-reducer) is a $T$-reducing system (resp $T$-reducer) if each component intersects $T$ in a single circle. A disk $M$-reducer whose boundary lies in $\partial_- A$ (resp $\partial_- B$) is an $A$-disk (resp $B$-disk).*

*Given an $M$-reducing system $E$, a positioning of $T$ so that $E$ is a $T$-reducing system is called a* solution *to the $M$-reducing system.*

Suppose $(E, \partial E) \subset (M, \partial M)$ is a $T$-reducing system in $M$.

**Definition 1.2.** *Let $S \subset M$ be a reducing sphere for $T$ that is disjoint from $E$ and cuts off a genus 1 stabilizing summand of $T$ inside a ball. Let $\gamma$ be an arc in $T$ with one end at a component $\overline{E}$ of $E$, the other end at $S \cap T$, and $\gamma$ is otherwise disjoint from both $E$ and $S$. Alter $\overline{E}$ by tube summing it to $S$ along a neighborhood of $\gamma$ and call the result $\overline{E}'$. Replace $\overline{E}$ by $\overline{E}'$ in $E$ and call the result $E'$. $E'$ is obtained by a* bubble move *on $E$ along $\gamma$ with bubble $S$.*

We can think of $S$ as a 'bubble' that passes through $E$ to create $E'$. Note that $E$ and $E'$ are properly isotopic in $M$.

**Definition 1.3.** *Let $(D_A, \partial D_A) \subset (A, \partial A)$ and $(D_B, \partial D_B) \subset (B, \partial B)$ be disjoint essential disks that are also disjoint from $E$. Let $\gamma$ be an arc in $T$ transverse to $E$, with one end at $\partial D_A$, other end at $\partial D_B$, and otherwise disjoint from $D_A \cup D_B$. Perform an eyeglass twist on $T$ using the eyeglass $D_A \cup \gamma \cup D_B$. The eyeglass twist returns $T$ to itself, but may alter $E$. The image $E'$ of $E$ is said to be obtained from $E$ by an eyeglass twist.*

Note that if $E$ is disjoint from the eyeglass then $E = E'$; in any case $E$ and $E'$ are properly isotopic in $M$.

**Definition 1.4.** *Suppose $E_0$ and $E_1$ are each a $T$-reducing system in $M$. An isotopy $E_s, 0 \leq s \leq 1$ from $E_0$ to $E_1$ in $M$ is an equivalence (and $E_0, E_1$ are equivalent) if $E_s$ is a $T$-reducing system for all $s$.*

**Definition 1.5.** *$E_0$ and $E_1$ are* congruent *if a sequence of equivalences, bubble moves and eyeglass twists carries $E_0$ to $E_1$.*

We intend to show:

**Theorem 1.6.** *Suppose every 2-sphere in $M$ separates. If $E_0, E_1$ are $T$-reducing systems that are properly isotopic in $M$, then $E_0$ and $E_1$ are congruent.*

In conjunction with [Sc], this means that, when $M$ contains no $S^1 \times S^2$ summand, any $M$-reducing system is isotopic in $M$ to a $T$-reducing system that is unique up to congruence.

For the purposes of the proof we will assume that $\partial M$ contains no sphere components; these add a small amount of complexity, which we leave for the reader to resolve.

**Example:** Theorem 1.6 is obvious for reducing spheres $E_0, E_1$ that intersect $T$ in disjoint circles: Then each component of $E_0 \cap E_1$ is a circle lying in either $A$ or $B$. An innermost one in $E_0 \cap A$, say, cuts off also a subdisk of $E_1 \cap A$. Since $A$ is irreducible, the latter disk can be isotoped to the former in $A$. Eventually such isotopies make $E_0$ and $E_1$ disjoint, so they are parallel in $M$. The splitting that $T$ induces inside of the collar between them is simply a sum of stabilizing pairs, per Waldhausen ([Wa], [R]), which can be passed through $E_0$ bubble by bubble until the spheres are equivalent. A similar argument applies if $E_0, E_1$ are $\partial$-reducing disks. So the interest focuses on cases in which $E_0 \cap E_1 \cap T \neq \emptyset$.

## 2. Sweepouts, spines, and labels for the graphic

Here we briefly review the classical sweep-out technology on $M = A \cup_T B$. $A$ (and also $B$) is a compression-body, which can be viewed (dually to the original definition in [Bo]) as a compact connected orientable 3-manifold obtained from a (possibly disconnected) surface $\times I$ by attaching 1-handles to surface $\times\{1\}$. The boundary of $A$ is the disjoint union of $\partial_- A =$ surface $\times\{0\}$ and a connected surface denoted $\partial_+ A$. From its construction we see that $A$ deformation retracts to the union of $\partial_- A$ and the cores of the 1-handles, where the latter are extended down through $\partial_- A \times I$ via the product structure.

More generally, a *spine* $\Sigma$ of $A$ is the union of $\partial_- A$ and a certain type of graph in $A$: all valence 1 vertices in the graph lie on $\partial_- A$, and $A$ deformation retracts to $\Sigma$; indeed $A - \Sigma \cong \partial_+ A \times (0, 1]$. (Sometimes we will not distinguish between $\Sigma$ and a thin regular neighborhood of $\Sigma$.) $A$ has many spines, but in an argument that goes back to Whitehead [Wh] (who was concerned with spheres, not disks) one can change one spine to any other by a sequence of "edge-slides", in which one edge is slid over others and along $\partial_- A$ [ST, Section 1].

A properly embedded annulus in $A$ is *spanning* if its two boundary components lie, one each, in $\partial_+ A$ and $\partial_- A$. Let $E_A \subset A$ be a properly embedded disjoint collection of spanning annuli and disks that compress $\partial_+ A$ in $A$. Essentially the same argument as in [ST, Section 1] shows that there is a spine $\Sigma$ for $A$ with the properties:

- Each disk in $E_A$ intersects $\Sigma$ in a single meridian of an edge.
- Each annulus in $E_A$ intersects $\Sigma$ only in $\partial_- A$.

Moreover, given $E_A$, one can choose the parameterization $A - \Sigma \cong \partial_+ A \times (0, 1]$ so that the half-open annuli $E_A - \Sigma$ are parameterized as $(E_A \cap \partial_+ A) \times (0, 1]$. We will say that such a spine and parameterization *comports with* $E_A$. Note that, via Hatcher's work [Ha1] ,[Ha2], the exact parameterization involves no choice, in the sense that its space of parameters is contractible.

Combining these ideas, if $E'_A \subset A$ is another such collection, then one can move from a spine (and associated parameterization) that comports with $E_A$ to one that comports with $E'_A$ via a sequence of edge slides.

