THE QUADRATIC ISOPERIMETRIC INEQUALITY FOR MAPPING TORI OF FREE GROUP AUTOMORPHISMS I: POSITIVE AUTOMORPHISMS

MARTIN R. BRIDSON AND DANIEL GROVES

Abstract. If F is a finitely generated free group and ϕ is a positive automorphism of F then $F \rtimes_\phi \mathbb{Z}$ satisfies a quadratic isoperimetric inequality.

Associated to an automorphism ϕ of any group G one has the algebraic *mapping torus* $G \rtimes_\phi \mathbb{Z}$. In this paper we shall be concerned with the case where G is a finitely generated free group, denoted F. We seek to understand the complexity of word problems for the groups $F \rtimes_\phi \mathbb{Z}$ as measured by their Dehn functions.

The class of groups of the form $F \rtimes_\phi \mathbb{Z}$ has been the subject of intensive investigation in recent years and a rich structure has begun to emerge in keeping with the subtlety of the classification of free group automorphisms [3], [4] [5], [18], [23], [26]. (See [1] and the references therein.) Bestvina–Feighn and Brinkmann proved that if $F \rtimes_\phi \mathbb{Z}$ doesn't contain a free abelian subgroup of rank two then it is hyperbolic [2], [15], i.e. its Dehn function is linear. Epstein and Thurston [17] proved that if ϕ is induced by a surface automorphism (in the sense discussed below) then $F \rtimes_\phi \mathbb{Z}$ is automatic and hence has a quadratic Dehn function. The question of whether or not all non-hyperbolic groups of the form $F \rtimes_\phi \mathbb{Z}$ have quadratic Dehn functions has attracted a good deal of attention.

Recall that an automorphism ϕ of a finitely generated free group F is called *positive* if there is a basis $a_1, \dots, a_n$ for F such that the reduced word representing each $\phi(a_i) \in F$ contains no inverses a_j^{-1}.

Main Theorem. *Let F be a finitely generated free group. If ϕ is a positive automorphism of F, then $F \rtimes_\phi \mathbb{Z}$ satisfies a quadratic isoperimetric inequality.*

Modulo a simple change in the interpretation of the symbols used, the proof of this theorem extends *verbatim* to automorphisms ϕ that have a power that admits a train track representative. Not all automorphisms of free groups admit such representatives. Nevertheless, in a subsequent article [11] we use

---

Date: 29 July, 2003.

2000 *Mathematics Subject Classification.* 20F65, (20F06, 20F28, 57M07).

Key words and phrases. free-by-cyclic groups, automorphisms of free groups, isoperimetric inequalities, Dehn functions.

The first author's work was supported in part by an EPSRC Advanced Fellowship.

the relative train-track technology developed by Bestvina, Feighn and Handel ([5] and [3]) and the architecture of the proof of the Main Theorem to establish the quadratic isoperimetric inequality for all groups of the form $F \rtimes \mathbb{Z}$.

Much of our modern understanding of the automorphisms of free groups has been guided by the analogy with automorphisms of surface groups, i.e. mapping classes of surfaces of finite type. This analogy provides a useful reference point when considering the word problems of mapping tori.

A self-homeomorphism of a compact surface S defines an outer automorphism of $\pi_1 S$ and hence a semidirect product $\pi_1 S \rtimes_\phi \mathbb{Z}$. This group is the fundamental group of a compact 3-manifold, namely the mapping torus M_ϕ of the homeomorphism. By using Thurston's Geometrization Theorem for Haken manifolds, Epstein and Thurston [17] were able to prove that $\pi_1 S \rtimes_\phi \mathbb{Z}$ is an automatic group and hence its Dehn function is either linear or quadratic. If S has boundary then only the quadratic case arises. A more geometric explanation for the existence of a quadratic isoperimetric inequality in the bounded case comes from the fact that M_ϕ supports a metric of non-positive curvature, as does any irreducible 3-manifold with non-empty boundary [8], [22].

If S has boundary, then $\pi_1 S$ is free. Thus the foregoing considerations give many examples of free-by-cyclic groups that have quadratic Dehn functions. But there are many types of free group automorphisms that do not arise from surface automorphisms, for example those ϕ that do not have a power leaving any non-trivial conjugacy class invariant, and those ϕ for which there is a word $w \in F$ such that the function $n \mapsto |\phi^n(w)|$ grows like a super-linear polynomial.

The non-automaticity of certain $F \rtimes_\phi \mathbb{Z}$ provides a more subtle obstruction to realising ϕ as a surface automorphism: in contrast to the Epstein-Thurston Theorem, Brady, Bridson and Reeves [6], [14] showed that certain mapping tori $F_3 \rtimes \mathbb{Z}$ are not automatic, for example that associated to the automorphism $[a \mapsto a,\ b \mapsto ab,\ c \mapsto a^2c]$. Such examples show that one cannot proceed via automaticity in order to prove the Main Theorem. Nor can one rely on non-positive curvature, because Gersten [19] showed that the above example $F_3 \rtimes \mathbb{Z}$ is not the fundamental group of any compact non-positively curved space. Thus one needs a new approach to the quadratic isoperimetric inequality.

A technique for dealing with classes of linearly growing automorphisms is described by Brady and Bridson in [6], and Macura [25] developed techniques for dealing with polynomially growing automorphisms. But these techniques apply only to restricted classes of automorphisms and do not speak to the core problem of establishing the quadratic isoperimetric inequality for mapping tori of general free group automorphisms. In the present article and its sequel we attack this core problem directly, undertaking a detailed analysis of the geometry of van Kampen diagrams over the natural presentations of free-by-cyclic groups.

This paper is organised as follows. In Section 1 we recall some basic definitions associated to Dehn functions. In Sections 2 and 3 we record some simple but important observations concerning the large-scale behaviour of the van Kampen diagrams associated to free-by-cyclic groups and in particular the geometry of *corridor* subdiagrams. (The automorphisms considered up to this point are not assumed to be positive.) These observations lead us to a strategy for proving the Main Theorem based on the geometry of the *time flow of corridors*. In Section 4 we state a sharper version of the Main Theorem adapted to this strategy and reduce to the study of automorphisms with stability properties that regulate the evolution of corridors. In Section 5 we develop the notion of *preferred future* which allows us to trace the trajectory of 1-cells in the corridor flow.

