AN ALGORITHM FOR CELLULAR MAPS OF CLOSED SURFACES

Autor(en): **DICKS, Warren / Glover, H. H.**

Objekttyp: **Article**

Zeitschrift: **L'Enseignement Mathématique**

Band (Jahr): **43 (1997)**

Heft 3-4: **L'ENSEIGNEMENT MATHÉMATIQUE**

PDF erstellt am: **21.07.2024**

Persistenter Link: https://doi.org/10.5169/seals-63277

L'Enseignement Mathématique, t. 43 (1997), p. 207–252

AN ALGORITHM FOR CELLULAR MAPS OF CLOSED SURFACES

by Warren DICKS and H. H. GLOVER

ABSTRACT. The purpose of this article is to use diagrammatic methods to give proofs, accessible to algebraists, of some important topological results of H. Kneser, A. L. Edmonds, and R. Skora; we then describe some consequences for homomorphisms between surface groups. Cellular maps between two-dimensional CW-complexes can be represented by diagrams which, in turn, can be interpreted algebraically in terms of fundamental groupoids. For diagrams representing cellular maps between closed surfaces, we show how to apply certain homotopy equivalences algorithmically to obtain a normal-form map, which is a branched covering, or a pinching followed by a covering, or a map which collapses a graph of punctured spheres to a graph immersed in the one-skeleton of the target surface. We then indicate how the algorithm can be expressed entirely in terms of formal manipulations with presentations of surface groupoids, yielding algebraic proofs of results about homomorphisms between surface groups.

1. INTRODUCTION

We begin by recalling some basic concepts.

1.1. DEFINITIONS. Let β be a map between closed surfaces (without boundary).

Then β is a *branched covering* if deleting finitely many points from the source and from the target yields a covering.

We say that β is a (possibly trivial) *pinching* if it is obtained by collapsing, to a point, a compact subsurface with a single boundary component.

The (geometric) *degree* of β, denoted $\mathcal{G}(\beta)$, is the least non-negative integer d such that there is a map β' homotopic to β, such that the inverse image under β' of some 2-disk consists of d 2-discs, each mapped homeomorphically by β' to the chosen disk.

Results of H. Kneser [11, p. 354], A.L. Edmonds [3], and R. Skora [16] show that if $\mathcal{G}(\beta)$ is nonzero, then β is homotopic either to a branched covering or to the composite of a pinching followed by a covering. In the case where $\mathcal{G}(\beta)$ is zero, β is homotopic to a map which is not surjective. Thus one of the following holds:

(a) β is homotopic to a map which is not surjective;

(b) β is homotopic to the composite of a pinching followed by a covering;

(c) β is homotopic to a branched covering.

In case (a), $\mathcal{G}(\beta) = 0$, and in case (b) (resp. (c)), $\mathcal{G}(\beta)$ is given by the degree of the covering (resp. branched covering).

This allows one to compute the degree via homological means, which is the essence of a classical result of Kneser [10], [11]. Edmonds and Skora further discuss cases where the surfaces are not necessarily closed, but we wish to restrict our attention to the closed case.

The main purpose of this article is to prove these Kneser-Edmonds-Skora results using diagrammatic techniques developed by van Kampen, Lyndon, and Ol'shanskii. We shall give an algorithm which applies homotopy equivalences to a cellular map between closed surfaces, and yields a map in normal form, that, in the non-zero degree case, is a branched covering, or, after pinching, is a covering, while, in the degree zero case, the source surface is expressed as a union of spheres with various punctures based at the poles, and these punctured spheres are collapsed to arcs, to give a graph immersed in the one-skeleton of the target surface. Recall that a graph map *immersion* is a locally injective graph map, so that the induced map of fundamental groups is injective. In particular, the algorithm yields the degree of the map.

In the non-zero degree case, we then have the Kneser-Edmonds-Skora result, and, in the zero degree case, we recover preliminary steps towards results previously obtained by several authors, notably Zieschang [17], Edmonds, Skora, Ol'shanskii [15], and Grigorchuk and Kurchanov [7]. The present article is very much in the spirit of Ol'shanskii's article.

The proofs by Edmonds and Skora are brief, simple, direct, and essentially algorithmic, but are not readily expressible in algebraic terms. Our proof, although substantially longer, has the feature that it deals throughout with closed surfaces, without cutting them up, and uses elementary homotopy operations which readily lend themselves to algebraic intrepretation. So we claim that we have fulfilled our main objective of giving algebraic proofs of the substantial group-theoretic consequences of these topological theorems. There is a natural motivation to have algebraic proofs of algebraic results, especially when they are obtained topologically, and our work fits into this scheme in a useful way; for example, it can be used in the (algebraic) proof of Theorem 4.9 of [2].

To give an idea of the sort of algebraic consequences of the algorithm, it is convenient to introduce some terminology.

1.2. DEFINITIONS. Recall that a group G is called a *surface group* if G is the fundamental group of a closed surface, or, equivalently, G has a *surface group presentation*, by which we mean a one-relator presentation $\langle S \mid r\rangle$ such that r is cyclically reduced, and each element of S occurs exactly twice in r, with exponent 1 or -1, and the face-adjacency relation on $S \cup S^{-1}$ has only one equivalence class. Here face-adjacency is the equivalence relation determined by identifying $s_1^{\epsilon_1}$ and $s_2^{\epsilon_2}$ whenever $s_1^{\epsilon_1} s_2^{-\epsilon_2}$ occurs in the cyclic expression of r.

It follows that G is a surface group if and only if G has a presentation

$$\left\langle x_1, y_1, \ldots, x_n, y_n, z_1, \ldots z_m \;\middle|\; (x_1, y_1) \cdots (x_n, y_n) z_1^2 \cdots z_m^2 \right\rangle,$$

for some non-negative integers m, n, and here one can arrange that either m or n is zero. Recall that (x, y) denotes $\bar{x}\bar{y}xy$, where overlines denote inverses.

There is an associated *orientation map* $\epsilon\colon G \to \{\pm 1\}$ which, for the latter presentation, sends the x_i and the y_i to 1, and the z_j to -1. The kernel of ϵ is denoted G^+. We say G is *orientable* (resp. *unorientable*) if $(G : G^+)$ is 1 (resp. 2). Thus a surface group is orientable if it is the fundamental group of an orientable closed surface. By the *orientation module* we mean the $\mathbf{Z}G$-module Ω which consists of the abelian group $\mathbf{Z}$, on which each $g \in G$ acts as $\epsilon(g) \in \{\pm 1\} = \operatorname{Aut}(\mathbf{Z})$.

The finite surface groups have order 1 or 2, and are the fundamental groups of the two-sphere and the projective plane, respectively. The infinite surface groups correspond to the surface group presentations in which the relator has length at least 4, and these are the fundamental groups of the *aspherical* closed surfaces, that is, closed surfaces whose universal covers are contractible.

A homotopy class of continuous maps between path-connected topological spaces determines an equivalence class of homomorphisms between their fundamental groups, where equivalence corresponds to composition with an inner automorphism of the target group. For aspherical closed surfaces, this correspondence between equivalence classes of morphisms is bijective, so, in quite a strong sense, the study of homotopy classes of continuous maps between aspherical closed surfaces is much the same as the study of homomorphisms between infinite surface groups.

1.3. DEFINITIONS. Let $\alpha\colon G_1 \to G_2$ be a homomorphism of surface groups.

If α together with the orientation maps of G_1 and G_2 form a commuting triangle, we say that α is *orientation-true*, and otherwise α is *orientation-false*. Thus α is orientation-true if and only if $\alpha^{-1}(G_2^+) = G_1^+$.

For any surface group presentations $G_1 = \langle S_1 \mid r_1 \rangle$, $G_2 = \langle S_2 \mid r_2 \rangle$, there exists a homomorphism of free groups $A\colon \langle S_1 \mid \ \rangle \to \langle S_2 \mid \ \rangle$ which induces α, and then there exist a non-negative integer d, elements $w_1, \dots, w_d$ of $\langle S_2 \mid \ \rangle$, and elements $\epsilon_1, \dots, \epsilon_d$ in $\{1, -1\}$, such that, in $\langle S_2 \mid \ \rangle$,

$$A(r_1) = \prod_{i=1}^{d} w_i r_2^{\epsilon_i} w_i^{-1}. \tag{1.1}$$

The *degree* of α, denoted $\mathcal{G}(\alpha)$, is the least value of d which occurs as we range over all the possible choices at our disposal. If G_1 or G_2 is finite, this concept is rather degenerate and we shall not be discussing this case. If G_1 and G_2 are infinite, the algorithm given in this article provides a lifting $A'\colon \langle S_1 \mid \ \rangle \to \langle S_2 \mid \ \rangle$ of α, and an expression of $A'(r_1)$ as a product of $\mathcal{G}(\alpha)$ conjugates of $r_2^{\pm 1}$.

Notice that if $\mathcal{G}(\alpha) = 0$, then α factors through the natural surjection $\langle S_2 \mid \ \rangle \to \langle S_2 \mid r_2 \rangle$; conversely, if α factors through any map from a free group F to $\langle S_2 \mid r_2 \rangle$, we can use the freeness of F to factor this map through the natural surjection. Thus $\mathcal{G}(\alpha) = 0$ if and only if α factors through a free group F. By replacing F with the image of α in F, we see that $\mathcal{G}(\alpha) = 0$ if and only if α factors through a surjective map to a free group.

Kneser's homological calculation of the degree, in the formulation of Skora [16], yields the following.

1.4. THEOREM (Kneser [10], [11]). *Let $\alpha\colon G_1 \to G_2$ be a homomorphism of infinite surface groups, and consider an equation* (1.1) *arising from some lifting of α.*

(i) *If α is orientation-true, then $\mathcal{G}(\alpha) = \left|\sum_{i=1}^{d} \epsilon_i\, \epsilon(w_i)\right|$, where the map $\epsilon\colon \langle S_2 \mid \ \rangle \to \{\pm 1\}$ is induced from the orientation map of G_2.*

(ii) *If α is orientation-false, and either d is even, or the index $(G_2 : \operatorname{Im}\alpha)$ is infinite, then $\mathcal{G}(\alpha) = 0$.*

(iii) *If α is orientation-false, and d is odd, and $(G_2 : \operatorname{Im}\alpha)$ is finite, then $\mathcal{G}(\alpha) = (G_2 : \operatorname{Im}\alpha)$.*

1.5. REMARK. Under pullback along α, the orientation module Ω_2 for G_2 becomes a G_1-module, again denoted Ω_2, and α induces a change of groups map in cohomology $H^2(\alpha, \Omega_2)\colon H^2(G_2, \Omega_2) \to H^2(G_1, \Omega_2)$. By Poincaré duality, $H^2(G_2, \Omega_2) \simeq \Omega_2 \otimes_{\mathbf{Z}G_2} \Omega_2 \simeq \mathbf{Z}$ with trivial G_2-action, and

$$H^2(G_1, \Omega_2) \simeq \Omega_1 \otimes_{\mathbf{Z}G_1} \Omega_2 \simeq \begin{cases} \mathbf{Z} & \text{if } \alpha \text{ is orientation-true,} \\ \mathbf{Z}/2\mathbf{Z} & \text{if } \alpha \text{ is orientation-false,} \end{cases}$$

with trivial G_1-action. Using a lifting A and an equation (1.1), and techniques such as those used in the proof of Theorem V.4.9 of [1], one can calculate that, up to sign, $H^2(\alpha, \Omega_2)$ acts as multiplication by $\sum\limits_{i=1}^{d} \epsilon_i\, \epsilon(w_i)$.

Hence, if α is orientation-true, the non-negative integer $\left|\sum\limits_{i=1}^{d} \epsilon_i\, \epsilon(w_i)\right|$ which occurs in Theorem 1.4 (a) is independent of the lifting chosen to get (1.1), and the theorem says that, in this case, there exists a lifting such that all the $\epsilon_i \epsilon(w_i)$ are equal.

Even if α is orientation-false, the parity of d (which is the parity of $\sum\limits_{i=1}^{d} \epsilon_i\, \epsilon(w_i)$) is independent of the lifting chosen to get (1.1), and will be called the *parity* of α, which is either *even* or *odd.*

In particular, if α is any homomorphism of infinite orientable surface groups, and $G_1 = \langle S_1 \mid r_1 \rangle$, $G_2 = \langle S_2 \mid r_2 \rangle$, are surface group presentations, then there exists a homomorphism of free groups $A\colon \langle S_1 \mid \ \rangle \to \langle S_2 \mid \ \rangle$ which induces α, such that $A(r_1)$ is a product of $\mathcal{G}(\alpha)$ conjugates of r_2 (or of r_2^{-1}), with no conjugate of r_2^{-1} (resp. r_2) needed in this expression.

The Kneser-Edmonds-Skora results give even more information about homomorphisms between infinite surface groups, but we shall postpone making the precise statements until Section 4.

In outline, the paper is structured as follows. In Section 2, we present some of the terminology we will use, describe some preliminary constructions, and recall how to associate, with a homomorphism between surface groups, a cellular map between surfaces which realizes the homomorphism. A cellular map between surfaces can be visualized as a labelled diagram, and, in Section 3, we give the algorithm for homotoping a diagram until a normal form is reached. In Section 4, we describe consequences for group homomorphisms, such as Kneser's Theorem determining degrees, and Nielsen's Theorem [14, Section 26] that every automorphism of a surface group lifts to an automorphism of the covering free group which sends the surface relator to a conjugate of itself or its inverse. In Section 5 we indicate how the algorithm can be described in terms of formal manipulations of presentations of surface groupoids, by describing a trivial example which illustrates the algorithm.

2. Diagrams of cellular surface maps

In this section we introduce the setting in which we shall work, and describe the connection with group theory.

2.1. Definitions. By a *two-dimensional CW-complex* X we shall mean a combinatorial CW-complex of dimension at most two, in which each cell has a preferred orientation. Formally we have the following situation.

As a set, X is the disjoint union of three sets V, E, F, whose elements are called the *vertices*, *edges*, and *faces*, of X, respectively.

There are given maps ι, τ, from E to V, and, for each edge e, the vertices ιe, τe are called the *initial* and *terminal* vertices of e, respectively. If $\iota e = \tau e$, we say that e is a *loop*. For each vertex v we understand that $\iota v = v = \tau v$.

We write $E^{\pm 1}$ for the Cartesian product $E \times \{1, -1\}$, and for any $(e, \epsilon) \in E^{\pm 1}$ we write e^{ϵ} for (e, ϵ). We identify $e = e^1$. We use the same conventions for the faces. For a vertex v, we understand that $v^1 = v = v^{-1}$.

For $e \in E$, we define $\iota(e^{-1}) = \tau e$, and $\tau(e^{-1}) = \iota e$.