Now we export all these ideas to the setting at hand: a Heegaard split $M = A \cup_T B$ and two $T$-reducing systems $E_0$ and $E_1$ that are isotopic in $M$. Each $E_i$ intersects each compression-body $A$ (resp $B$) in a collection of spanning annuli and essential disks $E_{i,A} = E_i \cap A$ (resp $E_{i,B} = E_i \cap B$). Choose spines $\Sigma_{i,A} \subset A$ (resp $\Sigma_{i,B} \subset B$) so that each comports with $E_{i,A}$ (resp $E_{i,B}$). For each $i = 0, 1$ combine the comporting parameterizations in each compression-body, to parameterize the entire complement of the spines in $M$ as $T \times (0, 1)$, picking the convention that the spine of $A$ is the limit of $T \times \{t\}$ as $t \to 0$. Then the complement of the spines in $M$ is swept-out by copies of $T$ in such a way that each copy of $T$ intersects each component of $E_i$ in a single circle. Denote the copy $T \times \{t\}$ in such a sweepout by $T_t$.

The core argument will mirror that of [FS, Section 4], with the isotopy $E_s, 0 \leq s \leq 1$ from $E_0$ to $E_1$ replacing what was there a sweepout of $S^3$ by level 2-spheres. In addition we use $s$ to simultaneously parameterize a movie of the sequence of edge slides on the spines that take $\Sigma_{0,A} \cup \Sigma_{0,B}$ to $\Sigma_{1,A} \cup \Sigma_{1,B}$. Together, this sweep-out and the isotopy $E_s$ (together with edge slides on the spines) produce a "graphic" $\Gamma$ in the $(t, s)$-square $I \times I$.

The graphic consists of open regions $R_i$ where $E_s$ and $T_t$ intersect transversely, edges or "walls" where the two have a tangency, and cusp points where two types of tangencies cancel. As argued in [RS] only domain walls corresponding to saddle tangencies need to be tracked. Cusps and tangencies of index 2 or 0 can be erased as they amount only to births/deaths of inessential simple closed curves of intersection in $E_s \cap T_t$. The most interesting event which occurs are transverse crossings of saddle walls; at this point two independent saddle tangencies occur.

Label a region of the graphic as follows:

- Ignore circles in $T_t \cap E_s$ that are inessential in $T_t$,
- Label the region $A$ if there is a circle $a$ of $T_t \cap E_s$ so that either

  - $a$ is innermost in $E_s$ among essential circles in $T_t \cap E_s$, and the disk in $E_s$ that it bounds lies in $A$ or
  - $a$ is parallel in $E_s$ to a boundary component, and the collar between them lies in $A$
- Label a region $B$ if there are *no* such circles $a$ as above, but there is *at least one circle* $T_t \cap E_s$ that is innermost in $E_s$ among essential circles in $T_t \cap E_s$ and the disk in $E_s$ that it bounds lies in $B$.

Note that the definition of the labeling breaks symmetry: A collar of $\partial A$ is counted as if it were an innermost disk but a collar of $\partial B$ is not; and a region in which there are essential disks in both $A$ and $B$ is labeled $A$.

In the figures illustrating our argument we will distinguish between the compression-bodies $A$ and $B$ by color: pinkish (nominally red) will denote $A$, while azure (nominally blue) will denote $B$. This distinction will color regions of $E$ cut out by $T$ alternately red and blue.

For example, consider Figure 2. The bi-colored horizontal plane shows part of a component of $E$. Two parts of $T$ are shown

- a large-diameter vertical annulus, separating the visible part of $E$ into a blue unbounded region and a red 'pair of pants'; and
- an inverted-U-shaped annulus that separates a blue 1-handle of $B$ (the 'blue tube') from the part of $A$ that contains the red pair of pants.

Figure 2 shows the blue tube bounded by part of $T$ being lowered through the reducer $E$. In so doing the $(s, t)$ parameter designating the two surfaces passes from one region of the graphic to another. An astute viewer will notice a gray area at the top of the blue tube reflecting ambiguity on what might lie there: Is the top of the tube just a disk, or does a chimney filled with blue ascend through it? This could be an important question, as we will discuss shortly.

We will also use the red-blue coloring scheme on the graphic: Regions that are labeled $A$ will be colored red; those labeled $B$ will be colored blue. Skip ahead to Figure 11 to see how the coloring scheme might then appear in $I \times I$ containing the graphic. The idea of the proof of Theorem 1.6 can already be discerned in this figure: Ultimately we will walk around the outside boundary of the big red region and observe that every step corresponds to some combination of an isotopy, a bubble move or an eyeglass twist.

Return now to the ongoing argument: The first labeling rule above – ignore circles of intersection that are inessential in $T_t$ – raises a *caveat*:

When we say that an essential circle $a$ bounds a disk in $A$ (similarly for $B$), what is technically meant is that there is a planar surface $P \subset A$ whose boundary consists of $a$ and, possibly, a collection of circles that are inessential in $T$.

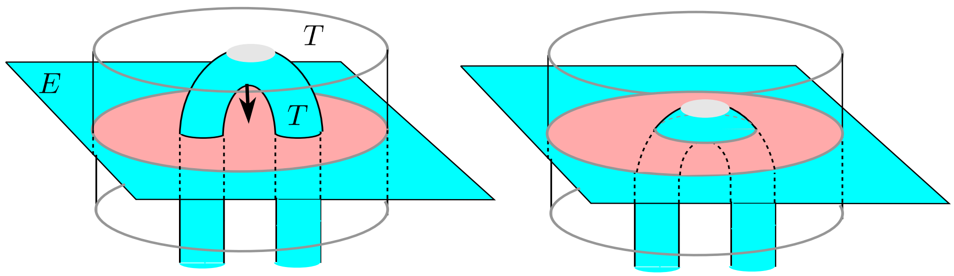


FIGURE 2. Label change or not?

As a consequence, crossing from one region of the graphic to another may change the label from $B$ to $A$ in surprising ways, as shown in Figure 2. As the saddle point in $T$ passes down through $E$, the label will change from $B$ to $A$ if the grey disk at the top is inessential in $T$. If the grey disk is essential in $T$, so $T$ ascends beyond it, the label remains $B$. The difference between the two situations cannot be determined just by examining the behavior of $E \cap T$.