The estimates that we establish in Sections 5 and 6 reduce us to the nub of the difficulties that one faces in trying to prove the Main Theorem, namely the possible existence of large blocks of "constant letters". A sketch of the strategy that we shall use to overcome this problem is presented in Section 7. The three main ingredients in this strategy are the elaborate global cancellation arguments in Section 8, the machinery of *teams* developed in Section 9, and the *bonus scheme* developed in Section 10 to accommodate a final tranche of cancellation phenomena whose quirkiness eludes the grasp of teams. In a brief final section we gather our many estimates to establish the bound required for the Main Theorem. A glossary of constants is included for the reader's convenience.

1. Van Kampen Diagrams

We recall some basic definitions and facts concerning Dehn functions and van Kampen diagrams.

1.1. **Dehn Functions and Isoperimetric Inequalities.**

Given a finitely presented group $G = \langle \mathcal{A} \mid \mathcal{R} \rangle$ and a word w in the generators $\mathcal{A}^{\pm 1}$ that represents $1 \in G$, one defines

$$\text{Area}(w) = \min\left\{ N \in \mathbb{N}^+ \mid \exists \text{ equality } w = \prod_{j=1}^{N} u_j^{-1} r_j u_j \text{ in } F(\mathcal{A}) \text{ with } r_j \in \mathcal{R}^{\pm 1} \right\}.$$

The *Dehn function* $\delta(n)$ of the finite presentation $\langle \mathcal{A} \mid \mathcal{R} \rangle$ is defined by

$$\delta(n) \;=\; \max\{\text{Area}(w) \mid w \in \ker(F(\mathcal{A}) \twoheadrightarrow G),\ |w| \leq n\},$$

where $|w|$ denotes the length of the word w. Whenever two presentations define isomorphic (or indeed quasi-isometric) groups, the Dehn functions of the finite presentations are equivalent under the relation $\simeq$ that identifies functions $[0, \infty) \to [0, \infty)$ that only differ by a quasi-Lipschitz distortion of their domain and their range.

For any constants $p, q \geq 1$, one sees that $n \mapsto n^p$ is $\simeq$ equivalent to $n \mapsto n^q$ only if $p = q$. Thus it makes sense to say that the "Dehn function of a group" is $\simeq n^p$.

A group Γ is said to *satisfy a quadratic isoperimetric inequality* if its Dehn function is $\simeq n$ or $\simeq n^2$. A result of Gromov [20], detailed proofs of which were given by several authors, states that if a Dehn function is subquadratic, then it is linear — see [13, III.H] for a discussion, proof and references.

See [9] for a thorough and elementary account of what is known about Dehn functions and an explanation of their connection with filling problems in Riemannian geometry.

1.2. **Van Kampen diagrams.** According to van Kampen's lemma (see [21], [24] or [13, I.8A]) an equality $w = \prod_{j=1}^{N} u_j r_j u_j^{-1}$ in the free group $\mathcal{A}$, with $N = \text{Area}(w)$, can be portrayed by a finite, 1-connected, combinatorial 2-complex with basepoint, embedded in $\mathbb{R}^2$. Such a complex is called a *van Kampen diagram* for w; its oriented 1-cells are labelled by elements of $\mathcal{A}^{\pm 1}$; the boundary label on each 2-cell (read with clockwise orientation from one of its vertices) is an element of $\mathcal{R}^{\pm 1}$; and the boundary cycle of the complex (read with positive orientation from the basepoint) is the word w; the number of 2-cells in the diagram is N. Conversely, any van Kampen diagram with M 2-cells gives rise to an equality in $F(\mathcal{A})$ expressing the word labelling the boundary cycle of the diagram as a product of M conjugates of the defining relations. Thus Area(w) is the minimum number of 2-cells among all van Kampen diagrams for w. If a van Kampen diagram Δ for w has Area(w) 2-cells, then Δ is a called a *least-area* diagram. If the underlying 2-complex is homeomorphic to a 2-dimensional disc, then the van Kampen diagram is called a *disc diagram.*

We use the term *area* to describe the number of 2-cells in a van Kampen diagram, and write Area Δ. We write $\partial\Delta$ to denote the boundary cycle of the diagram; we write $|\partial\Delta|$ to denote the length of this cycle.

Note that associated to a van Kampen diagram Δ with basepoint p one has a morphism of labelled, oriented graphs $h_\Delta : (\Delta^{(1)}, p) \to (\mathcal{C}_\mathcal{A}, 1)$, where $\mathcal{C}_\mathcal{A}$ is the Cayley graph associated to the choice of generators $\mathcal{A}$ for G. The map h_Δ takes p to the identity vertex $1 \in \mathcal{C}_\mathcal{A}$ and preserves the labels on oriented edges.

We shall need the following simple observations.

Lemma 1.1. *If a van Kampen diagram Δ is least-area, then every simply-connected subdiagram of Δ is also least-area.*

Recall that a function $f : \mathbb{N} \to [0, \infty)$ is *sub-additive* if $f(n+m) \leq f(n) + f(m)$ for all $n, m \in \mathbb{N}$. For example, given $r \geq 1$, $k > 0$, the function $n \mapsto kn^r$ is sub-additive.

Lemma 1.2. *Let $f : \mathbb{N} \to [0, \infty)$ be a sub-additive function and let $\mathcal{P}$ be a finite presentation of a group. If* Area $\Delta \leq f(|\partial\Delta|)$ *for every least-area disc diagram Δ over $\mathcal{P}$, then the Dehn function of $\mathcal{P}$ is $\leq f(n)$.*

1.3. **Presenting $F \rtimes \mathbb{Z}$.** We shall establish the quadratic bound required for the Main Theorem by examing the nature of van Kampen diagrams over the following natural (aspherical) presentations of free-by-cyclic groups.