Each face f of X has an associated *boundary cycle* which is a finite alternating sequence

$$\partial f = v_0, e_1^{\epsilon_1}, v_1, \dots, v_{n-1}, e_n^{\epsilon_n}, v_n, \tag{2.1}$$

where $n \geq 0$, the v_i are vertices, $v_n = v_0$, the e_i are edges, each ϵ_i is ± 1, and $\iota(e_i^{\epsilon_i}) = v_{i-1}$, $\tau(e_i^{\epsilon_i}) = v_i$. We define

$$\partial f^{-1} = v_n, e_n^{-\epsilon_n}, v_{n-1}, \dots, v_1, e_1^{-\epsilon_1}, v_0 .$$

It is thus implicit that we are assigning to each closed two-cell a polygonal structure, and a distinguished vertex where the boundary cycle begins and ends. Notice that we are allowing vertices of valence one, so a boundary cycle need not be reduced.

A one-dimensional CW-complex, that is, a two-dimensional CW-complex with no faces, will be called a *graph*.

2.2. DEFINITIONS.

Let X be a two-dimensional CW-complex.

The edges and vertices of X form a graph, denoted $X^{(1)}$ and called the *one-skeleton* of X. We say that X is *connected* if $X^{(1)}$ is a connected graph. The free groupoid on $X^{(1)}$ will be denoted $\pi X^{(1)}$.

Let e be an edge of X. A vertex v is said to be *incident* to e if $\iota e = v$ or $\tau e = v$, and in the former (resp. latter) case we call e^{-1} (resp. e^1) an edge with a *distinguished* incidence to v.

Let f be a face of X, and suppose that ∂f is as in (2.1). The distinguished vertex $v_0 = v_n$ will be denoted vert(f). There is associated an element of $\pi X^{(1)}$, denoted $\omega(f)$, which is vert(f) if $n = 0$, and is the product $e_1^{\epsilon_1} \cdots e_n^{\epsilon_n}$ if $n \geq 1$. A vertex v is said to be *incident* to f if some v_i equals v, and we then call the pair (f, i) a face with a *distinguished* incidence to v. An edge e is said to be *incident* to f if some e_i equals e, and we then call the pair (f, i) a face with a *distinguished* incidence to e. For $1 \leq i \leq n$, we say that $e_{i-1}^{\epsilon_{i-1}}$ and $e_i^{-\epsilon_i}$ are *adjacent* in f, where the subscripts are interpreted modulo n, that is, 0 is interpreted as n.

By the *groupoid* of X, denoted πX, we mean the groupoid obtained from $\pi X^{(1)}$ by imposing the relation $\omega(f) = $ vert(f) for each face f of X. For any vertex v of X the *fundamental group of X at v*, denoted $\pi(X, v)$, is the subgroup(oid) of πX consisting of all elements with initial and terminal vertex v. If X is connected then changing the choice of v gives an isomorphic group, and there is specified an isomorphism which is unique up to conjugacy.

We say that X is a (closed) *CW-surface* if it is a finite, connected, two-dimensional CW-complex such that for each edge e there are exactly two faces with a distinguished incidence to e, and for each vertex v the edges with a distinguished incidence to v form a single (non-empty) equivalence class under the equivalence relation generated by the relation of being adjacent in some face.

The former condition, on edges, implies that the edges with a distinguished incidence to v form cycles under the relation of being adjacent in some face, and the latter condition, on oriented edges, then requires that there be exactly one cycle at v, called the *edge cycle around* v.

2.3. EXAMPLES.

(i) A simplicial complex structure on a surface yields a CW-surface.

(ii) Any surface group presentation $\langle S \mid r \rangle$ has an *associated* CW-surface X with one vertex, denoted v, with edge set S, and with one two-cell, denoted

f, and the boundary cycle of f is the sequence in $S^{\pm 1}$ determined by r. Here $\pi X^{(1)} = \pi(X^{(1)}, v)$, $\pi X = \pi(X, v)$, and there are natural identifications $\pi X^{(1)} = \langle S \mid \ \rangle$ and $\pi X = \langle S \mid r \rangle$.

2.4. DEFINITIONS.

Let $X = (V, E, F)$ be a CW-surface.

The *dual surface* $X^* = (F^*, E^*, V^*)$ of X is defined to be any CW-surface constructed as follows. Let V^*, E^*, F^* be copies of V, E, F respectively, with bijective correspondence denoted by *. Then X^* has F^*, E^* and V^* as vertex set, edge set, and face set, respectively. For any $e \in E$, there are two different faces with a distinguished incidence to e. If we denote these by (f, i), (f', i'), with f, f' in F, then in X^*, the edge e^* is incident to the vertices f^*, f'^*. For any $v \in V$, the elements of X with a distinguished incidence to v are cyclically ordered, and this cyclically ordered set is called the *face-and-edge cycle around* v; by considering alternate terms we get the edge cycle around v and the *face cycle* around v. Applying * to the face-and-edge cycle around v gives a cyclic sequence which is taken to be the boundary cycle of v^*, once a distinguished vertex is chosen.

We say that X is *oriented* if each edge e occurs with opposite signs in the two faces with a distinguished incidence to e, and, in this event, we can use the signs to orient the dual surface X^* consistently.

We say that X is *orientable* if we can obtain an oriented CW-surface by replacing some faces with their inverses; otherwise X is *unorientable.*

In the remainder of this section and the next, all paragraphs which are devoted to the unorientable case are marked with a Maltese cross (✠), and, by skipping these, the reader interested primarily in orientable surfaces can follow the discussion for that case.

Consider a loop e in X, let v be the vertex incident to e, and consider a face f incident to e. Here two vertices in the boundary cycle of f are equal, which results in f getting attached to itself at a point.

If this attachment is performed without twisting, we say that e is an *orientable* or *two-sided* loop. If X is orientable than clearly all loops are orientable.

✠ If this attachment is performed with a twist, we say that e is an *unorientable*, or *one-sided* loop in X. This can be expressed in a more combinatorial manner by saying that e is unorientable if the boundary cycle of $f^{\pm 1}$, viewed cyclically, contains a subsequence e', v, e, v, e'', and the sequence e', e^{-1}, e, e''^{-1} of four distinct edges with a distinguished incidence to v is *not* in the correct order, with respect to the cyclic ordering by face-adjacency. Here the CW-complex resulting from collapsing e to a vertex is a CW-surface.

A useful way to codify a groupoid presentation of πX is to write $\langle E \mid \omega(\partial f)\,(f \in F)\rangle$, so $\pi X = \langle E \mid R\rangle$, where R is the set of words in $E^{\pm 1}$ determined by the boundary cycles, one word for each face. There is no need to specify the vertices, since they correspond to equivalence classes in $E^{\pm 1}$ under the equivalence relation generated by face-adjacency.

We can form a new CW-surface Y from X by successively erasing edges incident to two distinct faces (so melding two faces into one) until only one face f is left. The set E' of erased edges then corresponds to a maximal subtree in the one-skeleton of the dual complex of X. Here X and Y both have the same vertex set, and $Y^{(1)}$ can be viewed as the complement of E' in $X^{(1)}$, and πY is a subgroupoid of πX. One can even choose a retraction of πX onto πY by choosing a suitable image in $\pi Y^{(1)}$ of each erased edge. Notice that $\omega(\partial f)$ is an element of the free group $\pi(Y^{(1)}, v)$, where $v = \text{vert}(f)$, and there is an isomorphism $\pi(X, v) \simeq \pi(Y, v)$. Hence we have a homomorphism from a free group $\pi(Y^{(1)}, v)$ onto $\pi(Y, v) \simeq \pi(X, v)$, and the kernel is the normal subgroup generated by $\omega(\partial f)$.

Frequently we will want to alter the choice of E' by exchanging an edge b for some edge y of Y, such that b divides the face f into two faces f_1, f_2, each of which has a single occurrence of y in the boundary cycle. Either of these subfaces can be used to choose an element of the free groupoid $\pi Y^{(1)}$ which gets equated to b in the groupoid πX. We now have a new Y' with $Y'^{(1)} = Y^{(1)} \cup \{y\} - \{b\}$, and a map $Y'^{(1)} \to \pi Y^{(1)}$ which induces an isomorphism of free groupoids $\pi Y \simeq \pi Y'$. The single face f' of Y' is obtained by glueing together f_1 and f_2 along the two copies of b. It is straightforward to check that the isomorphism $\pi Y \simeq \pi Y'$ carries $\omega(f)$ to a conjugate of $\omega(f')$ or its inverse. The situation is amply illustrated in Figure 2.1, which depicts a labelled CW-subcomplex formed from two faces which are adjacent in two ways, so there are two ways to choose edges to be erased. In general, the symbol denoting an oriented path in $X^{(1)}$ is placed on the right of the path, and similarly for edges. Here, if we erase b, we get a face with clockwise boundary cycle $\bar{z}\bar{y}\bar{x}\bar{q}\bar{p}xyzrs$, where overlines denote inverses. But if we erase y, we get a face with clockwise boundary cycle $\bar{c}\bar{b}\bar{a}\bar{p}\bar{q}abcsr$. Algebraically, erasing b corresponds to using one of the small faces as a relation to eliminate b by identifying $b = \bar{a}qxyz\bar{s}\bar{c}$, and then $\bar{c}\bar{b}\bar{a}\bar{p}\bar{q}abcsr = \bar{c}(cs\bar{z}\bar{y}\bar{x}\bar{q}a)\bar{a}\bar{p}\bar{q}a(\bar{a}qxyz\bar{s}\bar{c})csr = s(\bar{z}\bar{y}\bar{x}\bar{q}\bar{p}xyzrs)\bar{s}$. Choosing whether to erase b or y affects the choice of free group mapping onto the surface group, but, as we have just seen, the free groups are related via an isomorphism which respects the given relators, up to conjugacy and inverse.

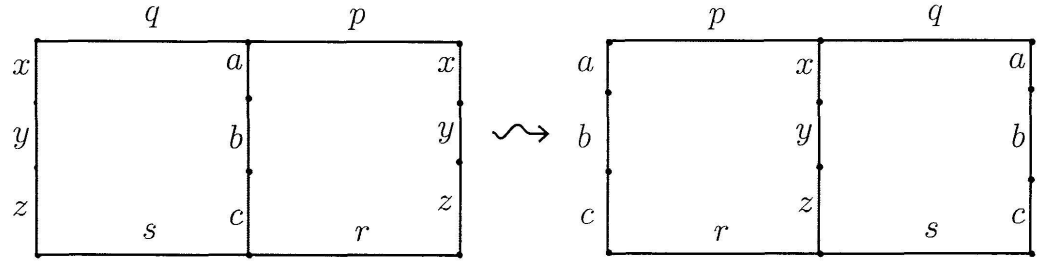


FIGURE 2.1
Changing edges to be erased

It usually happens that we are given a basis S of the free group $\pi(Y^{(1)}, v)$, and an expression of $\omega(\partial f)$ as a word r in $S^{\pm 1}$, so we get a presentation $\pi(X, v) = \langle S \mid r \rangle$. One standard method of choosing a basis S of $\pi(Y^{(1)}, v)$ is to choose a maximal subtree Y_0 of $Y^{(1)}$, and associate a free generator to each edge of $Y^{(1)} - Y_0$ in the natural way. This choice of S ensures that the above presentation is a surface group presentation. An alternative construction is to collapse the edges E_0 of Y_0 to get a new CW-surface Z with one vertex and one face, such that πZ is isomorphic to $\pi(X, v)$. Algebraically, in passing from the groupoid presentation $\pi X = \langle E \mid R \rangle$ to the groupoid presentation $\pi Y = \langle E - E' \mid \omega(\partial f) \rangle$, we successively use the erased edges to meld pairs of relators, and then annihilate the elements of E_0 to get a surface group presentation of πZ.

We will be interested in the situation where we are given a presentation to start with.

2.5. REMARKS. Let G be a surface group, and let $\langle S \mid r \rangle$ be a surface group presentation of G.

In this article we will be applying homotopy equivalences to a CW-surface X with fundamental group G, and we wish to ensure that the given presentation is always recoverable. Some of the difficulties arise from the choices involved. The choice of base point v affects the data only up to conjugacy. The choice of set of edges E' determining a maximal tree in the one-skeleton of the dual surface affects the data up to isomorphism of the covering free group $\pi(X^{(1)} - E', v)$, and we have seen that the isomorphism respects the relators up to conjugacy and inverse. Thus if S is associated to a basis of one of the free groups in such a way that r corresponds to a conjugate of the relator or its inverse, then S is associated to a basis of each of the free groups in such a way that r corresponds to a conjugate of the relator or its inverse. We want to ensure that each homotopy equivalence specifies an isomorphism of covering free groups so as to respect relators in this way.

We start with the CW-surface associated to the presentation $\langle S \mid r \rangle$, and apply homotopy equivalences using four operations called subdivision, erasing, collapsing and expanding.

Subdivision of edges and faces changes the covering free group by a simple isomorphism which preserves relators.

Erasing a set of edges E'' which determine a subtree of the one-skeleton of the dual complex changes, by a simple isomorphism which preserves relators, the covering free group corresponding to a choice of E' containing E''.

Collapsing, in the cases of interest to us, concerns the three elementary operations of collapsing to a vertex an edge which is not a loop, collapsing to an edge a two-edged face which is not a sphere, and collapsing a one-edged face to a vertex. If we want to collapse an edge which is not a loop, we first adjust the choice of E' to ensure that it does not contain the edge to be collapsed. It is then straightforward to check that, for each of the three elementary collapsing operations, the covering free group changes by a simple isomorphism which preserves relators.

Expanding is the reverse of collapsing, and changes the covering free group by a simple isomorphism which preserves relators.

At any stage we can lose the base vertex, and prior to its disappearance we have to change the covering free group by conjugating by a chosen path to a new base vertex.

2.6. DEFINITIONS. Let X_1 and X_2 be two-dimensional CW-complexes. A map of sets $\beta\colon X_1 \to X_2$ is said to be *cellular* if the following are satisfied:

If v is a vertex of X_1, then $\beta(v)$ is a vertex of X_2.

If e is an edge of X_1, then $\beta(e)$ is a vertex or an edge of X_2, and $\iota\beta(e) = \beta\iota(e)$, $\tau\beta(e) = \beta\tau(e)$.

If f is a face of X_1, exactly one of the following holds:

$\beta(f)$ is a vertex v, and all the terms of $\beta(\partial f)$ are v;

$\beta(f)$ is an edge e, one of the terms of $\beta(\partial f)$ is e, one is e^{-1}, and the rest are vertices;

$\beta(f)$ is a face, every edge incident to f is mapped to an edge, and $\partial\beta(f) = \beta(\partial f)$; here, with ∂f as in (2.1), we understand that

$$\beta(\partial f) = \beta(v_0), \beta(e_1)^{\epsilon_1}, \beta(v_1), \ldots, \beta(v_{n-1}), \beta(e_n)^{\epsilon_n}, \beta(v_n). \tag{2.2}$$

This definition of cellular map is more restrictive than the usual definition, but it does include simplicial maps, with suitably chosen orientations of simplices, so we do not lose any homotopy classes of maps.