## 3. LABELS AROUND THE BOUNDARY OF THE GRAPHIC

In thinking about the labeling scheme, consider first the situation near $s = 0, 1$. Since the parameterization $T \times (0, 1)$ in each case comports with $E$, any component of $E_i$ is swept out by a single circle. One side of the circle is a disk or annulus lying entirely in $A$. Thus all regions near $s = 0, 1$ are labeled $A$. Also, near $t = 0$, $T_t$ is near a spine of $A$, which must intersect each component of $E_s$, since $B$ is incompressible and $\partial_- B$ does not compress in $B$. Such an intersection point with the spine (or possibly a component of $\partial E_s$ in $\partial_- A$) means that near $t = 0$, $S_s \cap T_t$ will cut out from $E_s$ a small disk in $A$ (or a thin annulus near a component of $\partial E_s$ at $\partial_- A$). Thus the regions adjacent to $t = 0$ are again all labeled $A$.

The labeling of regions near $t = 1$ is more subtle and contains a warm-up for the general case. Because $T$ is near a spine of $B$, each circle of intersection with $E_s$ either bounds a disk in $B$ or is parallel in $B$ to a boundary component, as we have just noted. Consider first the three simple cases that can arise when $E = E_s$ is a single component:

- Suppose $E$ is a reducing sphere for $M$. Since each component of $E \cap T$ bounds a small disk in $B$, a region will be labeled $A$ if and only if there is a single circle of intersection, in other words, if and only if $E$ is a reducing sphere for $T$.

- If $E$ is a $\partial$-reducing disk whose boundary lies in $\partial_- B$ then $T$ intersects $E$ in at least a circle parallel to $\partial E$ and the collar lies in $B$. But if there is any other circle of intersection, the small disk it bounds lies in $B$, so there are no disks in $A$. Again, a region will be labeled $A$ if and only if $E$ is a $\partial$-reducing disk for $T$.
- Suppose $E$ is a $\partial$-reducing disk for $M$ whose boundary lies on $\partial_- A$. Since $\partial_- A$ is incompressible in $A$, not all of $E$ can lie in $A$ so there is at least one circle of intersection. There is exactly one circle if and only if that circle is $\partial$-parallel in $A$ and so labels the region $A$. So, once again, the region is labeled $A$ if and only if $E$ is a $\partial$-reducing disk for $T$.

Hence we have (see Figure 11):

**Proposition 3.1.** *Suppose $E$ has a single component and the regions adjacent to the side $t = 1$ are all labeled $A$. Then $E_0$ and $E_1$ are equivalent $T$-reducers.* □

When $E$ has many components, the argument is more complicated, since our labeling scheme assigns the label $A$ if just one component of $E_s$ is a $T$-reducer. So at this point we make a crucial inductive assumption:

**Assumption 3.2.** *Theorem 1.6 is true in all cases for which the genus of the splitting surface is less than the genus of $T$.*

Note that Theorem 1.6 is more or less obvious when $genus(T) = 1$.

Suppose in a region of the graphic a component of $E = E_s$ is a $T$-reducer $\overline{E}$. Then there is a natural way of isotoping $T$ rel $\overline{E}$ to a solution for all of $E$: Reduce or $\partial$-reduce $(M, T)$ along $\overline{E}$ to obtain a new Heegaard split manifold $M' = A' \cup_{T'} B'$ (disconnected if $E$ is separating). Each component of $T'$ has genus less than $genus(T)$. By [Sc], $T'$ can be isotoped in $M'$ so that the family $E - \overline{E}$ in $M'$ is $T'$-reducing, and by Assumption 3.2 this solution for $(M', T')$ is well-defined up to congruence. This solution, together with $\overline{E}$, constitutes a natural solution to $(M, T)$. Call it the *solution generated by $\overline{E}$.*

**Lemma 3.3.** *Suppose $\overline{E}'$ is another $T$-reducer in $E$. Then the solution generated by $\overline{E}'$ is congruent to that generated by $\overline{E}$.*

*If $E$ is a $T$-reducing system, then it is the solution generated by any of its members.*

*Proof.* The solution obtained by compressing (or $\partial$-compressing) along both $\overline{E}'$ and $\overline{E}$ is congruent to that generated by either, per Assumption 3.2. □

In view of Lemma 3.3 we can simply call such a solution in the region *internally generated* without naming the component of $E$ that generates it.

**Lemma 3.4.** *Suppose two regions of the graphic, adjacent along an edge of each, have internally generated solutions. Then these solutions are congruent.*

*Proof.* Let $\overline{E}$ and $\overline{E}'$ be generators in adjacent regions $R, R'$ respectively. If $\overline{E} = \overline{E}'$ congruence follows by definition, so we assume $\overline{E} \neq \overline{E}'$. The edge between regions $R$ and $R'$ represents $E$ passing through a single saddle tangency with $T$, a point that may lie on $\overline{E}$ or $\overline{E}'$ but not both. Thus at least one of the two is a generator in both regions, from which congruence of solutions follows by Lemma 3.3. □

Return now to the setting for Theorem 1.6 and we have:

**Proposition 3.5.** *Suppose the regions adjacent to the side $t = 1$ are all labeled $A$. Then $E_0$ and $E_1$ are congruent.*

*Proof.* The label $A$ implies that each region adjacent to the side $t = 1$ has a self-generated solution. Lemma 3.4 ensures that the congruence class of the solution doesn't change as we move along the side $t = 1$ from $E_0$ to $E_1$. □

## 4. A forbidden labeling around a vertex

Focus now how labels behave around a vertex in the interior of the graphic $\Gamma$. Such a vertex corresponds to a position of $T = T_t$ in which it has two simultaneous tangency points with $E = E_s$. The non-trivial cases arise when both points of tangency lie on a single component $\overline{E} \subset E$. If $\overline{E}$ is a disk, a simple combinatorial argument shows that there are 15 possible configurations of these tangency points, shown in Figure 3. The same diagram can be used when $\overline{E}$ is a sphere, but far fewer panels are needed because of the extra symmetry this introduces. For example, panels 10, 11 and 12 are the same in a sphere, as are 13 and 14. We will proceed assuming $\overline{E}$ is a disk; if it is a sphere, just delete an open disk near a point in $A$, converting it to an $A$-disk, and apply the arguments there.

There are typically many more circles in $\overline{E} \cap T$ than are shown in the panels of Figure 3; these only show the components containing tangency points. The two tangency points will be denoted $\rho = \rho_{\pm}$; the 4 quadrants near it correspond to the 4 ways of resolving the tangencies, each by perturbing $T$ slightly near $\rho$.