Given a finitely generated free group F and an automorphism ϕ of F, we fix a basis $a_1, \ldots, a_m$ for F, write u_i to denote the reduced word equal to $\phi(a_i)$ in F, and present $F \rtimes_\phi \mathbb{Z}$ by

$$\mathcal{P} \cong \langle a_1, \ldots, a_m, t \mid t^{-1}a_1tu_1^{-1}, \ldots, t^{-1}a_mtu_m^{-1} \rangle. \tag{1.1}$$

We shall work exclusively with this presentation.

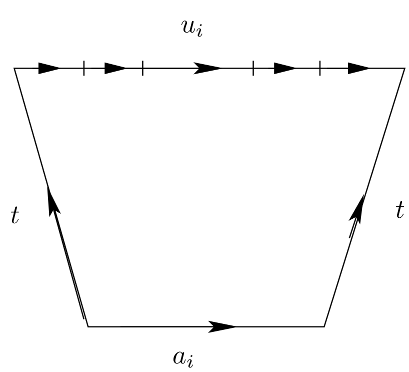

FIGURE 1. A 2-cell in a van Kampen diagram for $F \rtimes_\phi \mathbb{Z}$.

1.4. **Time and t-Corridors with naive tops.** The use of t-corridors as a tool for investigating van Kampen diagrams has become well-established in recent years. In the setting of van Kampen diagrams over the above presentation, t-corridors are easily described.

Consider a van Kampen diagram Δ over the above presentation $\mathcal{P}$ and focus on an edge in the boundary $\partial\Delta$ that is labelled $t^{\pm 1}$ (read with positive orientation from the basepoint). If this edge lies in the boundary of a 2-cell, then the boundary cycle of this 2-cell has the form $t^{-1}a_itu_i^{-1}$ (read with suitable orientation from a suitable point, see Figure 1). In particular, there is exactly one other edge in the boundary of the 2-cell that is labelled t; crossing this edge we enter another 2-cell with a similar boundary label, and iterating the argument we get a chain of 2-cells running across the diagram; this chain terminates at an edge of $\partial\Delta$ which (following the orientation of $\partial\Delta$ in the direction of our original edge labelled $t^{\pm 1}$) is labelled $t^{\mp 1}$. This chain of 2-cells is called a *t-corridor*. The edges labelled t that we crossed in the above description are called the *vertical* edges of the corridor. The vertical edge on

$\partial\Delta$ labelled t^{-1} is called the *initial* end of the corridor, and at the other end one has the *terminal* edge.

Formally, one should define a t-corridor to be a combinatorial map to Δ from a suitable subdivision of $[0,1]\times[0,1]$: the initial edge is the restriction of this map to $\{0\}\times[0,1]$; the vertical edges are the images of the 1-cells of the form $\{s\}\times[0,1]$, oriented so that the edge joining $(s,0)$ to $(s,1)$ is labelled t. The *naive top* of the corridor is the edge-path obtained by restricting the above map to $[0,1]\times\{1\}$, and the *bottom* is the restriction to $[0,1]\times\{0\}$.

Left/Right Terminology: The orientation of a disc diagram induces an orientation on its corridors. Whenever we focus on an individual corridor, we shall regard its initial as being *left*most and its terminal edge as being *right*most. (This is just a suggestive way of saying that the corridor map from $[0,1]\times(0,1)\subset\mathbb{R}^2$ to $\Delta\subset\mathbb{R}^2$ is orientation-preserving.)

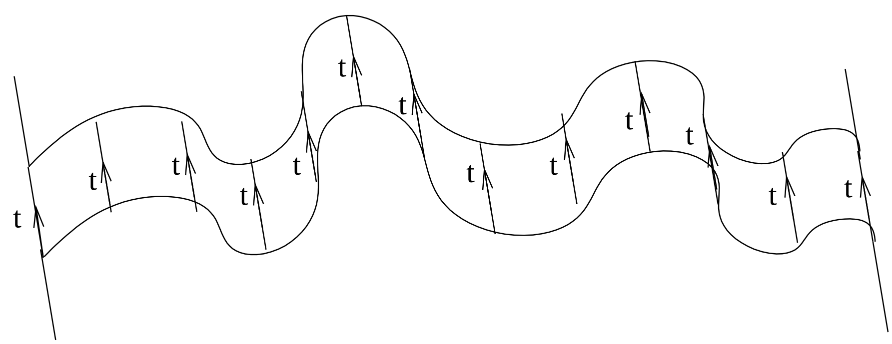

FIGURE 2. A t-corridor

See [10] for a detailed account of t-corridors. Here we shall need only the following easy facts:

(1) distinct t-corridors have disjoint interiors;
(2) if σ is the edge-path in Δ running along the (naive) top or bottom of a t-corridor, then σ is labelled by a word in the letters $\mathcal{A}^{\pm 1}$ that is equal in $F\rtimes\mathbb{Z}$ to the words labelling the subarcs of $\partial\Delta$ which share the endpoints of σ (given appropriate orientations);
(3) if we are in a least-area diagram then the word on the bottom of the corridor is freely reduced;
(4) the number of 2-cells in the t-corridor is the length of the word labelling the bottom side.
(5) In subsection 1.2 we described the map h_Δ associated to a van Kampen diagram. This map sends vertices of Δ to vertices of the Cayley graph $\mathcal{C}_\mathcal{A}$, i.e. elements of $F\rtimes\langle t\rangle$. If the initial vertex of a directed edge in Δ is sent to an element of the form wt^j, with $w\in F$, then the edge

is defined to occur at **time** j. Note that the vertical edges of a fixed corridor all occur at the same time.

We will consider the *dynamics* of the automorphism ϕ.

Definition 1.3 (Time and Length)**.** Item (5) above implies that the time of each t-corridor S is well-defined; we denote it time(S).

We define the *length* of a corridor S to be the number of 2-cells that it contains, which is equal to the number of 1-cells along its bottom. We write $|S|$ to denote the length of S.