2.7. CONSTRUCTION. Let $\beta\colon X_1 \to X_2$ be a cellular map of two-dimensional CW-complexes.

Then β induces a cellular map on the one-skeletons $\beta^{(1)}\colon X_1^{(1)} \to X_2^{(1)}$, and this determines a groupoid homomorphism $\pi(\beta^{(1)})\colon \pi X_1^{(1)} \to \pi X_2^{(1)}$. The latter then induces a groupoid homomorphism $\pi(\beta)\colon \pi X_1 \to \pi X_2$. If we specify a vertex v of X_1, then we obtain a group homomorphism $\pi(\beta, v)\colon \pi(X_1, v) \to \pi(X_2, \beta(v))$.

The *diagram associated to* β consists of the CW-complex X_1 together with a labelling of its cells, which labels each cell with its image cell in X_2. For any cells c_1 of X_1, c_2 of X_2, if $c_2 = \beta(c_1)$ we say that c_1 is a c_2*-cell.* With our definition of cellular map, there are three types of labelling that a face of X_1 can have, namely, we can have a v-face, an e-face, or an f-face, and these can be depicted as in Figure 2.2.

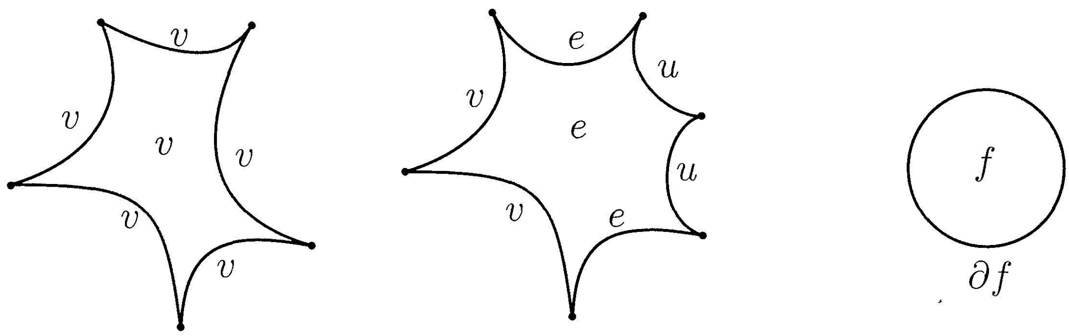


FIGURE 2.2
The three types of labelled cell

We remark that the labelled regions in Figure 2.2 are precisely the types of regions used by Ol'shanskii [15]. These diagrams, which represent elementary concepts in topology, can be viewed as Lyndon-van Kampen diagrams in which trivial relators play a larger part than usual. Here we have a groupoid setting which allows various vertices, rather than just the one vertex allowed when considering groups.

In terms of diagrams, β is a *covering* if for each vertex v of X_1 the cycle of labels (in X_2) of faces (in X_1) with a distinguished incidence to v is precisely the cycle of faces (in X_2) with a distinguished incidence to the label $\beta(v)$ of v. Also β is a *branched covering* if for each edge e of X_1 the labels (in X_2) of the two faces (in X_1) with a distinguished incidence to e are precisely the two faces (in X_2) with a distinguished incidence to the label $\beta(e)$ of e; here deleting all vertices leaves a covering. We will abuse notation and say that β is a *pinching* if X_1 is obtained from X_2 by slicing open non-loop edges and inserting punctured projective planes and punctured tori with cell structures and labels as depicted in Figure 3.1. Here β acts by collapsing these subsurfaces to edges, so is homotopic to a pinching, and our abuse of notation is reasonable.

Our main activity will be to apply operations to these diagrams. Let us note one which will be used frequently.

2.8. CONSTRUCTION (type: subdivision). Suppose we are given two-dimensional CW-complexes X_1, X_2, a face f of X_1, and a cellular map $\beta\colon X_1 - \{f\} \to X_2$, such that $\pi(\beta^{(1)})(\omega(f))$ is trivial in $\pi X_2^{(1)}$, that is, $\pi(\beta^{(1)})(\omega(f)) = \beta(\mathrm{vert}(f))$.

We wish to subdivide f to obtain a refinement X_1' of X_1, and an extension $\beta'\colon X_1' \to X_2$ of β.

Let ∂f be as in (2.1), and let $\beta(\partial f)$ denote the sequence in (2.2). Since $\pi(\beta^{(1)})(\omega(f))$ is trivial in $\pi X_2^{(1)}$, either every term of (2.2) is a vertex v, or some subsequence of (2.2) has the form $e^{\epsilon}, v, v, \dots v, e^{-\epsilon}$. In the former case we can extend β to X_1 by labelling f as v. In the latter case, we can subdivide f into two faces by adding an edge with label the vertex $u = \iota(e^{\epsilon})$ slicing off a piece of f with label e, as in Figure 2.3; here we use $\varnothing$ to indicate a region with boundary label which determines a trivial element of the free groupoid.

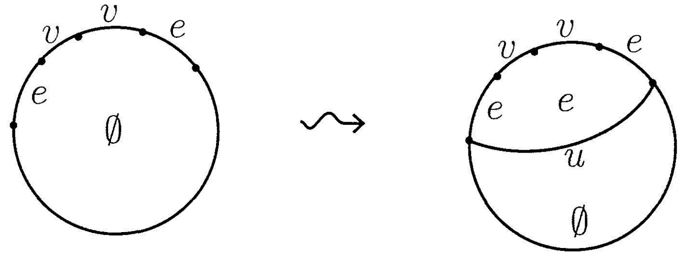


FIGURE 2.3
Subdividing

We can continue in this way, and by induction on the length of ∂f, we obtain a subdivision of f which allows an extension of β.

It is convenient to mention a more complicated operation at this stage.

2.9. CONSTRUCTION (type: subdivision). Suppose we are given two-dimensional CW-complexes X_1, X_2, and a cellular map $\beta^{(1)}: X_1^{(1)} \to X_2^{(1)}$ of the one-skeletons, such that, for each face f of X_1, $\pi(\beta^{(1)})(\omega(f))$ is trivial in πX_2, that is, there exists $d \geq 0$, and elements w_i of $\pi X_2^{(1)}$, and faces f_i of X_2 such that, in the free groupoid $\pi X_2^{(1)}$,

$$\pi(\beta^{(1)})(\omega(f)) = \prod_{i=1}^{d} w_i^{-1}\, \omega(f_i)^{\epsilon_i}\, w_i\,. \tag{2.3}$$

Essentially as in the previous construction, we wish to subdivide each face of X_1 to obtain a refinement X_1' of X_1, and an extension $\beta': X_1' \to X_2$ of $\beta^{(1)}$.

Let f be a face of X_1, and suppose (2.3) holds. We first subdivide f into $d+1$ two-cells by drawing in d balloons-on-sticks, as in Figure 2.4, which are subdivided and labelled in such a way that, if the boundary cycle is read clockwise, the labelling of the ith stick, starting at the basepoint, gives the word w_i, and the labelling of the boundary of the ith balloon, starting at the attaching point and reading counter-clockwise, is $\partial f_i^{\epsilon_i}$. We label the ith balloon f_i, and orient it in the manner dictated by the label on the boundary.

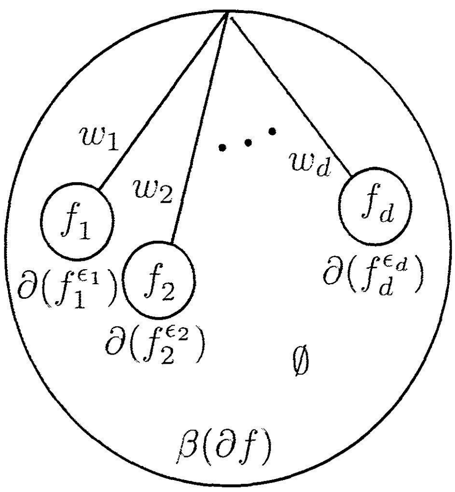


FIGURE 2.4
Subdividing for a relation

In the subdivided f there remains a single two-cell f' which is not labelled. Here, (2.3) implies that

$$\pi(\beta^{(1)})(\omega(f')) = \pi(\beta^{(1)})(\omega(f)) \prod_{i=1}^{d} w_i^{-1}\,\omega(f_i)^{-\epsilon_i}\,w_i = \beta^{(1)}(\mathrm{vert}(f'))$$

in $\pi X_2^{(1)}$. By Construction 2.8, we can subdivide f', and extend $\beta^{(1)}$. Thus we find we can subdivide each face of X_1 to obtain a refinement X_1' of X_1, and an extension $\beta' : X_1' \to X_2$ of $\beta^{(1)}$.

We conclude this section by recalling how one constructs a cellular map of CW-surfaces from a homomorphism of surface groups.

2.10. CONSTRUCTION (type: subdivision). Let $\alpha: G_1 \to G_2$ be a homomorphism of surface groups.

Let $\langle S_1 \mid r_1 \rangle$, $\langle S_2 \mid r_2 \rangle$ be surface group presentations of G_1, G_2, respectively, and choose a lifting $A: \langle S_1 \mid \ \rangle \to \langle S_2 \mid \ \rangle$ of α. Thus A is a homomorphism of free groups such that $A(r_1)$ lies in the normal subgroup generated by r_2, and the resulting homomorphism $\langle S_1 \mid r_1 \rangle \to \langle S_2 \mid r_2 \rangle$ is α.

Let $X_1 = (w, S_1, g)$ and $X_2 = (v, S_2, f)$ denote the CW-surfaces associated to the presentations $\langle S_1 \mid r_1 \rangle$ and $\langle S_2 \mid r_2 \rangle$, respectively.

We want to subdivide X_1 to obtain a CW-surface X_1', and a cellular map $\beta' : X_1' \to X_2$.

We begin by subdividing the one-skeleton $X_1^{(1)}$, to get a graph Γ, as follows. For each $s \in S_1$, $A(s)$ is a word in S_2, possibly empty, and hence $A(s)$ is either 1, or is a non-empty reduced word $e_1^{\epsilon_1} \cdots e_n^{\epsilon_n}$ in S_2, that is, $n \geq 1$, each e_i lies in S_2, and each ϵ_i is 1 or -1. If $A(s) = 1$ we label s with the vertex v; in the second case, we add $n-1$ new vertices to subdivide s into n edges, denoted $s_1^{\epsilon_1}, \ldots, s_n^{\epsilon_n}$, and label each s_i with e_i having the same orientation. Doing this for each element of S_1 gives us a labelled graph Γ, with the labels coming from $X_2^{(1)}$. Notice that the two-cell of the subdivided X_1 has as boundary cycle the subdivided r_1, and the labelling gives a word in $\pi(X_2^{(1)}, v)$ which corresponds to $A(r_1)$. Since this word equals v in $\pi(X_2, v)$, we can use Construction 2.9 to further subdivide X_1 and obtain a CW-surface X_1' and a cellular map $\beta' : X_1' \to X_2$. Moreover $\pi(\beta'^{(1)}, w): \pi(X_1'^{(1)}, w) \to \pi(X_2^{(1)}, v)$ can naturally be identified with $A: \langle S_1 \mid \ \rangle \to \langle S_2 \mid \ \rangle$, and $\pi(\beta', w): \pi(X_1', w) \to \pi(X_2, v)$ can naturally be identified with $\alpha: \langle S_1 \mid r_1 \rangle \to \langle S_2 \mid r_2 \rangle$. Thus we have a cellular map of CW-surfaces which realizes α.

Notice that the cellular map is constructed from an equation of the form (1.1). We can apply the algorithm of the next section to this map, to get a new cellular map, from which we can extract a new equation of the form (1.1), without altering the given presentations, since at each step we can choose isomorphisms of the covering free groups which respect the relator up to conjugacy and inverse. Each element of S_1 will be transformed into a path in a labelled one-skeleton without changing the homotopy class in the surface underlying X_1 ; this amounts to choosing a new labelling for each element of S_1, which, in turn, amounts to choosing a new lifting at the free group level. This whole process will then give non-trivial group-theoretical information, although not so much as in the topological situation.

3. The algorithm

Throughout this section let $\beta\colon X_1 \to X_2$ be a cellular map of CW-surfaces.

Let V, E, and F denote the sets of vertices, edges, and faces, respectively, of X_2. We then have a diagram with V-faces, E-faces, and F-faces, as depicted in Figure 2.2.

The aim of this section is to alter β by composing it with various cellular homotopy equivalences of X_1 and X_2 (based on the operations of contracting, expanding, erasing, and subdividing), until we arrive at the minimum possible number of F-faces. These alterations of β can be viewed as homotopies, since one is free to imagine that there is a surface X underlying X_1 that has lines inscribed on it, and that these lines can be deformed continuously. Abusing notation then, we will say that the altered forms of β are *homotopic* to β.

3.1. Construction (type: subdivision). If X_2 has a loop e, we subdivide e by adding a new vertex v, and, in X_1, subdivide each e-edge, and each e-face, by adding a new v-vertex, and a new v-edge, respectively.

By our definition of CW-surface, X_2 has an edge. Thus we have the following.

3.2. Condition. *There is at least one edge in X_2, but there are no loops. Hence, in X_1, no E-edge is a loop, or equivalently, all loops are V-loops.*

We will not make any further adjustments to X_2, except in the case of a branched covering where we may have to apply Construction 3.30.

We now want to describe the basic configurations.

3.3. DEFINITIONS.

Suppose that w_1, w_2 are two distinct vertices of X_1, and d_1, d_2, d_3 are three distinct one-cells of X_1, such that d_1 is a loop at w_1, and d_2, d_3 join w_1 to w_2. Let g be a face of X_1 with boundary label $w_1, d_1, w_1, d_2, w_2, d_3^{-1}, w_1$. In this event we say that (the closure of) g is a *loop triangle*. We say that g is an *annular triangle* if d_1 is an orientable loop.

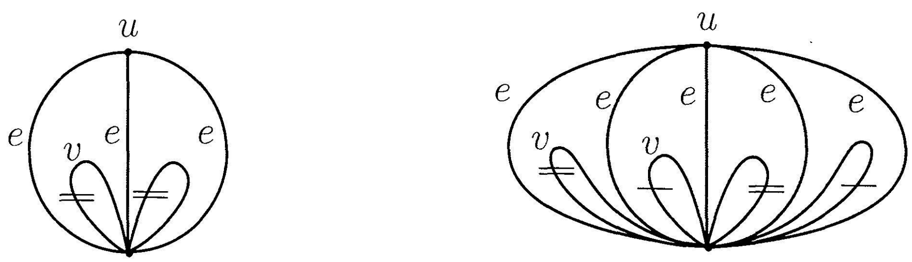
✠ We say that g is a *Möbius triangle* if d_1 is an unorientable loop.

Suppose further that β sends w_1 and d_1 to a vertex v of X_2, sends d_2, d_3 and g to a one-cell $e \in E^{\pm 1}$ of X_2, and sends w_2 to the other vertex u of e; in this event we say that g is a *loop e-triangle*. If d_1 is an orientable loop, we say that g is an *annular e-triangle*.

✠ If d_1 is an unorientable loop, we say that g is a *Möbius e-triangle*.