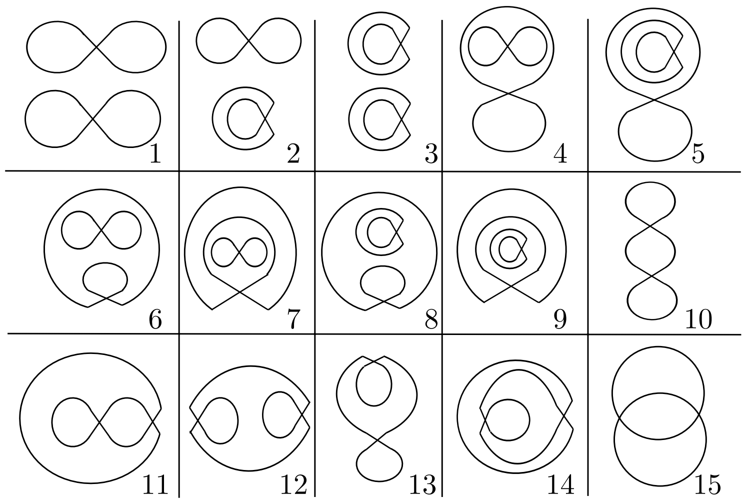


FIGURE 3. At a vertex in the graphic $\Gamma$

The picture in $T$ can be more complicated than these panels suggest. For example, panel 15 might look like Figure 4 in $T$.

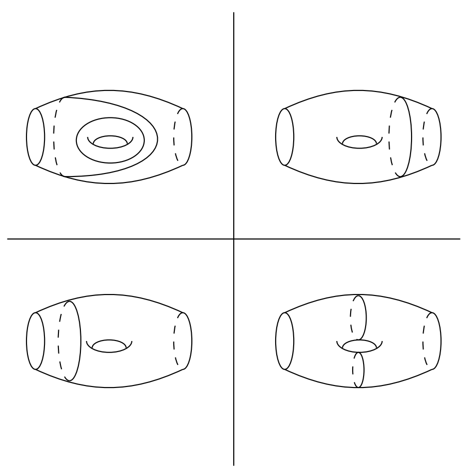

FIGURE 4. How panel 15 might be resolved in $T$

**Proposition 4.1.** *No vertex in the graphic is surrounded by labeling pattern* $\dfrac{A|B}{B|A}$.

*Proof.* The simply connected components of $\overline{E} - T$ that are shown in Figure 3 will each become a disk in some resolution of the tangency points; if such a disk contains no other essential circles of $T$ and is essential in $A$ we will call the component an $A$-leaf component and the disk it becomes an $A$-leaf. (The terminology is explained in the next section.) A component of $\overline{E} - T$ shown in the diagram is incident either to one of $\rho_\pm$ or to both.

**Lemma 4.2.** *If an $A$-leaf component is incident to only one of $\rho_\pm$, then the labeling around the vertex is not* $\begin{array}{c|c} A & B \\ \hline B & A \end{array}$

*Proof.* Resolve the single tangent point so that the component becomes an $A$-leaf. Either resolution at the other tangent point (these corresponding to two adjacent quadrants in $\Gamma$) leaves the $A$-leaf intact, so these two adjacent quadrants both get labeled $A$. □

**Lemma 4.3.** *At a vertex in $\Gamma$ with surrounding labels* $\begin{array}{c|c} A & B \\ \hline B & A \end{array}$ *the two $A$ labels cannot both come from $A$-leaf components.*

*Proof.* Following Lemma 4.2 each $A$-label must come from an $A$-leaf component that is incident to both $\rho_\pm$. This eliminates panels 1 through 9. Moreover, the two $A$-leaf components arise from different resolutions on each tangency point, since they are diagonally opposite. Only panels 11 and 15 have two 2-vertex components, but in panel 11 they are adjacent and so they can't both lie in $A$. In panel 15, a resolution of the tangencies that turns an $A$-leaf component into an $A$-leaf, when reversed, only gives disk components that contain points in $B$. □

This would seem to prove Proposition 4.1, until we remember that $A$-labels may arise in another way, as shown in Figure 2. For example in panel 4, the annulus component of $\overline{E} - T$ that is shown might lie in $A$, and the interior pair of circles might bound parallel disks in $B$, but when the pair is resolved into a single circle, it is inesential in $T$. Call such a component of $E - T$ in a panel an $A$-annulus.

**Lemma 4.4.** *At a vertex in $\Gamma$ with surrounding labels* $\begin{array}{c|c} A & B \\ \hline B & A \end{array}$*, neither $A$ label can come from an $A$-annulus.*

*Proof.* Suppose one of the quadrants gets its $A$-label via an $A$-annulus, as described. Such a component could arise in panels 4, 5, 6, 7, 8 and 9. In order to have the given labeling the opposite resolution at both $\rho_\pm$ should again generate an $A$-label. The label can't come from the same $A$-annulus since its inner boundary is no longer adjacent to an inessential disk. Thus the $A$-label must come from an $A$-leaf component and, by observation, each $A$-leaf component is incident to only one of $\rho_\pm$. This contradicts Lemma 4.2 $\square$

A final way in which $A$-labels might arise is via 'hidden components'. Remember that the panels only show components of $T \cap E$ that are incident to the tangency points. Imagine a circle $c$ of $E \cap T$ bounding a disk that contains the pair of components shown in panel 1. If both of $\rho_\pm$ resolve as in Figure 2, the resulting disk bounded by $c$ could generate a label $A$. The hidden component here is the 'pair of pants' bounded by $c$ and the two components in the panel; it is hidden because $c$ does not appear in the panel. But it is easy to see that hidden pairs of pants (which could arise in panels 1, 2, and 3) can't possibly give rise to the labeling $\begin{array}{c|c} A & B \\ \hline B & A \end{array}$.

Hidden annuli require more thought. Suppose a circle component $c$ of $E \cap T$ cobounds an annulus with a component $X$ from one of the panels. It may be possible to resolve the tangency points in $X$ so that the end of the annulus at $X$ bounds an inessential circle, so it might in this way be possible for $c$ to generate a label $A$. By the argument of Lemma 4.2 such a hidden annulus can be part of a labeling $\begin{array}{c|c} A & B \\ \hline B & A \end{array}$ only if the end at $X$ is incident to both tangency points $\rho_\pm$. This immediately rules out panels 1 through 9 as well as 11 and 14. The end at $X$ must also have the property that the opposite resolution at both $\rho_\pm$ will still give rise to an $A$-disk, and the new $A$-disk must be incident to both $\rho_\pm$, by Lemma 4.2. This eliminates panels 12 and 13. Panel 10 won't work: each leaf component shown has points in $B$, since the hidden annulus lies in $A$. Finally, these requirements can be fulfilled in panel 15 only if the middle sector lies in $A$ and the two other sectors are inessential in $B$. But in that case, there could be no label $B$ in any quadrant.