1.5. **Conditioning the Diagram.** We are working with the following presentation of $F \rtimes_\phi \mathbb{Z}$

$$\mathcal{P} = \langle a_1, \ldots, a_m, t \mid t^{-1}a_1tu_1^{-1}, \ldots, t^{-1}a_mtu_m^{-1} \rangle.$$

In the light of Lemma 1.2, in order to prove the main theorem it suffices to consider only *disc diagrams*. Therefore, henceforth we shall assume that all diagrams are topological discs. We shall also assume that all of the discs considered are *least-area* diagrams for freely reduced words.

Lemma 1.4. *Every least-area disc diagram over $\mathcal{P}$ is the union of its t-corridors.*

Proof. Since the diagram is a disc, every 1-cell lies in the boundary of some 2-cell. The boundary of each 2-cell contains two edges labelled t. Consider the equivalence relation on 2-cells generated by $e \sim e'$ if the boundaries of e and e' share an edge labelled t. Each equivalence class forms either a t-corridor or else a t-ring, i.e. the closure of an annular sub-diagram whose internal and external cycles are labelled by a word in the generators of F. If the latter case arose, then since F is a free group, the word u on the external cycle would be freely equal to the empty word (since it contains no edges labelled t). This would contradict the hypothesis that the diagram is least-area, because one could reduce its area by excising the simply-connected sub-diagram bounded by this cycle, replacing it with the zero-area diagram for u over the free presentation of F. □

1.6. **Folded Corridors.** In the light of the above lemma, we see that the diagrams Δ that we need to consider are essentially determined once one knows which pairs of boundary edges are connected by t-corridors. However, there remains a slight ambiguity arising from the fact that free-reduction in the free group is not a canonical process (e.g. $x = (xx^{-1})x = x(x^{-1}x)$).

To avoid this ambiguity, we fix a least area disc diagram Δ and assume that its corridors are *folded* in the sense of [7]. The topological closure $T \subset \Delta$ of each corridor is a combinatorial disc. The hypothesis "least area" alone forces the label on the *bottom* of the corridor to be a *freely reduced* word in the letters $a_i^{\pm 1}$. We define the *top* of the (folded) corridor to be the injective edge-path that remains when one deletes from the frontier of T the bottom and ends of the corridor. The word labelling this path is the freely reduced word in F that equals the label on the naive top of the corridor. Note that, unlike the bottom of the corridor, the top may fail to intersect the closure of some 2-cells — see Figures 3 and 4 (where the automorphism is $a \mapsto a, b \mapsto ba^2, c \mapsto ca$).

Notation 1.5. *We write* $\top(S)$ *and* $\bot(S)$*, respectively, to denote the top and bottom of a folded corridor* S*.*

Henceforth we shall refer to folded t*-corridors simply as "corridors".*

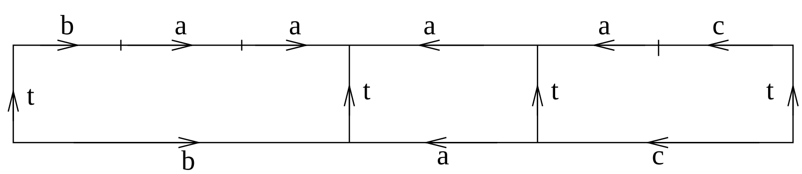

FIGURE 3. An unfolded corridor

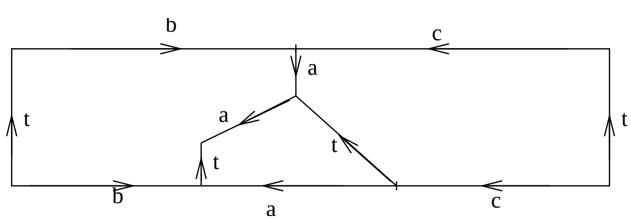

FIGURE 4. The corresponding unfolded corridor.

1.7. **Naive Expansion and Death.** For each generator $a_i \in F$ we have the reduced word $u_i = \phi(a_i)$. Given a reduced word $v = a_{i(1)} \dots a_{i(m)}$ we define the *naive expansion* of $\phi(v)$ to be the (unreduced) concatenation $u_{i(1)} \dots u_{i(m)}$.

Note that if v is the label on an interval of the bottom of a corridor, then the naive expansion of $\phi(v)$ is the label on the corresponding arc of the naive top of the corridor.

An edge ε on the bottom of a corridor S is said to *die* in S if the 2-cell containing that edge does not contain any edge of $\top(S)$. (Equivalently, if w is the label on $\bot(S)$ and a_i is the label on ε, then the subword $u_i = \phi(a_i)$ in the naive expansion of $\phi(w)$ is cancelled completely during the free reduction encoded in Δ.) In Figure 4 the edge labelled a on the bottom of the corridor dies.

2. Singularities and Bounded Cancellation

We have noted that the structure of a (folded, least-area disc) diagram over the natural presentation of a free-by-cyclic group is the union of its corridors. In this section we pursue an understanding of how these corridors meet.

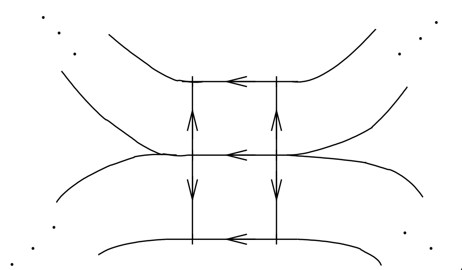

Figure 5. Corridors cannot meet this way in a least-area diagram

The first observation to make is that corridors cannot meet as in Figure 5.

Lemma 2.1. *If $S \neq S'$, then $\perp(S) \cap \perp(S')$ consists of at most one point.*

Proof. For each letter a, there is only one type of 2-cell which has the label a on its bottom side. Thus, if two corridors were to meet in the manner of Figure 5, then we would have a pair of 2-cells whose union was bounded by a loop labelled $u_i t^{-1} t u_i^{-1} t^{-1} t$, which is freely equal to the identity. By excising this pair of 2-cells and filling the loop with a diagram of zero area, we would reduce the area of Δ without altering its boundary label — but Δ is assumed to be a least-area diagram.