Two loop E-triangles are said to be *E-adjacent* if they have an E-edge in common.

By an *orientation-true prepinching*, we mean four consecutive E-adjacent triangles such that two pairs have orientable V-loops in common, and the V-loops have a common vertex, which determine a labelled CW-subcomplex as depicted in Figure 3.1(b). There are specified two edges which are incident to only one face; these edges are said to be the *boundary edges* of the prepinching. (We allow the possibility that the two boundary edges become identified in X_1, and in this case the prepinching is the whole diagram, a torus.) The remaining V- and E- edges are called the *interior* edges of the prepinching.



(a) (b)

FIGURE 3.1
The two elementary types of prepinching

✠ By an *orientation-false prepinching* we mean two E-adjacent triangles having an unorientable V-loop in common, which determine a labelled CW-subcomplex as depicted in Figure 3.1(a). There are specified two edges which are incident to only one face; these edges are said to be the *boundary edges* of the prepinching. (We allow the possibility that the two boundary edges become identified in X_1, and in this case the prepinching is the whole diagram, a projective plane.) The remaining V- and E- edges are called the *interior* edges of the prepinching.

3.4. DEFINITION. The *measuring quadruple* of the diagram consists of the following non-negative integers:

(1) the number of F-faces,

(2) the number of E-faces which have boundary length at least four,

(3) the number of V-faces,

(4) the number of edges which are not interior edges in prepinchings.

The quadruples are ordered lexicographically reading from (1) to (4); this is a well-ordering.

All subsequent operations will reduce the measuring quadruple, and since the quadruples are well-ordered, the procedure must eventually stop.

3.5. CONSTRUCTIONS (type: subdivision).

(a) If, for some $v \in V$, the diagram contains a v-face, then we choose an $e \in E$ which is incident to v in X_2, and subdivide the v-face into e-triangles by adding a vertex and e-edges, as depicted in Figure 3.2 (a).

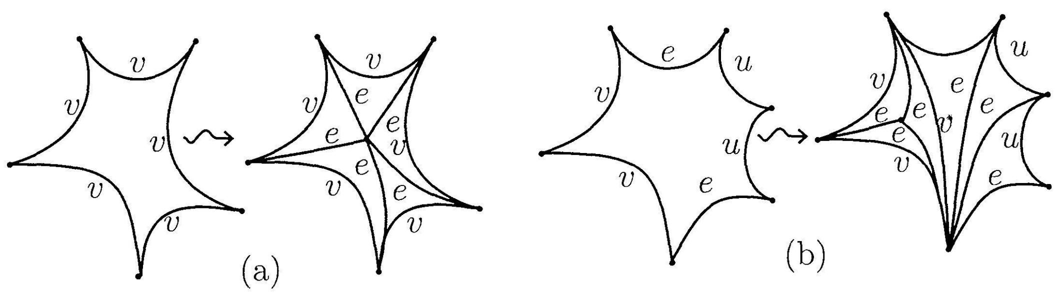


FIGURE 3.2
Subdividing into triangles

(b) If, for some $e \in E$, with vertices $u = \iota(e)$, $v = \tau(e)$, the diagram contains an e-face which has boundary length at least four, we subdivide the e-face into e-triangles as depicted in Figure 3.2 (b).

These operations reduce the measuring quadruple, since they reduce the second or third coordinate without affecting the preceding coordinates.

These are the first of several situations where we use subdivision, usually preceded by erasing, to express a homotopy between two maps which collapse a disc to a tree. Notice that it is important not to disturb the boundary of the disc, since we are not allowed to damage F-faces by collapsing E-edges.

We may now assume that we have the following.

3.6. CONDITION. *There are no V-faces.*
All E-faces have boundary length at most three.

Since an E-face has to have boundary length at least two, we then have only E-triangles and E-bigons, and no other E-faces.

Our next strategy is to eliminate some edges and faces.

3.7. CONSTRUCTION (type: collapsing). If the diagram contains a V-edge joining two distinct vertices of X_1, then we collapse the V-edge, and identify the two vertices.

This operation reduces the measuring quadruple, since it reduces the fourth coordinate, and does not increase any of the other coordinates.

We may therefore assume we have the following.

3.8. CONDITION. *All V-edges are loops.*

3.9. TERMINATING CASE. *If our diagram contains an E-face of boundary length two, and the two edges are identified in X_1, then X_1 is a sphere, and β is a degree zero map which collapses it to an edge which is not a loop, and we have the Normal Form 3.31(a).*

We may therefore assume we have the following.

3.10. CONDITION. *If an E-face has boundary length two, then the two edges are not identified in X_1.*

3.11. CONSTRUCTIONS (type: collapsing).

(a) If, for some $e \in E$, the diagram contains an e-face of boundary length two, whose edges are not identified in X_1, then the closure of the face is an e-disc, and we collapse it to an e-edge, as in Figure 3.3 (a).

(b) If the diagram contains an e-face of boundary length three, and two of the edges are identified in X_1, then, by Conditions 3.6 and 3.8, for some $v \in V$, the third edge is a v-loop, the e-face is a v-disc, and we then collapse this v-disc to a v-vertex, as in Figure 3.3 (b).

These operations reduce the measuring quadruple, since they reduce the fourth coordinate without affecting the preceding coordinates.

FIGURE 3.3
Two types of collapsing

Thus we may assume we have the following.

3.12. CONDITION. *Each E-face is a loop E-triangle.*

To summarize, in terms of closures in X_1, all E-edges are non-loops, all E-faces are loop triangles, there are no V-faces, and all V-edges are loops, and are incident to two loop E-triangles.

✠ We now turn our attention to unorientable V-loops.

✠ 3.13. CONSTRUCTION (Skora [16]; type: expand, erase, subdivide, collapse). If an unorientable V-loop is incident to two non E-adjacent E-faces, then we create an orientation-false prepinching by applying the steps depicted in Figure 3.4.

✠ This operation reduces the measuring quadruple, since it reduces the fourth coordinate without affecting the others.

✠ Thus we may assume the following.

✠ 3.14. CONDITION. *Each unorientable V-loop is the interior edge of an orientation-false pre-prinching.*

✠ This condition will enable us to ignore the unorientable V-loops, and treat them as if they were hidden within edges.

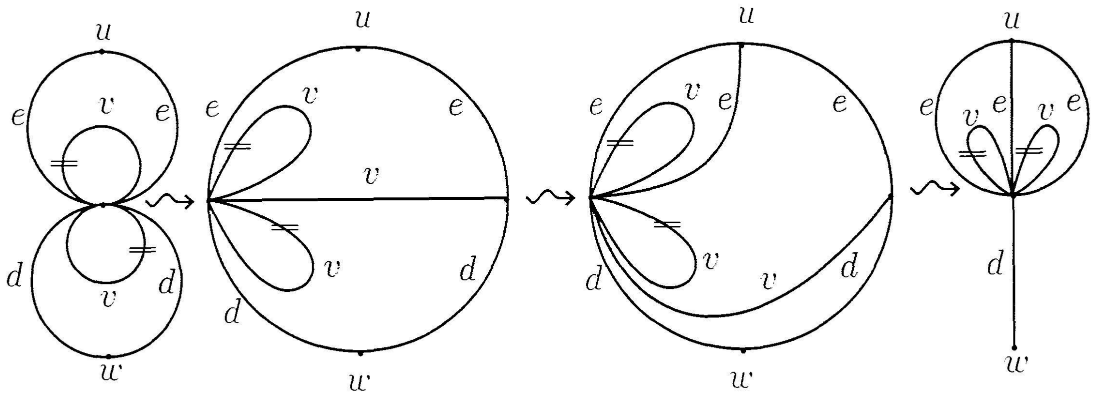


FIGURE 3.4
Normalizing a prepinching

We now want to examine how the orientable V-loops which meet at a vertex fit together.

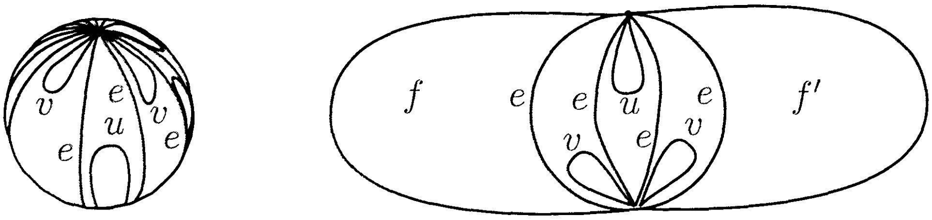
3.15. DEFINITIONS. Let w be a vertex in the diagram, and denote its label by $v \in V$.

By a *labelled cycle around* w we mean a finite sequence

$$g_1, d_1, g_2, \ldots, d_m, g_{m+1} = g_1,$$

where each d_i is an edge with a distinguished incidence to w, but is not a v-loop, each g_i is a face with a distinguished incidence to d_i, and d_i, d_{i+1} become adjacent in g_i after omitting v-loops. That is, we are listing face-adjacent edges, except where E-triangles have loops at w in which case we treat the v-loop as a vertex, and pass from one E-edge to the other. Thus we are looking at the face-and-edge cycles around w which arise when we collapse to w all the v-loops at w.

There are various types of labelled cycles. For any e in E, an e-triangle can be E-adjacent to another e-triangle or to an F-face. By finiteness of the diagram, every e-triangle lies in a labelled cycle of e-triangles, as in Figure 3.5 (a), or in a sequence joining together two F-faces, as in Figure 3.5 (b). In the case of Figure 3.5 (a), we say that the e-faces form a *punctured e-sphere*.



(a) (b)

FIGURE 3.5
Adjacent e-triangles

✠ Notice that, by Condition 3.14, any unorientable loops are identified in neighbouring pairs.

The whole subsurface is collapsed to e by β. Thus we see that one possibility for a labelled cycle around w consists of e-triangles and e-edges.

Consider now the case of Figure 3.5 (b). Here we get a sequence, starting at an f-face, for some $f \in F$ with a distinguished incidence (f, i) to e, and ending in an f'-face, for some $f' \in F$ with a distinguished incidence (f', i') to e. We say that the f-face and the f'-face are *e-joined*. There are two possibilities. Either $(f, i) \neq (f', i')$, so they are the two faces with a distinguished incidence to e (in X_2), or $(f, i) = (f', i')$. In the former case, we say the f-face and the f'-face are *well* joined, and in the latter case we say they are *badly* joined. Notice that if they are badly joined, then the two F-faces are both f-faces, and they must be distinct f-faces, since if they are equal, then their distinguished e-edges must be equal, and these e-edges are then incident to zero or two e-faces and one f-face, which contradicts the surface property.

From Figure 3.5 (b), we see that the second possibility for a labelled cycle around w consists of F-faces joined together cyclically by E-triangles.

If the F-faces in a labelled cycle around w are well-joined, then it is easy to see that the corresponding cycle of labels in F is given by repeating the face cycle around v in X_2 an integral number of times; if, moreover, there is only one labelled cycle around w, then the number of times the face cycle around v is repeated will be called the *branching degree* at w.

If all the faces are F-faces, and all E-joined F-faces are well joined, then the corresponding map is a branched covering, since it becomes a covering if all the vertices are deleted.

Here the number of F-faces is a multiple of the size of F, and the quotient is the degree d of the map, and the map is then a d-fold branched covering.

If $d = 1$ then the map is a homeomorphism. If the branching degree at each vertex is 1, then the map is a covering.

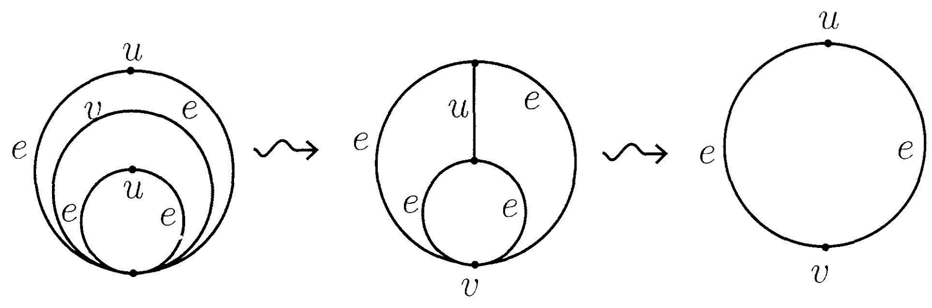


FIGURE 3.6
Collapsing

3.16. CONSTRUCTIONS. Suppose that two annular E-triangles have a common V-edge, and a common E-label, but do not lie in a prepinching.

Thus, for some $e \in E$, with vertices $u, v \in V$, two annular e-triangles have an orientable v-loop in common, and the v-loop is not an interior edge of a prepinching. We consider various possibilities for the intersection of (the closures of) the two annular e-triangles. Since the two e-triangles have an orientable loop in common, they get separated into different face cycles after collapsing the loop to a vertex, so they cannot have an e-edge in common. However, they may have a common u-vertex.

(a) (type: erasing, subdividing, collapsing). If the intersection of the two annular e-triangles is precisely the v-loop, then we erase the v-loop, and draw in a u-edge which is not a loop, and we can now collapse the resulting triangles, as depicted in Figure 3.6. Thus we take an annulus, which is a compact subsurface with two boundary components, and homotope it to a loop formed by two edges, in such a way that the boundaries are respected.

(b) (type: expanding, erasing, subdividing, collapsing). If the two annular e-triangles with a common v-loop also have a common u-vertex, then we create an orientation-true prepinching, as depicted in Figure 3.7. Thus we take a punctured torus, which is a compact subsurface with a single boundary component, and fit it into a sliced-open edge, in such a way that the boundary is respected.

Each of these two operations reduces the measuring quadruple, since it reduces the fourth coordinate without affecting the preceding coordinates.

We may therefore assume that we have the following.

3.17. CONDITION. *Each V-loop is either interior to a prepinching, or is an orientable V-loop incident to two E-faces with different E-labels. Moreover, all E-faces are loop triangles and there are no V-faces.*

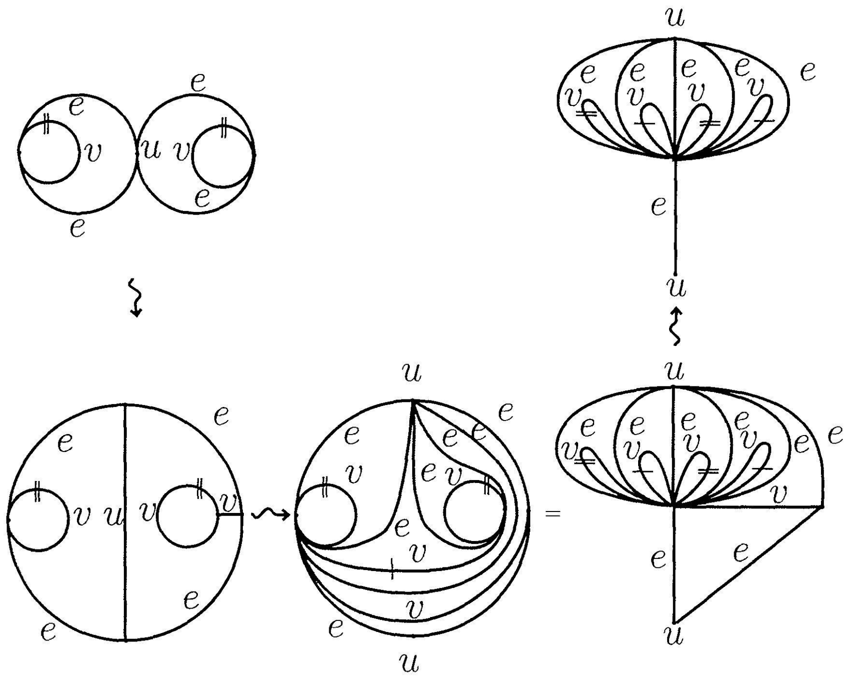


FIGURE 3.7
Normalizing a prepinching

3.18. CONSTRUCTION (type: expanding-collapsing). Suppose that, at some vertex of the diagram, there are two distinct labelled cycles having an E-label in common.