A technical note: our labeling convention assigns the label $A$ also if one of the regions in $\overline{E} - T$ is a collar of the boundary in an $A$-disk. The argument in this case is identical to that given above for the case in which there is a hidden circle that completely surrounds the figure in each panel. $\square$

## 5. From weakly reducing to reducing

In [CG] Casson-Gordon introduced the notion of a weakly reducible Heegaard splitting, rejuvenating Heegaard theory. They showed that if there are disjoint essential disks in $A$ and $B$, then simultaneous compression on a maximal family of such disjoint disks in $A$ and $B$ will produce either a reducing sphere for $T$ or an incompressible surface or both. In considering uniqueness, the choice of a 'maximal family' is problematic, since such a family is far from unique. In this section we avoid this problem of choice, by deriving from the entire pattern of circles $T \cap E$ in $E$ a recipe to move from a weakly reducing system for $T$ to a full $T$-reducing system, in a series of steps that is well-defined up to congruence.

Suppose $E$ is an $M$-reducing system for $M = A \cup_T B$. $E$ will be called a *weak solution* if, among the components of $E - T$, there are $\partial$-reducing disks for both $A$ and $B$. Continuing under Assumption 3.2, we will describe a natural algorithm that transforms a weak solution $E$ into a $T$-reducing system, an algorithm that is well-defined up to congruence.

Denote by $\mathcal{D}_A$ (resp $\mathcal{D}_B$) the collection of all disk components of $E$ cut off by $T$ that lie in $A$ (resp $B$). Figure 5 illustrates the idea in a $B$-disk component of $E$.

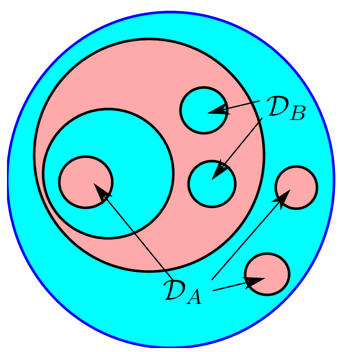


FIGURE 5. The view in $E$

Consider the surface $T_c \subset M$ obtained by compressing $T$ along $\mathcal{D}_A \cup \mathcal{D}_B$. $T_c$ divides $M$ into two (possibly disconnected) 3-manifolds $M_A$ and $M_B$. Imagine thickening $T_c$ by expanding it into a bi-collar as shown in Figure 6. This would induce Heegaard splitting surfaces $T_A \subset M_A$, obtained from the original $T$ by compressing only along $\mathcal{D}_A$ and then pushing towards the $A$-side of the bicollar. The symmetric construction gives a Heegaard splitting surface $T_B$ in $M_B$,

Both $T_A$ and $T_B$ have lower genus than $T$, so our inductive Assumption 3.2 applies. In particular, given any $M_A$-reducing system of disks and spheres, $T_A$ can be isotoped, uniquely up to congruence, so that the system becomes $T_A$ reducing (and similarly for $(M_B, T_B)$). Such

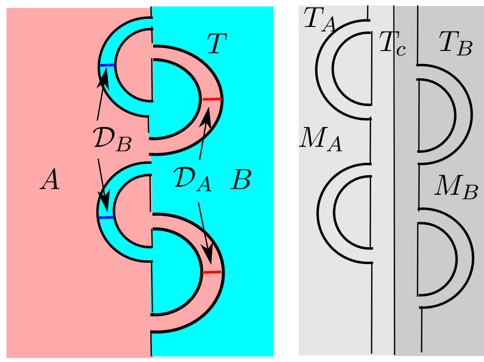


FIGURE 6. The view in $M$, and a mental image

an isotopy of $T_A$ can be described [Sc] as a sequence of handle-slides of and over the handles whose cocores are the $\mathcal{D}_A$ disks. But these same handle-slides could have been done on and over the handles as they actually lie on $T_c$, avoiding (by general position) the attaching disks for the handles on the other side, those with cocores the disks $\mathcal{D}_B$. In thinking of this as an isotopy of the original Heegaard surface $T$, the exact trajectory which the handle-slides follow across $T_c$ in order to avoid the disks $\mathcal{D}_B$ is, for our purposes, unimportant: one choice can be moved to another by eyeglass twists of $T$. The symmetric argument applies to $M_A$. The upshot is:

**Proposition 5.1.** *Suppose $E_A \subset M_A$ and $E_B \subset M_B$ are embedded collections of $\partial$-reducing disks whose boundaries lie on $T_c \subset M$. Then there is an isotopy of $T$, keeping $T_c$ setwise fixed, to a position in which the boundary of each disk remains unchanged in $T_c$ and the interior of each disk is disjoint from $T$. The isotopy is well-defined up to congruence.* □

Consider a component $P$ of $E - T$ which is next to innermost, i.e. all but one of its boundary component is an innermost circle in $E \cap T$, and so each bounds a disk in $\mathcal{D}_A$ (or each bounds a disk in $\mathcal{D}_B$). Then the exceptional boundary component $\partial_0 P$ lies in $T_c$ and bounds a disk $D_P$ in $M_B$ (or $M_A$), through which the 1-handles dual to $\mathcal{D}_A$ or $\mathcal{D}_B$ may pass.

The algorithm is then:

(1) Apply Proposition 5.1 to the collection of all such components $P$ of $E - T$, isotoping $T$ without changing $T_c$ so that afterwards the interior of each disk $D_P$ is disjoint from $T$.

(2) Add each such disk $D_p$ to $\mathcal{D}_B$ or $\mathcal{D}_A$ as appropriate, compressing $T_c$ to $T_c'$
(3) Repeat the process until at least one component $\overline{E}$ of $E$ is a $T$-reducer.
(4) The output is the solution generated by $\overline{E}$.

It will be important for its application that the algorithm is robust: a minimal change in input information will result in the same output. To understand more fully how the algorithm operates, we can describe it schematically.

The pattern of circles $T \cap E$ in $E$ defines a tree in each component $\overline{E}$ of $E$, with a vertex for each component of $E - T$ and an edge connecting two components if there is a single circle of $E \cap T$ between them. The tree has a natural base or root when $\overline{E}$ is a disk, namely the component of $\overline{E} - T$ containing the boundary. Let $Y$ denote the forest that is the whole collection of trees. The innermost disks of $E - T$ can be thought of as leaves in the forest $Y$. One measure of the complexity of each tree is the diameter of the tree, when $\overline{E}$ is a sphere, or the height of the tree when $\overline{E}$ is a disk. (Tree height is the edge-distance from the root of the tree to the most distant leaf. See Figure 7).