Thus $\perp(S) \cap \perp(S')$ contains no edges. To see that it cannot contain more than one vertex, follow the proof of Proposition 2.3(1). □

Definition 2.2. A *singularity* in Δ is a non-empty connected component of the intersection of the tops of two distinct folded corridors. A 2-cell is said to *hit* the singularity if it contains an edge of the singularity.

The singularity is said to be degenerate if it consists of a single point, and otherwise it is *non-degenerate.*

Let M be the maximum of the lengths of the words u_i in our fixed presentation $\mathcal{P}$ of $F \rtimes_\phi \mathbb{Z}$.

Proposition 2.3 (Bounded singularities)**.**

1. *If the tops of two corridors in a least-area diagram meet, then their intersection is a singularity.*
2. *There exists a constant B depending only on ϕ such that less than B 2-cells hit each singularity in a least-area diagram over $\mathcal{P}$.*

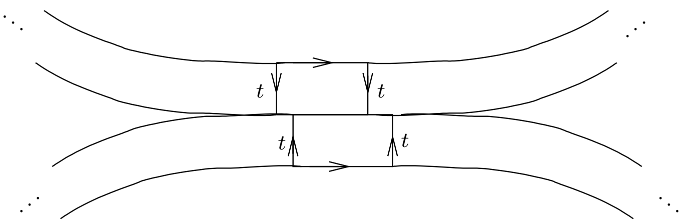

FIGURE 6. A 'singularity'

3. If Δ is a least-area diagram over $\mathcal{P}$, then there are less than $2|\partial\Delta|$ non-degenerate singularities in Δ, and each has length at most MB.

Proof. Suppose that the intersection of the tops of two corridors S and S' contains two distinct vertices, p and q say. Consider the unique subarcs of $\top(S)$ and $\top(S')$ connecting p to q. Each of these arcs is labelled by a reduced word in the generators of F; since the arcs have the same endpoints in Δ, these words must be identical. If the arcs did not coincide, then we could excise the subdiagram that they bounded and replace it with a zero-area diagram, contradicting our least-area hypothesis. This proves (1).

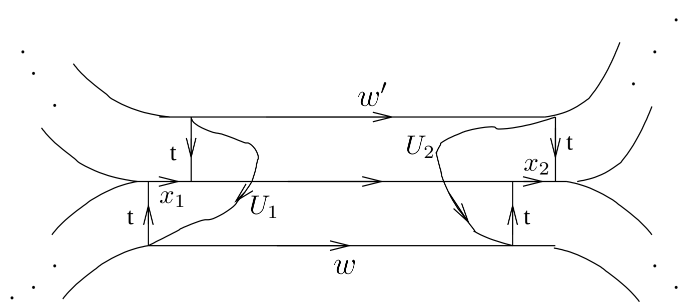

FIGURE 7. The proof of Proposition 2.3

Figure 7 portrays the argument we use to prove (2). In S (respectively S'), we choose an outermost pair of oriented edges $\varepsilon_1, \varepsilon_2$ (resp. $\varepsilon_1', \varepsilon_2'$) labelled t whose termini lie on the singularity. We then connect their endpoints by shortest arcs in the singularity as shown. Note that each of the arcs labelled x_1 and x_2 is contained in the top of a single 2-cell, and hence has length at most M. We write α_i to denote the concatenation of ε_i, the arc labelled x_i and the inverse of ε_i'.

Let $U_i^{-1} \in F$ be the reduced word representing $\phi^{-1}(x_i)$. In $F \rtimes_\phi \mathbb{Z}$ we have $tx_it^{-1}U_i = 1$; let Δ_i be a least-area van Kampen diagram portraying this equality.

Let w (resp. w') be the label on the edge-path in $\bot(S)$ (resp. $\bot(S')$) that connects the initial point of ε_1 (resp. ε_1') to the initial point of ε_2 (resp. ε_2').

If we excise from Δ the subdiagram bounded by the loop whose label is $t^{-1}wtx_2t^{-1}w'^{-1}tx_1^{-1}$, then we reduce the area of Δ by $|w|+|w'|$. (Recall that the edges on the bottom of a corridor are in 1-1 correspondence with the 2-cells of the corridor.) We may then attach a copy of Δ_i along α_i and fill the resulting loop labelled $U_1wU_2^{-1}w'^{-1}$ with a diagram of zero area, because this word is equal to 1 in the free group F. Thus we obtain a new van Kampen diagram whose boundary label is the same as that of Δ and which has area

$$\mathrm{Area}(\Delta)+\mathrm{Area}(\Delta_1)+\mathrm{Area}(\Delta_2)-|w|-|w'|.$$

Since Δ is assumed to be least-area, this implies that $\mathrm{Area}(\Delta_1)+\mathrm{Area}(\Delta_2)\geq |w|+|w'|$.

Let B_0 be an upper bound on the area of all least-area van Kampen diagrams portraying equalities of the form $txt^{-1}\phi^{-1}(x)^{-1}=1$ with $|x|\leq M$. (It suffices to take $B_0=MM_{inv}$, where M_{inv} is the maximum of the lengths of the reduced words $\phi^{-1}(a_i)$.) By definition, $\mathrm{Area}(\Delta_1)+\mathrm{Area}(\Delta_2)\leq 2B_0$, and hence $|w|+|w'|\leq 2B_0$. Thus for (2) it suffices to let $B=2B_0+1$.

The length of the singularity in the above argument is less than the sum of the lengths of the naive expansions of $\phi(w)$ and $\phi(w')$. Since $|w|+|w'|\leq B$, the singularity has length less than MB.

It remains to bound the number of non-degenerate singularities in Δ. To this end, we consider the subcomplex $\Gamma\subset\Delta$ formed by the union of the tops of all folded corridors. Arguing as in (1), we see that the graph Γ contains no non-trivial loops, i.e. it is a forest. Let V denote the set of vertices in Γ that have valence at least 3 or else lie on $\partial\Delta$. (Thus V is the set of degenerate singularities, endpoints of non-degenerate singularities, and endpoints of the tops of corridors.) Let E be the set of connected components of $\Gamma\smallsetminus V$.