Thus there exist two e-edges incident to a vertex, and these two e-edges are separated into different labelled cycles by some orientable v-loop at the vertex, as in the left diagram in Figure 3.8.

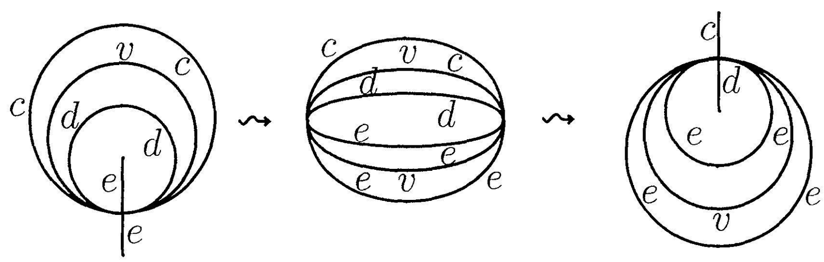


FIGURE 3.8
Readjusting

We expand the two e-edges into e-triangles, and then collapse the resulting c- and d-triangles to c- and d-edges, as in Figure 3.8, without disturbing the two boundary components at any stage. Possibly the interior e-edge equals one of the d-edges, and possibly, the exterior e-edge equals one of the c-edges; in this event, the corresponding identifications must be maintained throughout the operation.

Thus we have applied a homotopy which does not affect the measuring quadruple, and we can now apply the Constructions 3.16 (a) and (b), which will reduce the measuring quadruple.

Hence we may assume the following.

3.19. CONDITION. *At each vertex, distinct labelling cycles have disjoint label sets.*

3.20. CONSTRUCTION (type: expanding, erasing, subdividing, collapsing). Suppose there is a vertex w, such that there is only one labelling cycle around w, and there is a V-loop incident to w which is not interior to a prepinching.

Since collapsing the V-loop at w separates the face cycle around w into two disjoint cycles, and, by hypothesis, collapsing all the V-loops at w leaves a single face cycle, there must be some V-loop which connects up the two face cycles, giving us the situation depicted in the left diagram in Figure 3.9.

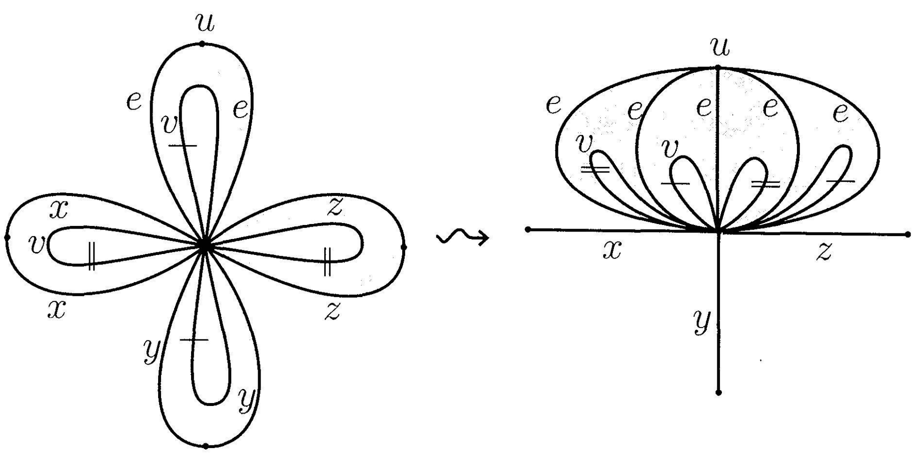


FIGURE 3.9
Normalizing a prepinching

It is now straightforward to homotope this configuration into an orientation-true prepinching, as depicted in Figure 3.9, without affecting the single boundary component.

This reduces the measuring quadruple in the fourth coordinate, without affecting the other coordinates.

Hence we may assume the following.

3.21. CONDITION. *At each vertex with a single labelling cycle, all V-loops are interior to prepinchings.*

3.22. TERMINATING CASE. *If there is no F-face then we have the following situation.*

All the faces are E-triangles, and the diagram is formed by amalgamating punctured E-spheres along the V-loops, and the E-spheres which meet at a vertex have distinct E-labels.

The algorithm now terminates, as we have the Normal Form 3.31(a).

Hence we may assume the following.

3.23. CONDITION. *The diagram has at least one F-face.*

3.24. CONSTRUCTION (type: erasing and subdividing). Suppose that two F-faces are badly E-joined, as in the left diagram of Figure 3.10.

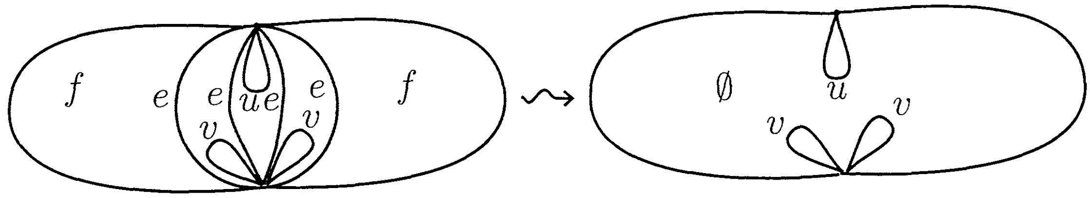


FIGURE 3.10
Starting over

Here we abandon all the progress we have made. We erase all the E-edges involved, to join up distinct faces, and so obtain a disc with a boundary label which determines a trivial element of $\pi X_2^{(1)}$, as in the second diagram of Figure 3.10. Now we apply Construction 2.7 to fill in the disc with V- and E-faces, and so get a new diagram representing a map homotopic to β.

This procedure reduces the first coordinate of the measuring quadruple by two.

The foregoing construction disturbs the conditions we have obtained so far, and we return to Construction 3.5.

Repeating the procedure up to this stage a finite number of times, we eventually eliminate all pairs of F-faces which are badly joined. Thus we may assume the following.

3.25. CONDITION. *Each E-joined pair of F-faces is well-joined. Hence the only faces are E-triangles forming prepinchings, and F-faces. Moreover, the branching degree is defined at each vertex.*

Proof. By Condition 3.23, there is at least one F-face.

Let us consider a vertex w incident to an F-face, and look at the labelled cycle around w containing the F-face. Since F-faces are well joined, we see, as in Definition 3.15, that as we run through the labelled cycle, the labels run through the face-and-edge cycle around v, where $v \in V$ is the label of w. Thus every edge incident to v occurs as a label in the labelled cycle around w. It follows, from Condition 3.19, that there is only one labelled cycle around w. Now, by Condition 3.21, all the V-loops at w are interior to prepinchings. Hence all the faces incident to w are E-triangles forming prepinchings and F-faces. It follows that the compact subsurface formed by the F-faces and the E-triangles occurring in prepinchings is closed under edge adjacency, so is the whole surface.

Hence the branching degree is defined at each vertex. □

3.26. TERMINATING CASE. *If the branching degree is 1 at each vertex then after the pinching, consisting of collapsing to edges the prepinching regions which are as depicted in Figure 3.1, we have a diagram in which all faces are F-faces, and all E-joined F-faces are well joined, and the branching degree at each vertex is 1, so it is a diagram representing a covering.*

The algorithm terminates since we have the Normal Form 3.31(b).

Hence we may assume the following.

3.27. CONDITION. *The branching degree is at least two at some vertex w.*

Let $v \in V$ denote the label of w. Since w has branching degree at least 2, the face labels around w run at least twice through the faces round v. Thus we can choose an $f \in F$ with a distinguished incidence to v, and choose the first two terms of the labelled face sequence around w with this label. Clearly these two faces with distinguished vertex are distinct. Choose an $e \in E$ in the boundary cycle of f next to the distinguished occurrence of v. Thus we have the situation occurring in the left diagram in Figure 3.11, and the two e-edges are distinct.

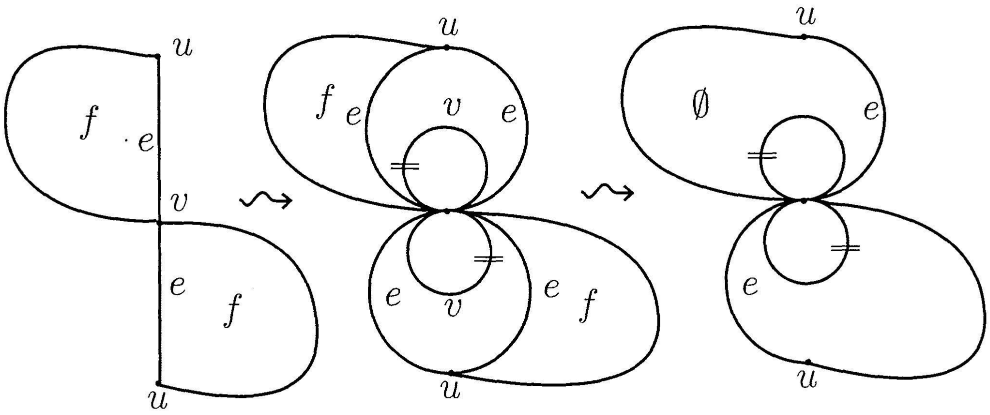


FIGURE 3.11
Starting over

✠ 3.28. CONSTRUCTION (Skora [16]; type: erase-subdivide). Suppose the diagram has at least one unorientable V-loop.

✠ We claim that unorientable V-loops are highly mobile, in the sense that two Möbius E-triangles, attached along an unorientable V-loop (and possibly an E-edge) forming a Möbius band (that is, a punctured projective plane), can slide around; this sliding has the same effect as cutting open a pair of incident edges (resp. an edge) to form a loop, which is then identified with the boundary of the Möbius band, while the reverse of such an operation is performed somewhere else. This can be shown using Construction 3.13, and its reverse, and similar arguments, and we will not go into details since they are straightforward. These operations do not affect the first coordinate of the measuring quadruple, which is the coordinate the operation will eventually reduce. Thus we can move one of the unorientable V-loops into the two e-edges, as depicted in the middle diagram of Figure 3.11.

✠ Now we have two f-faces, and two e-faces, which give four distinct faces, and we erase the three edges along which they are joined, to get an open disc, as in Figure 3.11, and the boundary label is $\partial f, e, e^{-1}, \partial f^{-1}$, which determines a trivial element in the free groupoid $\pi X_2^{(1)}$. Notice that the boundary cycle has a repeated vertex which causes the closure of the disc to be attached to itself with a twist, and there may be other boundary identifications. We now apply Construction 2.7 to subdivide the disc into V- and E-faces.

✠ The first coordinate of the measuring quadruple drops by two.

✠ As happened after Construction 3.24, we have to return to Construction 3.5 and repeat all the steps. Since the measuring quadruple is reduced in the first coordinate, eventually it reaches a stage where it cannot drop any more. Now we have the following.

✠ 3.29. CONDITION. *The diagram has no unorientable V-loop.*

Once the algorithm arrives here, it stops, and we do not consider the measuring quadruple any more. We now perform a tidying operation, which alters F, and increases the number of F-faces.

3.30. CONSTRUCTION (Edmonds [3]; type: subdivide X_2, erase-subdivide-relabel). Suppose that there is an orientable V-loop, and hence a pair of orientable V-loops in a prepinching region, by Condition 3.25.

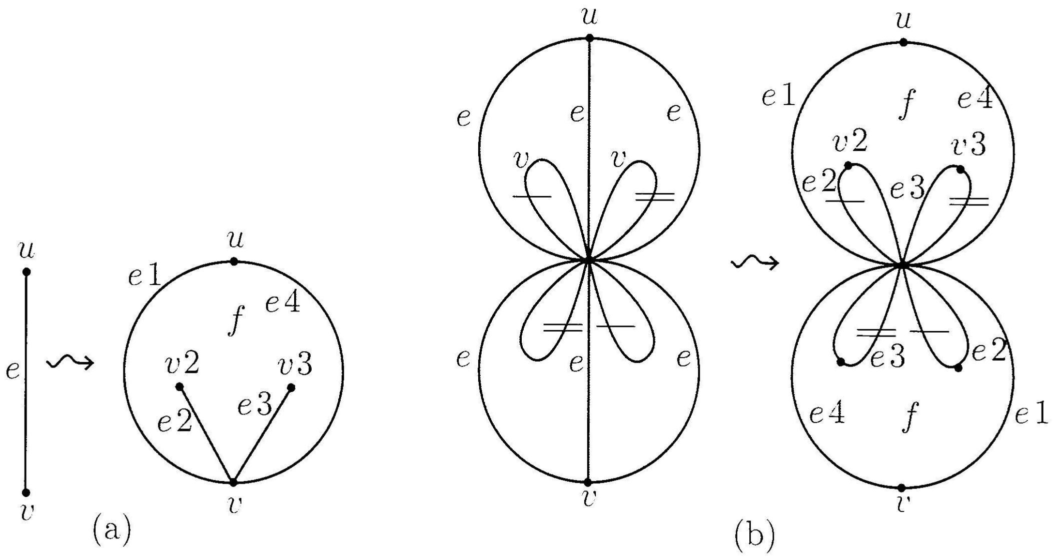


FIGURE 3.12
Changing prepinchings to get a branched cover

It follows from Construction 3.16, and its reverse, that pairs of orientable V-loops are highly mobile, and we can move them into the pair of e-edges incident to w that is given by Condition 3.27, as in the left diagram in Figure 3.12 (b). Now we can subdivide a face of X_2 incident to e, to create a new face f, and new faces and vertices, as in the right diagram in Figure 3.12 (a), or, equivalently, expand the edge e into a disc. We expand each e-edge in

X_1 to a disc in a corresponding manner, with the exception of the six e-edges occurring in the left diagram in Figure 3.12 (b), where we erase, subdivide edges, and relabel, as indicated in Figure 3.12 (b).

We see that all F-faces are well E-joined, and the branching degree is unchanged at each old vertex, and that, at the two added vertices, the branching degree is 2.

Repeating this operation once for every prepinching region, we eliminate all the V-loops, and obtain a diagram in which all faces are F-faces, and all E-joined F-faces are well joined, so it is a diagram representing a branched covering.