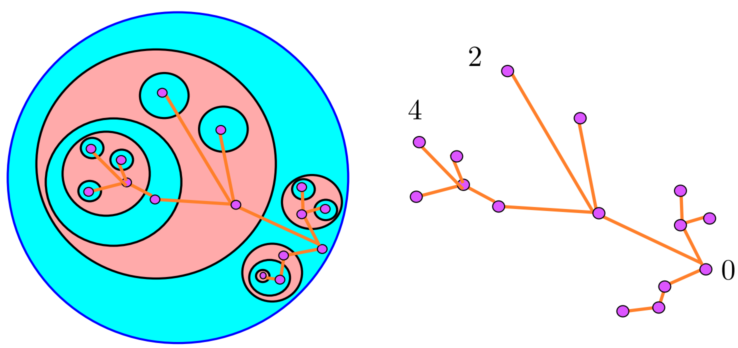


FIGURE 7. Tree height is 4

The $B$-leaves of $Y$ correspond to $\mathcal{D}_B$, and $A$-leaves to $\mathcal{D}_A$. The *branch-structure* $Y_c$ of $Y$ is obtained from $Y$ by removing all leaves; alternatively, it is the forest determined by the circles $T_c \cap E$ in $E$.

The leaves of the branch structure correspond to the "second-innermost" circles in the algorithm described above or, in terms of the original forest, they are the outermost forks. Applying the algorithm described above replaces the original leaves with new leaves, corresponding to what were originally outermost forks. Since we have no control over how the 1-handles of $T_A \subset M_A$ and $T_B \subset M_B$ intersect the non-disk components of $E - T_c$, leaves may also sprout from every other vertex

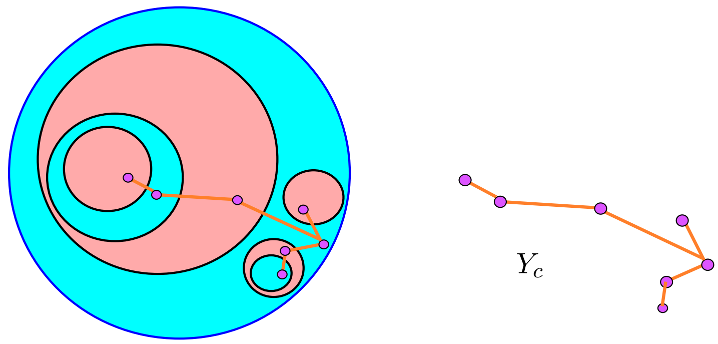


FIGURE 8. Branch structure from $T_c \cap E$

in $Y_c$. But one iteration of the algorithm described will decrease the height (or diameter) of each tree. This is shown schematically in Figure 9, where new leaves sprout in non-disk components of $E - T_c$.

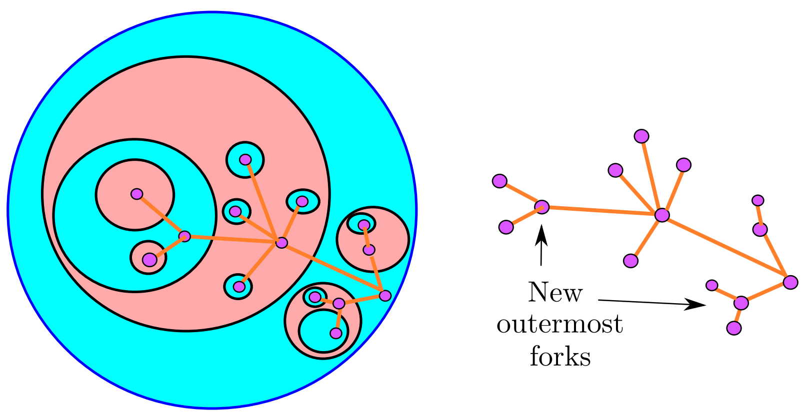


FIGURE 9. De/refoliation of $B$-leaves; height is now 3

Since $genus(T_A) < genus(T)$, the inductive hypothesis says that the new $A$-leaves of $Y'$ (the ones contributed by what were previously outermost forks) are well-defined in $T_A$ up to congruence, so similarly well-defined in $T$. Add them to $\mathcal{D}_A$, compressing $T_c$ into $A$ and effectively $\partial$-reducing $T_A \subset M_A$. See Figure 10, Call the augmented collection $\mathcal{D'}_A$.

A similar argument applies in $M_B$, resulting in new surfaces $T'_A, T'_B$ and $T'_c$, the latter dividing $M$ now into $M'_A, M'_B$.

Continue with the algorithm until the height or diameter in some component $\overline{E}$ is 1. (We pause to note the last step). $T$ now divides $\overline{E}$ into a planar surface in $B$, say, (which is incident to $\partial\overline{E}$ if $\overline{E}$ is a disk) and a collection of disks in $A$, all of them lying in a submanifold of $M$

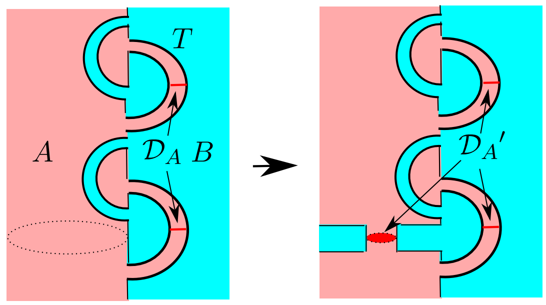


FIGURE 10. New $A$-leaves added to $\mathcal{D}_A$

with a lower genus Heegaard splitting surface. Once again apply the Strong Haken Theorem [Sc] together with the inductive hypothesis on this lower genus splitting to isotope $T$ so that $\overline{E}$ is a solution. This completes the algorithm.

## 6. Appendages and convergence

Section 5 described an algorithm which proceeds by well-defined iteration from a weak solution of an $M$-reducing system $E$ into a $T$-reducing system for $M = A \cup_T B$, at each step augmenting the number of disjoint weakly reducing disks.

**Definition 6.1.** *Two weak solutions for $E$* converge *if the algorithm results in congruent $T$-reducing systems.*

The term is meant to convey that, after perhaps some iterations in the algorithm, the two weak solutions may become indistinguishable, even before they each become full $T$-reducing systems.