$|V|-|E|$ is the number π_0 of connected components of the forest Γ. The valence 1 vertices $V^1\subset\Gamma$ are a subset of the endpoints of the tops of corridors, so there are less than $|\partial\Delta|$ of them. One can calculate $|E|$ as half the sum of the valences of the vertices $v\in V$, so $3(|V|-|V^1|)+|V^1|\leq 2|E|$. Hence

$$|E|=|V|-\pi_0\leq\frac{2}{3}\big(|E|+|V^1|\big)-\pi_0<\frac{2}{3}\big(|E|+|\partial\Delta|\big).$$

Therefore $|E|<2|\partial\Delta|$.

Each non-degenerate singularity determines an element of E, so the (crude) estimate in (3) is established. □

Lemma 2.4 (Bounded Cancellation Lemma)**.** *There is a constant B, depending only on ϕ, such that if I is an interval consisting of $|I|$ edges on the bottom of a (folded) corridor S in a least-area diagram over $\mathcal{P}$, and every edge of I dies in S, then $|I|<B$.*

Proof. The argument is entirely similar to that given for part (2) of the previous proposition. □

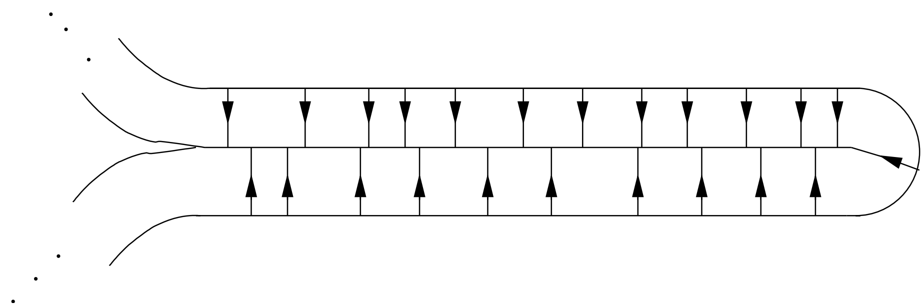

FIGURE 8. Bounded cancellation lemma

The above lemma is a reformulation of the Bounded Cancellation Lemma from [16], which Cooper attributes to Thurston.

Remark 2.5. *'Singularities are only 1 pixel large.'* The reader may find it useful to keep in mind the following picture: think of a least-area van Kampen diagram rendered on a computer screen and assume that the length of the boundary of the diagram is large, so large that the constant B in Proposition 2.3 has to be scaled to something less than 1 pixel in order to fit the picture on to the computer's screen. In the resulting image one sees blocks of t-corridors as shown in Figure 9 below, and the singularities take on the appearance of classical k-prong singularities in the time-flow of t-corridors.

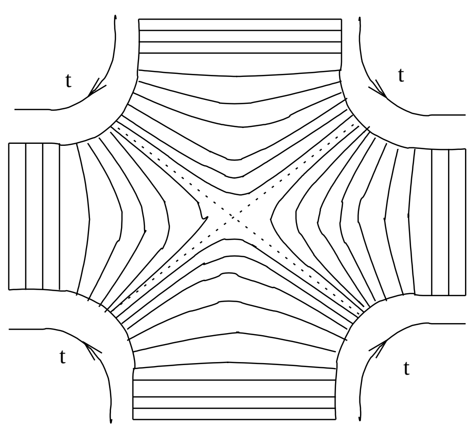

FIGURE 9. Schematic depiction of a singularity

3. PAST, FUTURE AND COLOUR

Our investigations thus far have led us to regard van Kampen diagrams over $\mathcal{P}$ as flows of corridors (at least schematically). We require some more vocabulary to pursue this approach.

We continue to work with a fixed disc diagram Δ over $\mathcal{P}$.

Definition 3.1 (Ancestors and Colour)**.** Each edge ε_1 on the bottom of a corridor either lies in the boundary of Δ, or else lies in the top of a unique 2-cell, the bottom of which we denote ε_0. We consider the partial ordering on the set $\mathcal{E}$ of edges from the bottom of all corridors generated by setting $\varepsilon_0 < \varepsilon_1$ whenever edges are related in this way.

If $\varepsilon' < \varepsilon$ then we call ε' an *ancestor* of ε. The *past* of ε is the set of its ancestors, and the *future* of ε is the set of edges ε'' such that $\varepsilon < \varepsilon''$.

Two edges are defined to be of the same *colour* if they have a common ancestor. Since every edge has a unique ancestor on the boundary, colours are in bijection with a subset[1] of the edges in $\partial\Delta$ whose label is not t; in particular there are less than $|\partial\Delta|$ colours.

Each 2-cell in Δ has a unique edge in the bottom of a corridor. Thus we may also regard $\leq$ as a partial ordering on the 2-cells of Δ and define the past, future and colour of a 2-cell.

We define the past (resp. future) of a *corridor* to be the union of the pasts (resp. futures) of its closed 2-cells.

Remark 3.2. Each $e \in \mathcal{E}$ and each 2-cell has at most one immediate ancestor (i.e. one that is maximal among its ancestors). Consider the graph $\mathcal{F}$ with vertex set $\mathcal{E}$ that has an edge connecting a pair of vertices if and only if one is the immediate ancestor of the other. Note that $\mathcal{F}$ is a forest (union of trees).

The *colours* in the diagram correspond to the connected components (trees) of this forest.

There is a natural embedding of $\mathcal{F} \hookrightarrow \Delta$: choose a point ('centre') in the interior of each 2-cell and connect it to the centre of its immediate ancestor by an arc that passes through their common edge.