The algorithm terminates since we have the Normal Form 3.31 (c).

The foregoing algorithm has proved the main result.

3.31. NORMAL FORM THEOREM (Kneser [11], Edmonds [3], Skora [16]). *If $\beta: X_1 \to X_2$ is a cellular map of CW-surfaces, then there exists a cellular map $\beta': X_1' \to X_2'$ of CW-surfaces homotopic to β whose diagram satisfies one of the following.*

(a) (Degree zero) *The diagram is a union of punctured E-spheres with bouquets of V-loops at the two poles, and the V-loops are identified in pairs in such a way that no two distinct E-spheres with a vertex in common have the same E-label. The identifications have the property that, at each vertex, there is only one face-and-edge cycle, but there can be various labelled cycles. If X_1 is orientable, the identifications can be chosen to respect orientations of the spheres. After collapsing the spheres to edges, there are no faces, and all incident edges have distinct E-labels, yielding an immersion of graphs. Here β has degree zero.*

(b) (Pinching followed by covering) *After pinching, consisting of collapsing the prepinchings to edges, all faces are F-faces, all edge-adjacent pairs are well E-joined, and the branching degree at each vertex is 1, yielding a d-fold covering for some positive integer d. Here β has degree d.*

(c) (Branched covering) *All faces are F-faces, and all edge-adjacent pairs are well E-joined, yielding a d-fold branched covering for some positive integer d. Here β has degree d.*

3.32. COROLLARY (Kneser [10], [11]). *Any cellular map of CW-surfaces of degree 1 is homotopic to a (possibly trivial) pinching.*

The following is an interesting illustration of Theorem 3.31.

3.33. EXAMPLE: SELF-MAPS OF THE REAL PROJECTIVE PLANE. By considering the Puppe exact sequence [12, p. 238], [13, p. 3] associated to a map $S^1 \to S^1$ of degree 2, one finds that each pointed homotopy class of maps from a real projective plane to a real projective plane is determined by its degree, and the possible values are 0, 2, and the odd positive integers. In particular, the same holds for the (unpointed) homotopy classes of maps.

A degree zero map is given by collapsing the source surface to a point. This is of type (a).

A degree two map is given by collapsing an unorientable loop to a point to obtain a two-sphere, and then composing with a double covering of the projective plane. This is an orientation-false pinching composed with a double covering, so is of type (b).

An odd positive integer degree map is given by taking an odd positive degree covering of one Möbius band by another, and then collapsing the boundaries to points. This is a branched covering with a single branch point, so is of type (c).

In the usual way, the homotopy classes of self-maps of the real projective plane form a monoid under composition; to calculate composites one need only calculate the degree, and that can be done easily, even using the algorithm given here. Thus we can identify each homotopy class with its degree, and examine the binary operation induced by composition. We find that the monoid is obtained from the usual multiplicative monoid of non-negative integers by identifying two distinct non-negative integers if and only if they are even and are equal modulo 4.

4. HOMOMORPHISMS OF SURFACE GROUPS

Throughout this section, let $\alpha\colon G_1 \to G_2$ be a homomorphism of infinite surface groups, and $G_1 = \langle S_1 \mid r_1 \rangle$, $G_2 = \langle S_2 \mid r_2 \rangle$ be surface group presentations.

4.1. REVIEW. The arguments of Sections 2, 3 give us a method for finding a normal form for α, and hence for calculating the degree of α.

Let us itemize the steps performed.

We choose a homomorphism of free groups $A\colon \langle S_1 \mid \ \rangle \to \langle S_2 \mid \ \rangle$ which induces α, and we choose, for some non-negative integer d, elements $w_1, \dots, w_d$ of $\langle S_2 \mid \ \rangle$, and elements $\epsilon_1, \dots, \epsilon_d$ in $\{1, -1\}$, such that

$$A(r_1) = \prod_{i=1}^{d} w_i\, r_2^{\epsilon_i}\, w_i^{-1} \text{ in } \langle S_2 \mid \ \rangle. \tag{4.1}$$

We then use A and (4.1) in Construction 2.10 to construct a cellular map $\beta\colon X_1 \to X_2$ realizing α. Here S_1 (resp. S_2) is identified with a basis of the free fundamental group of a specified subgraph of X_1 (resp. the whole one-skeleton of X_2) with a specified base vertex; also, r_1 (resp. r_2) corresponds to the boundary cycle of a certain subdivided face (resp. the unique face). By construction, β restricts to a graph morphism between the specified subgraphs, and the resulting homomorphism of free fundamental groups agrees with A.

We apply the algorithm of Section 3 to β to obtain a new cellular map $\beta'\colon X_1' \to X_2'$ which is in the normal form given by Theorem 3.31. Here X_2' is obtained from X_2 through Constructions 3.1 and 3.30, and there is a natural map from the one-skeleton of X_2' to the one-skeleton of X_2, and both complexes have natural base vertices, and both base vertices will be denoted v_2. By Remarks 2.5, we can trace through the steps of the algorithm and for each transform of X_1, we can identify S_1 with a basis of the fundamental group of a subgraph with a base vertex. Thus for any set E' of edges which corresponds to a maximal subtree of the one-skeleton of the dual complex of X_1', we can identify S_1 with a basis of the fundamental group of the one-skeleton of $X_1' - E'$. Throughout the algorithm S_1 is altered only up to homotopy and change of base vertex. Moreover, up to conjugacy and inverse, r_1 agrees with the boundary cycle of the resulting subdivided large face expressed in terms of the basis S_1. Now β' gives a new lifting A' of α, via the labelling. The boundary label of each face in X_1' corresponds to a conjugate of $r_2^{\pm 1}$, and the subdivided large face gives a description of $A'(r_1)$ as a product of conjugates of $r_2^{\pm 1}$, by viewing the large face as a compressed version of Figure 2.4. Now we get an expression

$$A'(r_1) = \prod_{i=1}^{d'} w_i'\, r_2^{\epsilon_i'}\, w_i'^{-1} \tag{4.2}$$

in which d' is the number of F-faces, and hence equals the degree of α.

For some purposes, it is convenient to have a new generating set S_1' of G_1 adapted to the normal-form map. This can be thought of as a change of basis within the free group, but we prefer to think of it as giving a new free group mapping onto G_1, with a specified isomorphism to the old free group, with the property that the new relator r_1' arising from the boundary cycle of the subdivided face corresponds to r_1, up to conjugacy and inverse.

To define S_1', we first choose a set of edges to erase in X_1' as follows. Choose a maximal forest of E-edge-adjacent faces in X_1, and erase the E-edges, and then choose a maximal tree of V-edge-adjacent faces and erase the V-edges. It is clear from Figure 3.1 that, in the prepinching regions, the interior E-edges get erased, and the interior V-edges do not. In the one-skeleton of the resulting CW-surface, choose a base vertex v_1 which maps to v_2, and choose a maximal tree, and collapse the edges; notice that these are all E-edges, since the V-edges are loops. This gives us a surface group presentation, $\pi(X_1', v_1) = \langle S_1' \mid r_1' \rangle$. Now $\pi(\beta'^{(1)}, v_1)\colon \pi(X_1'^{(1)}, v_1) \to \pi(X_2'^{(1)}, v_2)$ determines a homomorphism $A''\colon \langle S_1' \mid \ \rangle \to \langle S_2 \mid \ \rangle$ of free groups, and we get an equation

$$A''(r_1') = \prod_{i=1}^{d''} w_i'' \, r_2^{\epsilon_i''} \, w_i''^{-1} \tag{4.3}$$

closely related to the normal form, in which d'' is the degree of α.

Here all unerased V-loops, which include all the V-loops occurring in prepinchings, determine elements of S_1' which are sent to 1 under A'. Thus the algorithm gives us a distinguished set $K \subseteq S_1'$ of generators which go to 1.

We now want to examine in detail what can be said in each of the three types of normal norm.

4.2. THE DEGREE ZERO CASE. Suppose case (a) of Theorem 3.31 holds.

Here r_2 loses its significance, and we are studying a homomorphism from a surface group to the free group $\pi(X_2^{(1)}, v_2) = \langle S_2 \mid \ \rangle$.

Form a labelled graph Γ by collapsing each E-sphere to an edge. The labelling immerses Γ in the graph $X_2^{(1)}$, since no two E-spheres at a vertex have the same E-label. In particular, if the induced map of fundamental groups $\pi(\beta, v_1)\colon \pi(X_1, v_1) \to \pi(X_2^{(1)}, v_2)$ is surjective, then the labelling identifies $\Gamma = X_2^{(1)}$.

Our erasing procedure erases all but one E-edge in each punctured E-sphere, and then erases V-loops incident to distinct faces as often as possible, leaving a single face. The one-skeleton is then a copy of Γ with bouquets of V-loops at each vertex. We then collapse a maximal subtree of Γ to a vertex, to obtain the surface group presentation $G_1 = \langle S_1' \mid r_1' \rangle$. Every element of S_1' is either an edge of the collapsed Γ, or is an unerased V-loop. Recall that K denotes the set of elements of S_1' corresponding to unerased V-loops. Then the complement, $S_1' - K$, is in bijective correspondence with the edge set of the collapsed Γ. If we were to collapse the unerased V-loops to vertices, we would have a face with boundary label a relation in the fundamental groupoid of Γ, but this is a free groupoid, so the relation represents a trivial element. That is, r_1' lies in the normal closure of $K \subset S_1'$.

This proves that any surjective homomorphism from a surface group to a free group can be expressed in the form $\langle S_1' \mid r_1' \rangle \to \langle S_1' \mid r_1', K \rangle$ where K is a subset of S_1' whose normal closure contains r_1'.

One can extract even more information from the diagram. For example, it is natural to divide in half all those edges of Γ which lie outside the maximal subtree, and subdivide the edges and faces of X_1' which map to these. This introduces an orientable V-loop around the equator of certain punctured E-spheres, and we can erase one old V-loop for each equator we add. The surface obtained by deleting these equators from X_1' maps to the subtree of Γ obtained by deleting a point from each edge outside the maximal subtree. Hence we have a punctured subsurface which maps to a tree, so its fundamental group is collapsed. The surface X_1' can be recovered from the punctured surface by identifying boundary components in pairs. The effect on the fundamental group is to form an HNN-extension which adds a new generator conjugating one of the boundary components to the other, and the new generator corresponds to one of the specified generators of the fundamental group of Γ. This can be used to give quite a precise normal form, but we are still some distance from recovering all the information that is currently known. Zieschang [17, Satz 2] showed that any surjective homomorphism from an orientable surface group onto a free group can be expressed in the form

$$\begin{aligned} &\langle x_1, y_1, \dots, x_n, y_n \mid (x_1, y_1) \cdots (x_n, y_n) \rangle \\ &\quad \to \langle x_1, y_1, \dots, x_n, y_n \mid (x_1, y_1) \cdots (x_n, y_n), x_1, \dots, x_n, y_{r+1}, \dots, y_n \rangle, \end{aligned}$$

where $0 \leq r \leq n$. Grigorchuk and Kurchanov [7, Theorem 1] showed that any surjective homomorphism from an unorientable surface group onto a free group can be expressed in exactly one of two forms

$$\langle z_1, z_2, \ldots, z_n \mid z_1^2 z_2^2 \cdots z_n^2 \rangle \\ \to \langle z_1, z_2, \ldots, z_r \mid z_1^2 z_2^2 \cdots z_r^2, z_1 z_2, z_3 z_4, \ldots, z_{2r-1} z_{2r} \rangle,$$

where $z_{2r+1}, \ldots, z_n$ are either all sent to z_{2r}, where n is even and $0 < 2r \leq n$, or all sent to 1, where $0 \leq 2r < n$. An elegant proof can be found in [8].

Ol'shanskii [15, Section 2] used diagram techniques similar to those used here to obtain some of the above results independently.

It is interesting to note that V-loops frequently occur in the literature. Edmonds [3] and Skora [16], in the course of their arguments, find it necessary to prove that, for any surface map of degree zero, there exists a non-separating V-loop; Skora uses a non-separating point of the graph Γ, except in the case where Γ is a tree and the map is trivial. Ol'shanskii's arguments for maps from surface groups to free groups are based on proving that there exists a non-collapsable V-loop. Gabai [4] used three-dimensional topology to show that every non-injective homomorphism between surface groups can be represented by a diagram with a non-collapsable V-loop.

We now turn to the nonzero degree case, and describe the group-theoretic formulation of branched covers.

4.3. THE BRANCHED COVERING CASE. Consider any non-negative integers n, m, p, with $m = 0$ or $n = 0$, and positive integers $d_1, \ldots, d_p$. Let

$$G = \langle x_1, y_1, \ldots, x_n, y_n, z_1, \ldots, z_m, t_1, \ldots, t_p \\ \mid (x_1, y_1) \cdots (x_n, y_n) z_1^2 \cdots z_m^2 t_1 \cdots t_p, t_1^{d_1}, \ldots, t_p^{d_p} \rangle.$$

There is a canonical map from G to the surface group

$$G_2 = \langle x_1, y_1, \ldots, x_n, y_n, z_1, \ldots z_m \mid (x_1, y_1) \cdots (x_n, y_n) z_1^2 \cdots z_m^2 \rangle,$$

obtained by annihilating the t_k.

The *Euler characteristic* of G is defined as

$$\chi(G) = 2 - m - 2n - p + \sum_{i=1}^{p} \frac{1}{d_i}.$$

For example, if $p = 0$ then $G = G_2$, and here the Euler characteristic plus the rank equals 2, where the *rank* is the minimum number of generators, or equivalently, the size of the generating set in the surface group presentation.

It is known that G acts, with compact quotient, as a group of isometries on a sphere, plane or hyperbolic disc, depending as $\chi(G)$ is positive, zero, or negative, respectively. Any subgroup H of finite index is again of this form, and the Riemann-Hurwitz formula says that $\chi(H) = (G : H)\chi(G)$.

If we choose a surface subgroup G_1 in G of finite index, then we get a homomorphism of surface groups $G_1 \to G_2$. A homomorphism arising in this way is called a *branched covering homomorphism of surface groups*. It is not difficult to construct the corresponding cellular map of CW-surfaces in normal form, and find that it is a d-fold branched covering, where $d = (G : G_1)$, and the $d_1, \dots, d_p$ can be taken as the branching degrees. Conversely, any cellular map of CW-surfaces which is a branched covering has an associated group homomorphism of this form.

There is an *orientation map* from G to $\{\pm 1\}$ which sends the x_i, y_i, t_i to 1, and the z_i to -1. It follows that branched covering homomorphisms of surface groups are orientation-true, so for infinite surface groups, the value of $\left|\sum_{i=1}^{d} \epsilon_i\, \epsilon(w_i)\right|$ in (4.1) is independent of the lifting chosen.