An important example of convergent weak solutions begins with this easy corollary of Theorem 1.6:

**Corollary 6.2.** *With hypotheses as in Theorem 1.6, suppose that $\mathcal{D} \subset M$ is a collection of $\partial$-reducing disks for $T$ that is disjoint from the two systems $E_i$. Then the sequence of eyeglass twists and bubble moves creating the congruence may be taken to be disjoint from $\mathcal{D}$.*

*Proof.* The Corollary follows from Assumption 3.2, which allows us to apply Theorem 1.6 to the lower-genus Heegaard split manifold $(M', T')$ obtained from $(M, T)$ by $\partial$-reducing along the disks $\mathcal{D}$. □

Suppose $E$ is an $M$-reducing system whose intersection with $T$ is a weak solution, and $\mathcal{D}_A \subset M_A$ and $\mathcal{D}_B \subset M_B$ are the leaves, as described above.

**Proposition 6.3.** *Suppose* $(D, \partial D) \subset (M_B, \partial M_B)$ *(resp* $(M_A, \partial M_A)$ *is a properly embedded essential disk that is disjoint from* $E$*. Then the solution to* $E$ *given by the algorithm is unaffected by adding* $D$ *to* $\mathcal{D}_B$ *(resp* $\mathcal{D}_A$*) at the start.*

*Proof.* We will show that in both $M_A$ and $M_B$ the algorithm is unaffected by the addition of $D$ to $\mathcal{D}_B$. (Of course if $D$ is parallel in $T$ to an element in $\mathcal{D}_B$ there is nothing to show.)

This follows from Corollary 6.2 for $M_B$, since $D$ can be regarded as a $\partial$-reducing disk in $M_B$.

In $M_A$ the addition of $D$ changes the status of its dual 1-handle from being part of the boundary of $M_A$ to being a 1-handle in the splitting of $M_A$ by $T_A$. This is a profound change, but the original algorithm describes passing 1-handles past $\partial D$, i. e. over the new 1-handle, and this can still be done. Since $D$ is disjoint from $E$ the algorithm never requires the new 1-handle to move. And so the algorithm can proceed step after step, never noticing the new 1-handle, until a solution is achieved. $\square$

**Definition 6.4.** *Because of its inactivity in the proof, a disk* $D$ *as in Proposition 6.3 is called an* appendage disk*.*

**Lemma 6.5.** *Suppose there are multiple components in a weakly reducing system* $E$*, and the system* $E'$ *obtained by removing one of them remains weakly reducing. Then the solution provided by the algorithm is unaffected by the removal. (That is, the two weak solutions converge.)*

*Proof.* The case is in which $E$ consists of only two components $E = \overline{E}_\pm$ is definitive; the general case is no harder. At the beginning of the algorithm on $E$, $M_A$ and $M_B$ are defined by compressing $T$ along the disk components of $E - T$. Now remove $\overline{E}_-$ and note that the algorithm applied to $\overline{E}_+$ would call for compressing only along the disk components of $\overline{E}_+ - T$. But the outcome of that algorithm is unaffected by further compressing by disk components on $\overline{E}_-$, by Proposition 6.3. So, at the initial stage, there is no difference between the eventual solutions. Just continue in this manner, using how the algorithm behaves on the 'virtual' componoent $\overline{E}_-$ to present extra disks to be included as appendages (under Proposition 6.3) as the algorithm is applied to $\overline{E}_+$ alone.

Eventually the parallel algorithms stop, when one of $\overline{E}_\pm$ becomes a $T$-reducer. If it stops because $\overline{E}_+$ is a $T$-reducer, then we have shown that the solutions on $E$ and $\overline{E}_+$ result in the same solution, as required. If it stops because $\overline{E}_-$ is a $T$-reducer, then the algorithm for $E$ declares that a solution consists of $\overline{E}_-$, together with *any* solution for $\overline{E}_+$ in the manifold obtained by reducing (or $\partial$-reducing) $(M, T)$ along $\overline{E}_-$. An example of such a solution is given by the output of the algorithm further played out on $\overline{E}_+$. □

Suppose an edge in the graphic lies between a region labeled $A$ and a region labeled $B$. The edge indicates a saddle tangency of $E$ with $T$. Let $\overline{E}$ be the component of $E$ that contains the saddle tangency. Let $\overline{E}_A, \overline{E}_B$ be slight push-offs of $\overline{E}$ corresponding to the regions labeled $A$ and $B$ respectively.

**Definition 6.6.** *The* edge weak solution *is the weak solution obtained by deleting $\overline{E}$ from $E$ and replacing with $\overline{E}_A \cup \overline{E}_B$.*

**Proposition 6.7.** *An edge weak solution converges with a weak solution (if any) determined by either adjacent region.*

*Proof.* Suppose there is a weak solution for the adjacent region labeled $A$ (resp $B$). That weak solution is obtained from the edge weak solution by just deleting $\overline{E}_B$ (resp $\overline{E}_A$). The result then follows from Lemma 6.5. □

## 7. Return to the graphic

Return now to the proof of Theorem 1.6 by examining the graphic more closely, inspired by [FS, Subsection 4.5] and adopting similar conventions. An edge in the graphic that lies between a region labeled $A$ and a region labeled $B$ will be called a *border edge*. Following Section 4, any vertex in the graphic that is incident to a border edge is incident to exactly two border edges (or to the boundary of the graphic). Thus the collection of border edges constitute a properly embedded 1-manifold in the graphic that separates $A$ regions from $B$ regions.

We have shown earlier that three sides of the graphic ($s = 0, 1$ and $t = 0$) are adjacent to $A$-regions. Since the union of the three sides is connected, there is a single component $\mathcal{A}$ of the complement of the border edges that contains all three sides in its boundary. $\mathcal{A}$ consists entirely of regions labeled $A$. See Figure 11.

We focus on the 1-manifold component $C$ of $\partial\mathcal{A}$ that contains the three sides $s = 0, 1$ and $t = 0$. If $C$ contains the fourth side $t = 1$ then per Proposition 3.5 we are done, so our interest will focus on the arc in $C$ whose ends are at the corners $s \in \{0, 1\}, t = 1$ of the graphic, or more specifically, the border arcs that $C$ contains. (See Figure 11).

It follows from Proposition 6.7 that the lowest border edge (i. e. minimal $s$) on the lowest border arc in $C$ generates the solution $E_0$ and the highest border edge on the highest border arc of $C$ generates the solution $E_1$. If we can show that the weak solutions given by successive

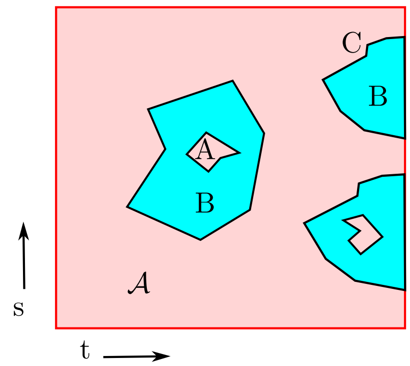


FIGURE 11. Graphic labels: red = A, blue = B

border edges in $C$ always generate congruent solutions, then we will have shown that the solutions $E_0$ and $E_1$ are congruent, as required. So we examine how passing through a vertex of the graphic that lies on a border arc affects the weak solutions given by the incident edges. We will show the following, from which Theorem 1.6 then follows.