If the future of a corridor S' intersects a corridor S then the intersection is connected:

Lemma 3.3 (Connected Pasts)**.** *If a pair of 2-cells α and β in a corridor S have ancestors α' and β' in a corridor S', then every 2-cell γ that lies between α and β in S has an ancestor γ' that lies between α' and β' in S'.*

Proof. Connect the centres of α and β by an arc in the interior of S that intersects only those 2-cells lying between α and β, and connect the centres of α' and β' by a similar arc in the interior of S'. Along with these two arcs, we consider the embedded arcs connecting α to α' and β to β' in the forest $\mathcal{F}$ described in Remark 3.2. These four arcs together form a loop, and the disc that this loop encloses does not intersect the boundary of Δ. (Recall that Δ is a disc.)

[1]namely, those edges of $\partial\Delta$ that lie on the bottom of some 2-cell

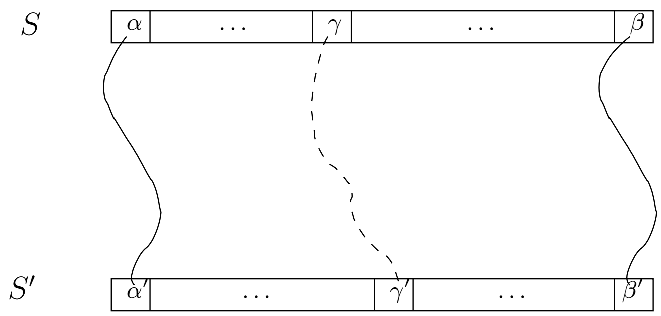

Figure 10. The 'loop' picture

Consider the tree from $\mathcal{F}$ that contains γ. We may assume that the arc in this tree that connects γ to its ancestor on the boundary does not intersect the arc we chose in S. It must therefore intersect our loop either in S', yielding the desired ancestor γ' in S', or else in one of the arcs connecting α to α', or β to β'. If the latter alternative pertains, α' or β' is an ancestor of γ, and we are done. □

We highlight the degenerate case where the 2-cells α' and β' are equal and have their bottom on $\partial\Delta$:

Corollary 3.4. *Within a corridor, the 2-cells of each colour form a connected region.*

4. Strategy, Strata and Conditioning

Everything that has been said up to this point has been true for mapping tori of arbitrary automorphisms of finitely generated free groups. *Henceforth, we assume that the automorphism ϕ is positive.*

A van Kampen diagram whose boundary cycle has length n contains at most $n/2$ corridors. Thus our Main Theorem is an immediate consequence of:

Theorem 4.1. *There is a constant K depending only on ϕ such that each corridor in a least-area diagram Δ over $\mathcal{P}$ has length at most $K\,|\partial\Delta|$.*

In order to establish the desired bound on the length of corridors, we must analyse how corridors grow as they flow into the future, and assess what cancellation can take place to inhibit this growth. In the remainder of this section we shall condition the automorphism to simplify the discussion of growth.

Remark 4.2. The mapping torus $F \rtimes_{\phi^k} \mathbb{Z}$ is isomorphic to a subgroup of finite index in $F \rtimes_\phi \mathbb{Z}$, namely $F \rtimes_\phi k\mathbb{Z}$. Thus, since the Dehn functions of commensurable groups are $\simeq$ equivalent, we are free to replace ϕ by a convenient positive power in our proof of the Main Theorem.

4.1. **Strata.** In the following discussion we shall write x to denote an arbitrary choice of letter from our basis $\{a_1, \ldots, a_m\}$ for F.

Naturally associated to any positive automorphism one has *supports* and *strata*. The support $\mathrm{Supp}(x)$ associated to x is the set of all letters which appear in the freely reduced word $\phi^j(x)$ for some $j \geq 0$. The stratum $\Sigma(x) \subset \mathrm{Supp}(x)$ associated to x consists of those $y \in \mathrm{Supp}(x)$ such that $\mathrm{Supp}(x) = \mathrm{Supp}(y)$.

Note that $y \in \mathrm{Supp}(x)$ implies $\mathrm{Supp}(y) \subseteq \mathrm{Supp}(x)$, and $y \in \Sigma(x)$ implies $\Sigma(y) = \Sigma(x)$.

There are two kinds of strata. The first are *parabolic*[2] *strata*, which are those of the form $\Sigma(x)$ with $x \notin \mathrm{Supp}(y)$ for all $y \in \mathrm{Supp}(x) \smallsetminus \{x\}$. The second kind are *exponential strata*, where one has $\Sigma(x) = \Sigma(y)$ for some distinct x and y. The letter x is defined to be *parabolic* or *exponential* according to the type of $\Sigma(x)$.

If x is exponential then $|\phi^j(x)|$ grows exponentially with j. If all the edges of $\mathrm{Supp}(x)$ are parabolic then $|\phi^j(x)|$ grows polynomially with j. However, it may also happen that x is a parabolic letter but $|\phi^j(x)|$ grows exponentially; this will be the case if $\mathrm{Supp}(x)$ contains letters y such that $\Sigma(y)$ is exponential.

Example 4.3. Define $\phi : F_3 \to F_3$ by $a_1 \mapsto a_1^2 a_2,\ a_2 \mapsto a_1 a_2,\ a_3 \mapsto a_1 a_2 a_3$. Then $\Sigma(a_1) = \Sigma(a_2) = \{a_1, a_2\}$ is an exponential stratum, while $\Sigma(a_3) = \{a_3\}$ is a parabolic stratum with $\mathrm{Supp}(a_3) = \{a_1, a_2, a_3\}$.

Remark 4.4. The relation $[y < x$ if $\Sigma(y) \subset \mathrm{Supp}(x) \smallsetminus \Sigma(x)]$ generates a partial ordering on the letters $\{a_1, \ldots, a_m\}$. For each x, the subgroup of F generated by $\mathrm{Pre}(x) = \{y \mid y < x\}$ is ϕ-invariant. Let $F\lfloor x \rfloor$ denote the quotient of $\langle \mathrm{Supp}(x) \rangle$ by the normal closure of $\mathrm{Pre}(x) \subset \mathrm{Supp}(x)$, and let $F\lceil x \rceil$ denote the quotient of F by the normal closure of $\mathrm{Pre}(x) \subset F$. Note that $F\lfloor x \rfloor$ is a free group with basis (the images of) the letters in $\Sigma(x)$, and $F\lceil x \rceil$ is the free group with basis $\{a_1, \ldots, a_m\} \smallsetminus \mathrm{Pre}(x)$.