Let us take presentations and diagrams corresponding to the branched covering. Consider an edge e in E, and a distinguished occurrence of e in the boundary cycle of the single face f in F, and two distinct e-adjacent f-faces, denoted f_i, f_j. These have associated a w_i and a w_j representing paths back to the base vertex, so $w_i^{-1}w_j$ represents a path between the base vertices of f_i and f_j, and for the purposes of checking signs, we may assume the base vertex is incident to e. Since the f-faces are well e-joined, $\epsilon(w_i^{-1}w_j)$ describes whether the two (distinguished) e-edges in the two f-faces would be identified with a twist, or not, that is, have the same, or different, signs, respectively, in the two occurrences in the boundary cycle of f. But $\epsilon_i^{-1}\epsilon_j$ describes whether the two adjacent f-faces have the same orientation of not. Thus $\epsilon(w_i^{-1}w_j) = \epsilon_i^{-1}\epsilon_j$. Hence, for this choice of presentation, $\left|\sum_{i=1}^{d} \epsilon_i\, \epsilon(w_i)\right| = d = (G : G_1)$. In summary, the degree of a branched homomorphism of infinite surface groups is given by $(G : G_1)$.

Let N denote the kernel of $G \to G_2$. Then

$$(G : G_1) \geq (G : G_1N) = (G/N : G_1N/N) = (G_2 : \operatorname{Im} G_1),$$

so the degree is at least $(G_2 : \operatorname{Im} G_1)$.

Notice that the degree is exactly $(G_2 : \operatorname{Im} G_1)$ if and only if $G_1 = G_1 N$, that is, N lies in G_1. But N is generated by torsion elements, and G_1 is torsion-free, so the latter holds if and only if N is trivial, that is, $G = G_2$, which is the case $p = 0$. Here we simply have an inclusion of finite index, which corresponds to an (unbranched) covering.

In case (b) of Theorem 3.31, adding the relations to $\pi(X_1', v_1) = G_1$ which annihilate the pinched generators leaves a surface group presentation. Let us invent terminology to express this.

4.4. THE PINCHING CASE. A *pinching homomorphism of surface groups* is a homomorphism which can be put in the form $\langle S \mid r\rangle \to \langle S \mid r, K\rangle$ where $\langle S \mid r\rangle$ is a surface group presentation, and K is a subset of S such that deleting the occurrences of elements of K from r leaves a word r' such that $\langle S - K \mid r'\rangle$ is a surface group presentation. Notice that the parity of the homomorphism is odd. If some element of K occurs twice with the same sign in r, then the homomorphism is orientation-false, and otherwise it is orientation-true.

It can be shown that a pinching homomorphism of surface groups can be uniquely expressed in the form $\langle S \mid r\rangle \to \langle S \mid r, K\rangle$ where $K \subseteq S$ and exactly one of the following holds:

$$S = \{x_1, y_1, \dots, x_n, y_n\},\ K = \{x_1, y_1, \dots, x_m, y_m\},$$
$$\text{and } r = (x_1, y_1)\cdots(x_n, y_n), \text{ where } 0 \le m \le n;$$
$$S = \{x_1, y_1, \dots, x_n, y_n, z_1, \dots, z_m\},\ K = \{x_1, y_1, \dots, x_n, y_n\},$$
$$\text{and } r = (x_1, y_1)\cdots(x_n, y_n) z_1^2 \cdots z_m^2, \text{ where } 0 \le m,\ 1 \le n;$$
$$S = \{x_1, y_1, \dots, x_n, y_n, z_1, \dots, z_m\},\ K = \{z_1, \dots, z_m\},$$
$$\text{and } r = (x_1, y_1)\cdots(x_n, y_n) z_1^2 \cdots z_m^2, \text{ where } 0 \le m,\ 1 \le n;$$
$$S = \{z_1, \dots, z_m\},\ K = \{z_1, \dots, z_n\}$$
$$\text{and } r = z_1^2 \cdots z_m^2 \text{ where } 0 \le n < m.$$

The first two types are orientation-true, and the last two types are orientation-false.

Suppose now that $\alpha\colon G_1 \to G_2$ factors as a pinching homomorphism of surface groups $\alpha'\colon G_1 \to \operatorname{Im}\alpha$, followed by an inclusion of finite index $\alpha''\colon \operatorname{Im}\alpha \to G_2$. We wish to verify that $\mathcal{G}(\alpha) = (G_2 : \operatorname{Im}\alpha)$.

It is straightforward to construct a lifting A and an equation (4.1) with $d = (G_2 : \operatorname{Im}\alpha)$, so we have $\mathcal{G}(\alpha) \le (G_2 : \operatorname{Im}\alpha)$, and we may assume $\mathcal{G}(\alpha) < (G_2 : \operatorname{Im}\alpha)$. We wish to obtain a contradiction.

Notice that the parity of α' is odd, so α' does not factor through a free group, and hence α itself cannot factor through a free group. Thus $\mathcal{G}(\alpha) > 0$.

Let $d = \mathcal{G}(\alpha)$. We may assume that we started with a lifting A, and an equation (4.1), that is, d is smallest possible. Thus in the process of applying the algorithm of Section 3, we perform no cancellation of F-faces, and we finish with a Normal Form map of degree d. We are not in case (a), since $d > 0$, and we are not in case (b) or (c), since $d < (G_2 : \operatorname{Im}\alpha)$. This is impossible, as desired.

Thus we have proved the following.

4.5. THEOREM (Kneser-Edmonds-Skora). *If $\alpha\colon G_1 \to G_2$ is a homomorphism between infinite surface groups, then exactly one of the following holds.*

(a) *The homomorphism α factors through a surjective homomorphism from G_1 to a free group; here $\mathcal{G}(\alpha) = 0 < (G_2 : \operatorname{Im}\alpha)$.*

(b) *For some positive integer d, α factors as a pinching homomorphism followed by an index d inclusion; here $\mathcal{G}(\alpha) = d = (G_2 : \operatorname{Im}\alpha)$.*

(c) *For some positive integer d, α is a non-injective d-fold branched covering homomorphism of surface groups; here $\mathcal{G}(\alpha) = d > (G_2 : \operatorname{Im}\alpha)$.* $\square$

Notice that in type (b) we have the usual factorization as a surjection followed by a (finite index) inclusion, while in type (c) we have a rather unusual finite index inclusion followed by a surjection. In type (a), we have a special surjection to a free group, with kernel generated by at least half the generators in a suitable surface group presentation, followed by a homomorphism which need not be injective.

4.6. COROLLARY (Kneser [10], [11]). *If a homomorphism between infinite surface groups has degree 1 then it is a (possibly bijective) pinching homomorphism.* $\square$

4.7. COROLLARY. *If G is a surface group with negative Euler characteristic, and α is an endomorphism of G, then either α is an automorphism, or the image of α has infinite index in G, and the kernel of α is generated as normal subgroup by a set consisting of at least half the generators in some surface group presentation of G.*

Proof. If $(G : \operatorname{Im}\alpha)$ is infinite, then $\operatorname{Im}\alpha$ is a free group, and by a Grigorchuk-Kurchanov-Zieschang result recovered in Case 4.2, the kernel of α is generated as normal subgroup by a set consisting of at least half the generators in some surface group presentation of G.

This leaves the case where $\operatorname{Im}\alpha$ has finite index n in G. To see that $n = 1$, we suppose that $n > 1$ and obtain a contradiction as follows. By the Riemann-Hurwitz formula, and the fact that $\chi(G) < 0$, we see that $\chi(\operatorname{Im}\alpha) = n\chi(G) > \chi(G)$, so the rank of $\operatorname{Im}\alpha$ is less than the rank of G. This is impossible, since $\operatorname{Im}\alpha$ is a quotient of G, so $n = 1$. Hence α is surjective.

Since α cannot factor through a group of rank strictly smaller than that of G, we see that α cannot factor through a non-trivial pinching homomorphism. By Theorem 4.5, we see that α is a branched covering homomorphism. Thus G has finite index m in some group

$$H = \Big\langle x_1, y_1, \dots, x_n, y_n, z_1, \dots z_m, t_1, \dots, t_p \\ \Big|\ (x_1, y_1)\cdots(x_n, y_n)z_1^2\cdots z_m^2 t_1\cdots t_p, t_1^{d_1}, \dots, t_p^{d_p}\Big\rangle,$$

where $G = \big\langle x_1, y_1, \dots, x_n, y_n, z_1, \dots z_m \ \big|\ (x_1, y_1)\cdots(x_n, y_n)z_1^2\cdots z_m^2\big\rangle$. By the Riemann-Hurwitz formula, and the fact that $\chi(G) < 0$, we see that $\chi(G) = m\chi(H) \leq \chi(H)$, so $0 \geq \chi(G) - \chi(H) = p - \sum_{i=1}^{p} \frac{1}{d_i} \geq 0$. It follows that $m = 1$, and that α is bijective. $\square$

We can also recover Kneser's description of degree.

4.8. THEOREM (Kneser [10], [11]). *Let $\alpha\colon G_1 \to G_2$ be a homomorphism of infinite surface groups, and consider an equation* (4.1) *arising from some lifting of α.*

(i) *If α is orientation-true, then $\mathcal{G}(\alpha) = \big|\sum_{i=1}^{d} \epsilon_i\,\epsilon(w_i)\big|$, where the map $\epsilon\colon \langle S_2 \mid \ \rangle \to \{\pm 1\}$ is induced from the orientation map of G_2.*

(ii) *If α is orientation-false, and either d is even, or the index $(G_2 : \operatorname{Im}\alpha)$ is infinite, then $\mathcal{G}(\alpha) = 0$.*

(iii) *If α is orientation-false, and d is odd, and $(G_2 : \operatorname{Im}\alpha)$ is finite, then $\mathcal{G}(\alpha) = (G_2 : \operatorname{Im}\alpha)$.*

Moreover, the lifting A can be chosen so that $d = \mathcal{G}(\alpha)$, with the original choice of presentations. $\square$

This result can be used to prove a theorem of Nielsen's which predates Kneser's result.

4.9. THEOREM (Nielsen [14, Section 26], [9]). *If $\langle S \mid r \rangle$ is a surface group presentation of a (surface) group G, and α is an automorphism of G, then there exists an automorphism A of the free group $\langle S \mid \ \rangle$ which maps r to a conjugate of r or r^{-1}, such that the induced map on G is α.*

Proof. This is clear if G is finite, so we may assume that G is infinite. Since α is an automorphism, the kernel is trivial, so α does not have degree zero, and no pinching takes place. Thus α must be a branched covering homomorphism, by Theorem 4.5. We saw in Definition 4.3 that branched covering homomorphisms are orientation-true. It follows from Theorem 4.8 (a), that, among orientation-true maps, the degree is multiplicative with respect to composition. Thus $\mathcal{G}(\alpha)\mathcal{G}(\alpha^{-1}) = \mathcal{G}(1) = 1$. Thus $\mathcal{G}(\alpha) = 1$. By the final part of Theorem 4.8, we can choose a lifting of α to an endomorphism A of the free group on S which sends r to a conjugate of r or r^{-1}. A theorem of Zieschang [17] then shows that A is an automorphism. (A simple proof of surjectivity, using Fox derivatives, is given in Theorem V.4.11 of [1], and injectivity is proved using Nielsen reductions, as in Theorem I.10.5 of [1].) □

The foregoing argument contains elements of the original proof by Nielsen, and of the algebraic proof by Zieschang [17], [18, Corollary 5.4.3].

4.10. REMARKS. Recall that for two groups G_1 and G_2, the set of group homomorphisms from G_1 to G_2 is partitioned into orbits under the natural action of the group $\mathrm{Aut}(G_1)$ via composition. Two homomorphisms in the same orbit are said to be *strongly equivalent*.

Without going into details, let us describe some known results.

Case 4.2, above, mentions surjective homomorphisms from surface groups to free groups. Such homomorphisms have been thoroughly analyzed by algebraic techniques, starting with the work of Zieschang [17], and Ol'shanskii [15], and culminating in the work of Grigorchuk and Kurchanov [7]. This work is distilled in [8] where it is shown that if $\alpha_1, \alpha_2 \colon G_1 \to G_2$ are homomorphisms from a surface group to a free group, then they are strongly equivalent if and only if $\alpha_1(G_1) = \alpha_2(G_1)$ and $\alpha_1(G_1^+) = \alpha_2(G_1^+)$. Together with knowing the maps described in Case 4.1, this allows one to calculate the exact number of strong equivalence classes of surjective homomorphisms from a given surface group to a given free group.

Important work of Gabai and Kazez [5], [6] which uses three-dimensional topology shows that, if $\alpha_1, \alpha_2 \colon G_1 \to G_2$ are nonzero-degree homomorphisms between infinite surface groups, they are strongly equivalent if and only if $\mathcal{G}(\alpha_1) = \mathcal{G}(\alpha_2)$, $\alpha_1(G_1) = \alpha_2(G_1)$ and $\alpha_1(G_1^+) = \alpha_2(G_1^+)$. They also show that, if $\alpha_1, \alpha_2 \colon G_1 \to G_2$ are homomorphisms between surface groups at least one of which is finite, then α_1, α_2 are strongly equivalent if and only if $\mathcal{G}(\alpha_1) = \mathcal{G}(\alpha_2)$, $\alpha_1(G_1) = \alpha_2(G_1)$ and $\alpha_1(G_1^+) = \alpha_2(G_1^+)$.

5. A WORKED EXAMPLE

In this section we will apply the algorithm to a rather trivial example to illustrate the algebraic manipulations involved.

Consider the homomorphism $\alpha \colon \langle a, b, c, d \mid (a, b)(c, d) \rangle \to \langle x, y \mid (x, y) \rangle$ induced by the homomorphism of free groups $A \colon \langle a, b, c, d \mid \ \rangle \to \langle x, y \mid \ \rangle$ determined by $(a, b, c, d) \mapsto (x, y, x, y^{-1})$.

We have

$$\begin{aligned} A((a,b)(c,d)) &= (x,y)(x,y^{-1}) = (x,y)x^{-1}yx(x,y)^{-1}x^{-1}y^{-1}x \\ &= (x,y)^{1-x^{-1}y^{-1}x}. \end{aligned}$$

Since α is orientation-true, Kneser's Theorem 4.8 implies that $\mathcal{G}(\alpha)$ is obtained by applying the orientation map to $1 - x^{-1}y^{-1}x$, so $\mathcal{G}(\alpha) = 0$. Thus we want to apply the algorithm to transform A into a map A' inducing α, such that $A'((a,b)(c,d)) = 1$.

Form the CW-surfaces associated with the given surface group presentations, so the free group generators can be viewed as loops.

Let us subdivide y into two edges, one again called y, and the other called z. We will call the vertices u and v, so that x is a loop at v, y joins v to u, and z joins u to v. The algorithm requires us to subdivide x, but, in order to keep the example simple, we shall not do this. Now we subdivide b and d into two edges labelled $y1$, $z1$, and $y2$, $z2$ respectively. Here the first letter indicates the image label, while, since we plan to depict the moves in planar diagrams, we also want a label to identify equal edges, and it is convenient to use integers for this identification. Similarly, we label c and d as $x1$ and $x2$, respectively.