**Proposition 7.1.** *At any vertex in a border arc, the weak solutions given by the incident border edges converge.*

*Proof.* There is an important feature distinguishing between the two diagonals in a labeling diagram around a vertex in $\Gamma$. Put $T$ in the position determined by the vertex of $\Gamma$, so that $E = E_s$ and $T = T_t$ are tangent at two points $\rho = \rho_\pm$. We assume that $\rho_\pm$ lie on the same component $\overline{E}$ of $E$; if they lie on different components of $E$ the proof is easier.

Let $\overline{E}_\pm$ be slight push-offs of $\overline{E}$ to each of its sides. Then the disks $\overline{E}_\pm$ correspond to positionings of $\overline{E}$ that lie in diagonally opposite quadrants of the graphic, since in moving from one to the other, the resolution of each of the tangencies at $\rho_\pm$ is changed. The curves of $\overline{E}_+ \cap T$ and $\overline{E}_- \cap T$ are visibly disjoint in $T$, since the disks $\overline{E}_\pm$ are disjoint in $M$. Call this the *aligned* diagonal. (The other diagonal was called the *dangerous diagonal* in [FS]. In Figure 4 the antidiagonal is aligned and the main diagonal is dangerous.)

The argument will be symmetric in $A$ and $B$ and also indifferent to symmetries of the quadrants about the vertex, so, following Proposition 4.1, there are just two cases to consider, corresponding to the labelings: $\begin{array}{c|c} A & A \\ \hline A & B \end{array}$ and $\begin{array}{c|c} A & A \\ \hline B & B \end{array}$.

**Case 1:** The labelings around the vertex are $\begin{array}{c|c} A & A \\ \hline A & B \end{array}$; and the antidiagonal $\begin{array}{c|c} & \bullet \\ \hline \bullet & \end{array}$ is aligned.

In this case, replace the component $\overline{E}$ in $E$ by three parallel components, namely, the two components $\overline{E}_{\pm}$ representing the antidiagonal, and a component $\overline{E}_B$ representing the quadrant labeled $B$. Call the resulting system $E^+$.

Deleting exactly $\overline{E}_+$ from $E^+$ gives the weakly reducing system for one boundary edge and deleting exactly $\overline{E}_-$ gives the weakly reducing system for the other boundary edge. Now apply Lemma 6.5: both of these converge with the system $E^+$.

**Case 2:** The labelings around the vertex are $\begin{array}{c|c} A & A \\ \hline A & B \end{array}$; and the main diagonal $\begin{array}{c|c} \bullet & \\ \hline & \bullet \end{array}$ is aligned.

In a similar fashion, replace $\overline{E}$ in $E$ by three parallel components: $\overline{E}_+$ representing the upper left quadrant, $\overline{E}_-$ representing the lower right quadrant and a component $\overline{E}^{12}$ representing the upper right quadrant. Call the resulting system $E^{12}$.

Deleting exactly $\overline{E}_+$ from $E^{12}$ gives the weakly reducing system for the right boundary edge; deleting exactly $\overline{E}_{12}$ gives a weakly reducing system we call here the *diagonal system.* By Lemma 6.5 the two systems converge.

Now repeat the argument using the system $E^{21}$ obtained by replacing $\overline{E}^{12}$ with a component $\overline{E}^{21}$ representing the lower left quadrant. The argument shows that the diagonal system also converges to the weakly reducing system for the lower boundary edge. Therefore the weak solutions representing the two boundary edges converge to each other.

**Case 3:** The labelings around the vertex are $\begin{array}{c|c} A & A \\ \hline B & B \end{array}$. In this case we may as well assume the main diagonal is aligned. Then a minor variant of the argument for Case 2 suffices. $\square$

## References

[Bo] F. Bonahon, Cobordism of automorphisms of surfaces. *Ann. Sci. École Norm. Sup.* **16** (1983) 237?270.

[CG] A. Casson and C. McA. Gordon, *Reducing Heegaard splittings*, Topology and its applications, **27** (1987), 275-283.

[FS] M. Freedman and M. Scharlemann, *Powell moves and the Goeritz group*, arXiv:1804.05909.

[Ha] W. Haken, *Some results on surfaces in 3-manifolds*, Studies in Modern Topology, Math. Assoc. Am., Prentice Hall, 1968, 34-98.

[Ha1] A. Hatcher, Homeomorphisms of sufficiently large $P^2$-irreducible 3-manifolds, *Topology* **11** (1976) 343-347.

[Ha2] A.. Hatcher, A proof of the Smale Conjecture, *Annals of Mathematics*, **117** (1983) 553–607.

[Po] J. Powell, Homeomorphisms of $S^3$ leaving a Heegaard surface invariant, *Trans. Amer. Math. Soc.* **257** (1980) 193–216.

[R] Y. Rieck, A proof of Waldhausen's uniqueness of splittings of S3 (after Rubinstein and Scharlemann). Workshop on Heegaard Splittings, 277–284, Geom. Topol. Monogr., 12, Geom. Topol. Publ., Coventry, 2007.

[RS] H. Rubinstein, Hyam and M. Scharlemann, Comparing Heegaard splittings of non-Haken 3-manifolds. *Topology* **35** (1996), 1005–1026.

[Sc] M. Scharlemann, A Strong Haken Theorem, Arxiv 2003.08523.

[ST] M. Scharlemann and A. Thompson, *Heegaard splittings of (surface)* $\times$ *I are standard*, Math. Ann., **295** (1993), 549-564.

[Wh] J. H. C. Whitehead, On equivalent sets of elements in a free group, *Ann. of Math.* **37** (1936), 782–800.

[Wa] F. Waldhausen, Heegaard-Zerlegungen der 3-Sphäre. *Topology* **7** (1968) 195-203.

Microsoft Station Q, University of California, Santa Barbara, CA 93106-6105, and, Mathematics Department, University of California, Santa Barbara, CA 93106-3080 USA

*Email address*: michaelf@microsoft.com

Martin Scharlemann, Mathematics Department, University of California, Santa Barbara, CA 93106-3080 USA

*Email address*: mgscharl@math.ucsb.edu