We first use Construction 2.9 to get a cellular map, and hence a diagram, and then, after some simple applications of Construction 3.5 and 3.11, we can obtain the first diagram in Figure 5.1. Now we can apply the two-step

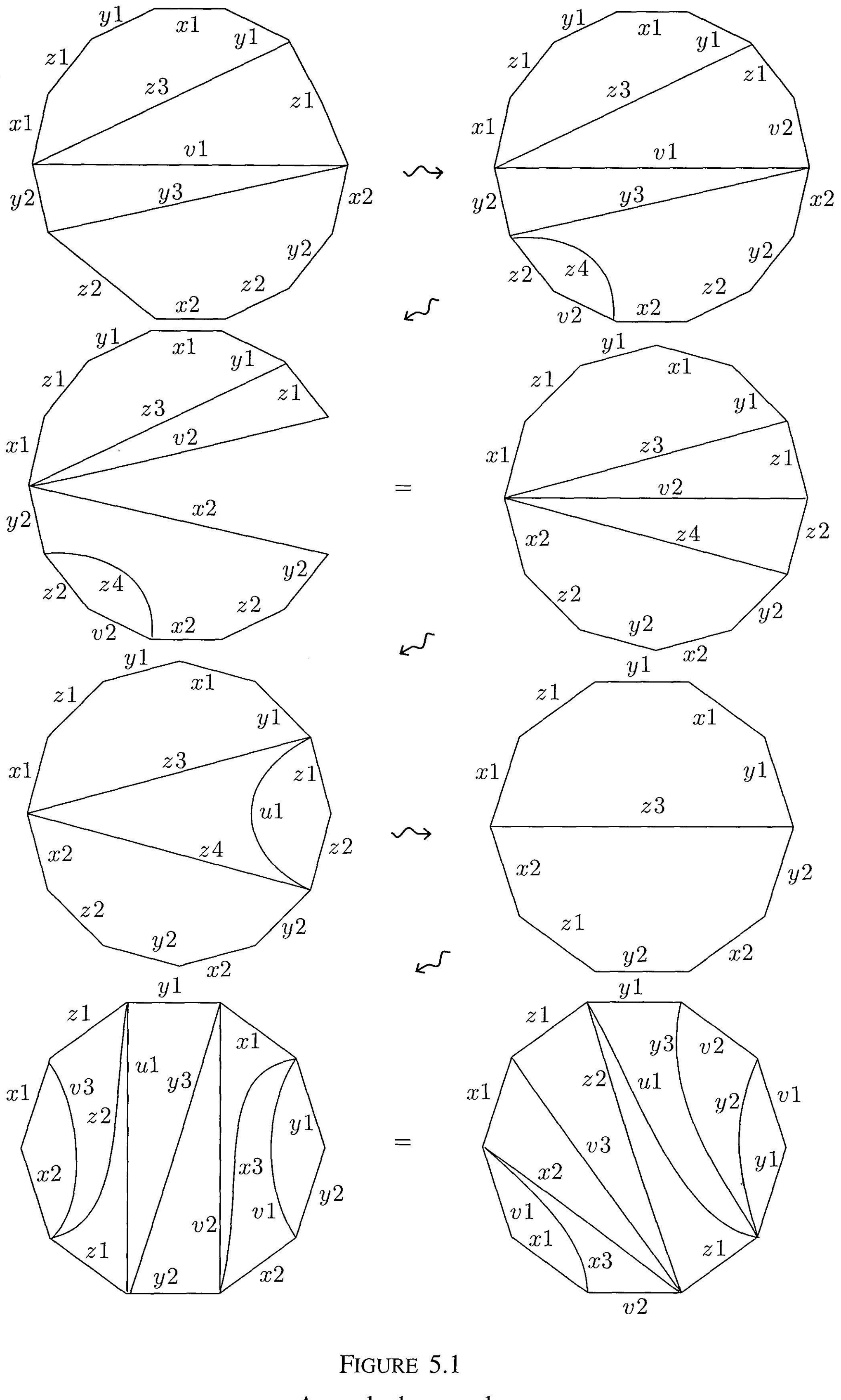


FIGURE 5.1
A worked example

Construction 3.18, to pass from the first to the second, and the second to the third diagram. Thus the first and third diagrams are obtained from the second, by first collapsing $v1$ and $v2$, respectively, and then identifying $z2 = z4$, and $y2 = y3$, respectively. The fourth diagram is a convenient redrawing of the third diagram. Now we apply Construction 3.16 (a) to identify $z1 = z2$ and $z3 = z4$, and arrive at the sixth diagram in Figure 5.1.

Now we can apply Construction 3.24 to arrive at the seventh diagram in Figure 5.1, where we have three punctured spheres, as depicted in Figure 5.2, and we see that $v1$, $v2$ are non-separating non-trivial V-loops, and cutting along these leaves a sphere with four punctures, which can be opened up into a disc by cutting along $x1$, $y1$, and $z1$. Thus we can rearrange the seventh diagram in Figure 5.1 to obtain the eighth diagram.

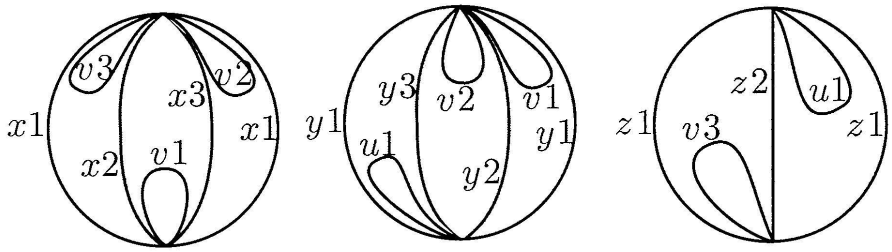


FIGURE 5.2

A normal form

To see what this says about our original group homomorphism, we express all the steps algebraically, by manipulating groupoid presentations.

Here X_2 has only one face, and we have to choose a maximal subtree, and it is natural to choose $\{u, v, z\}$. Let us express this by writing $\langle x, y; z \mid (x, yz)\rangle$, where the edges after the semicolon specify (the edge set of) a maximal subtree among the boundary edges. Recall that for CW-surface fundamental groupoid presentations we do not specify vertices, since they correspond to face-adjacency cycles.

In the same spirit, we express the first diagram in Figure 5.1 as

$$\big\langle x1, y1, x2, y2; z1, z2; v1, y3, z3 \;\big|\; \overline{x1}\,\overline{z1}\,\overline{y1}x1y1z3,\ \overline{z3}\,z1v1, \overline{v1}y3\overline{y2},\ \overline{y3}\,\overline{x2}y2z2x2\overline{z2}\big\rangle,$$

where the edges after the second semicolon specify the edges to be erased to form a single face, and overlines indicate inverses. Here we can identify $a = x1$, $b = y1z1$, $c = x2$, $d = \overline{z2}\,\overline{y2}$.

Now we introduce a new vertex, two new edges $v2$, $z4$, and a new face $z4 = z2v2$, so the second diagram is expressed as

$$\big\langle x1, y1, x2, y2; z1, z2, v2; v1, y3, z3, z4 \;\big|\; \overline{x1}\,\overline{z1}\,\overline{y1}x1y1z3,\ \overline{z3}\,z1v2v1, \overline{v1}y3\overline{y2},\ z4\overline{v2}\,\overline{z2},\ \overline{y3}\,\overline{x2}y2z2x2\overline{z4}\big\rangle,$$

and here $x2$ is an arc, and we identify $a = x1$, $b = y1z1$, $c = x2\overline{v2}$, $d = \overline{z2}\,\overline{y2}$. (We can recover the first diagram by collapsing $v2$ in the maximal subtree, and cashing in the new face relation to identify $z4 = z2$.)

We now collapse the edge $v1$ in the maximal subtree, and cash in the face relation $(v1)y2 = y3$, to identify the two eges $y2 = y3$. This obliges us to choose new edges to erase, and we find that the fourth, and third, diagrams are expressed as

$$\begin{aligned}&\langle x1, y1, x2, y2; z1, z2; z3, z4, v2\\ &\qquad\qquad \mid \overline{x1}\,\overline{z1}\,\overline{y1}x1y1z3,\ \overline{z3}\,z1v2, z4\overline{v2}\,\overline{z2},\ \overline{y2}\,\overline{x2}y2z2x2\overline{z4}\rangle,\end{aligned}$$

and here we identify $a = x1$, $b = y1z1$, $c = x2\overline{v2} = \overline{z2}\,\overline{y2}\,x2y2z2$, $d = \overline{z2}\,\overline{y2}$.

We now re-triangulate, and the fifth diagram can be expressed as

$$\begin{aligned}&\langle x1, y1, x2, y2; z1, z2; z3, z4, u1\\ &\qquad\qquad \mid \overline{x1}\,\overline{z1}\,\overline{y1}x1y1z3,\ \overline{z3}\,\overline{u1}\,z4, u1z1\overline{z2},\ \overline{y2}\,\overline{x2}y2z2x2\overline{z4}\rangle,\end{aligned}$$

and here we identify $a = x1$, $b = y1z1$, $c = \overline{z2}\,\overline{y2}\,x2y2z2$, $d = \overline{z2}\,\overline{y2}$.

We now collapse the edge $u1$, and make identifications using the face relations $z2 = (u1)z1$, $z4 = (u1)z3$, and the sixth diagram can be expressed as $\langle x1, y1, x2, y2; z1; z3 \mid \overline{x1}\,\overline{z1}\,\overline{y1}x1y1z3, \overline{y2}\,\overline{x2}y2z1x2\overline{z3}\rangle$, and here we identify $a = x1$, $b = y1z1$, $c = \overline{z1}\,\overline{y2}\,x2y2z1$, $d = \overline{z1}\,\overline{y2}$.

We now retriangulate, to express relations which map to relations in the free group.

Notice that we have now lifted α to the homomorphism

$$A'\colon \langle a, b, c, d \mid \ \rangle \to \langle x, y \mid \ \rangle$$

determined by $(a, b, c, d) \mapsto (x, y, \bar{y}xy, \bar{y})$, and $A'((a, b)(c, d)) = 1$.

Moreover, by changing presentations, we can now express α in a more natural form. We take the non-separating v-loops $v1 = y1\overline{y2}$ and $v2 = x1y1\overline{y2x2}$, and get the presentation

$$\langle x1, y1, v1, v2; z1 \mid \overline{x1}\,\overline{z1}, \overline{y1}v2\overline{v1}y1z1\overline{v2}x1v1\rangle,$$

and here we identify $x2 = \overline{v2}x1v1$, $y2 = \overline{v1}y1$, so $a = x1$, $b = y1z1$, $c = \overline{z1}\,\overline{y2}\,x2y2z1 = \overline{z1}\,\overline{y1}v1\overline{v2}x1y1z1$, $d = \overline{z1}\,\overline{y2} = \overline{z1}\,\overline{y1}v1$.

Now we can collapse the maximal subtrees, and we have a description of our group homomorphism as follows. We have the genus two surface group $\langle x1, y1, v1, v2 \mid \overline{x1}\,\overline{y1}v2\overline{v1}y1\overline{v2}x1v1\rangle$, we first impose relations annihilating the two generators $v1$, $v2$, to get a free group, and we then impose a relation to get the genus one surface group. Here we can identify $a = x1$, $b = y1$, $c = \overline{y1}v1\overline{v2}x1y1$, $d = \overline{y1}v1$, and thus $bd = v1$, $ab\bar{c}d = v2$. This represents α in one of the normal forms described in Case 4.2.

ACKNOWLEDGEMENTS. We thank Allan Edmonds, Rostislav Grigorchuk, John Guaschi, Boju Jiang, Gilbert Levitt and Joan Porti for information about the literature. We especially thank Carles Casacuberta for showing us Example 3.33, and the referee for much useful advice.

The first author is grateful to the DGICYT (Spain) for support through grant PB93-0900. The second author is grateful to the Centre de Recerca Matemàtica, Barcelona, for hospitality during the Spring term of 1994.

REFERENCES

[1] DICKS, W. and M. J. DUNWOODY. *Groups Acting on Graphs*. Cambridge Studies in Advanced Mathematics, vol. 19. Cambridge University Press, 1989.

[2] DICKS, W. and E. FORMANEK. Automorphism subgroups of finite index in algebraic mapping class groups. *J. Algebra 189* (1997), 58–89.

[3] EDMONDS, A. L. Deformation of maps to branched coverings in dimension two. *Ann. of Math. (2) 110* (1979), 113–125.

[4] GABAI, D. The simple loop conjecture. *J. Differential Geom. 21* (1985), 143–149.

[5] GABAI, D. and W. H. KAZEZ. The classification of maps of surfaces. *Invent. Math. 90* (1987), 219–242.

[6] GABAI, D. and W. H. KAZEZ. The classification of maps of nonorientable surfaces. *Math. Ann. 281* (1988), 687–702.

[7] GRIGORCHUK, R. I. and P. F. KURCHANOV. The classification of epimorphisms from fundamental groups of surfaces to free groups. *Mat. Zametki 48* (1990), 26–35.

[8] GRIGORCHUK, R. I., P. F. KURCHANOV and H. ZIESCHANG. Equivalence of homomorphisms of surface groups to free groups and some properties of 3-dimensional handlebodies. *Cont. Math. 131* (1992), 521–530.

[9] HANSEN, V. L.(ed.) *Jakob Nielsen: Collected Mathematical Papers, Volume 1 (1913-1932)*. Birkhäuser, 1986.

[10] KNESER, H. Glättung von Flächenabbildungen. *Math. Ann. 100* (1928), 609–616.

[11] —— Die kleinste Bedeckungszahl innerhalb einer Klasse von Flächenabbildungen. *Math. Ann. 103* (1930), 347–358.

[12] MAUNDER, C. R. F. *Algebraic Topology*. Cambridge University Press, 1980.

[13] NEISENDORFER, J. Primary homotopy theory. *Mem. Amer. Math. Soc. 25* (1980).

[14] NIELSEN, J. Untersuchungen zur Topologie der geschlossenen zweiseitigen Flächen. *Acta Math. 50* (1927), 189–358.

[15] OL'SHANSKII, A. Y. Diagrams of homomorphisms of surface groups. *Siberian Math. J. 30* (1989), 961–979.

[16] SKORA, R. The degree of a map between surfaces. *Math. Ann. 276* (1987), 415–423.

[17] ZIESCHANG, H. Alternierende Produkte in freien Gruppen. *Abh. Math. Sem. Univ. Hamburg 27* (1964), 13–31.

[18] ZIESCHANG, H., E. VOGT and H.-D. COLDEWEY. *Surfaces and Planar Discontinuous Groups*. Lecture Notes in Mathematics, vol. 835. Springer-Verlag, 1980.

(Reçu le 9 décembre 1996)

Warren Dicks

Departament de Matemàtiques
Universitat Autònoma de Barcelona
08193 Bellaterra (Barcelona)
Spain
e-mail: dicks@mat.uab.es

H. H. Glover

Department of Mathematics
Ohio State University
Columbus, OH 43210
U. S. A.
e-mail: glover@math.ohio-state.edu