The automorphisms of $\mathrm{Pre}(x)$, $F\lfloor x \rfloor$ and $F\lceil x \rceil$ induced by ϕ are positive with respect to the obvious bases, and their strata are images of the strata of ϕ.

4.2. **Conditioning the automorphism.** In the following proposition, the strata considered are those of ϕ^k. (These may be smaller than the strata of ϕ; consider the periodic case for example.)

Proposition 4.5. *There exists a positive integer k such that $\phi_0 := \phi^k$ has the following properties:*

1. *Each letter x appears in its own image under ϕ_0.*
2. *Each exponential letter x appears at least* 3 *times in its own image under ϕ_0.*

[2] Bestvina *et al.* [3] use the terminology *non-exponentially-growing* strata

3. *For all x, each letter $y \in \mathrm{Supp}(x)$ appears in $\phi_0(x)$.*
4. *For all x and all $j \geq 1$, the leftmost and rightmost letters of $\phi_0^j(x)$ are the same as those of $\phi_0(x)$.*
5. *For all x, all $j \geq 1$ and all strata $\Sigma \subseteq \mathrm{Supp}(x)$, the leftmost (respectively, rightmost) letter from Σ in the reduced word $\phi_0^j(x)$ is the same as the leftmost (resp. rightmost) letter from Σ in $\phi_0(x)$.*

Proof. Items (1) to (3) can be seen as simple facts about positive integer matrices, read-off from the action of ϕ on the abelianization of F. (By definition $a_j \in \Sigma(a_i)$ if and only if the (i, j) entry of some power of the matrix describing this action is non-zero.)

Assume that ϕ_1 is a power of ϕ that satisfies (1) to (3). Note that (3) implies that the strata of ϕ_1 coincide with those of any proper power of it.

Replacing ϕ_1 by a positive power if necessary, we may assume that if $\phi_1^j(x)$ begins with the letter x, for any $j \geq 1$, then $\phi_1(x)$ begins with x. This ensures that [$y \preceq_L x$ *if some* $\phi^j(x)$ *begins with* y] is a partial ordering, for if $\phi_1^{j_k}(x_k)$ begins with x_{k+1} for $k = 1, \dots, r$ and if $x_{r+1} = x_1$, then $\phi_1^{\Sigma j_k}(x_1) = x_1$ and hence $x_1 = x_2 = \cdots = x_r$.

If $\phi_1(x)$ begins with z then $z \preceq_L x$, so by raising ϕ_1 to a suitable power we can ensure for all x that $\phi_1(x)$ begins with a letter that is $\preceq_L$-minimal. The $\preceq_L$-minimal letters y are precisely those such that $\phi_1(y)$ begins with y. An entirely similar argument applies to the relation [$y \preceq_R x$ *if some* $\phi^j(x)$ *ends with* y]. This proves (4).

Now assume that ϕ_0 satisfies (1) to (4). The assertion in (5) concerning leftmost letters from Σ is clear for those x where $\phi_0(x)$ begins with x. If $\phi_0(x)$ begins with $y \neq x$, then either $\Sigma \subset \mathrm{Supp}(y)$ or else the occurrences of letters from Σ in $\phi_0^j(x)$ are in 1-1 correspondence with the occurrences in the image of $\phi_0^j(x)$ in $F\lceil y \rceil$. (Notation of Remark 4.4.) In the latter case, arguing by induction on the size of $\mathrm{Pre}(y)$ we may assume that the induced automorphism $\lceil \phi_0 \rceil_y : F\lceil y \rceil \to F\lceil y \rceil$ has the property asserted in (5); the desired conclusion for $\phi_0^j(x)$ is then tautologous. In the former case, if we replace ϕ_0 by ϕ_0^2 then the conclusion becomes as immediate as it was when $\phi_0(x)$ began with x.

An entirely similar argument applies to rightmost letters. □

Remark 4.6. Although we shall have no need of it here, it seems worth recording that item (5) of the above proposition remains true if one replaces strata $\Sigma \subset \mathrm{Supp}(x)$ by supports $\mathrm{Supp}(y) \subset \mathrm{Supp}(x)$.

We now fix an automorphism $\phi = \phi_0$ and assume that is satisfies conditions (1)-(5) above. All of the constants discussed in the sequel will be calculated with respect to this ϕ.

5. Preferred Futures, Fast Letters and Cancellation

Having conditioned our automorphism appropriately, we are now in a position to analyse the fates of (blocks of) edges as they evolve in time.

Definition 5.1 (Preferred futures)**.** For each basis element $x \in \{a_1, \dots, a_n\}$, we choose an occurrence of x in the reduced word $\phi(x)$ to be the (immediate) *preferred future of* x: if x is a parabolic letter, there is only one possible choice; if x is an exponential letter, we choose an occurrence of x that is neither leftmost nor rightmost (recall that we have arranged for x to appear at least three times in $\phi(x)$). More generally, we make a recursive definition of the *preferred future of* x *in* $\phi^n(x)$: this is the occurrence of x in $\phi^n(x)$ that is the preferred future of the preferred future of x in $\phi^{n-1}(x)$.

The above definition distinguishes an edge ε_1 on the top of each 2-cell in our diagram Δ, namely the edge labelled by the preferred future of the label at the bottom ε_0 of the 2-cell. We define ε_1 to be the (immediate) *preferred future* of ε_0. As with letters, an obvious recursion then defines a preferred future of ε_0 at each step in its future (for as long as it continues to exist).

Note that ε_0 has at most one preferred future at each time. (It has exactly one until a preferred future dies in a corridor, lies on the boundary, or hits a singularity.)

If the bottom edge of a 2-cell is ε_0, then we define the preferred future of that 2-cell at time t to be the unique 2-cell at time t whose bottom edge is the preferred future of ε_0.

5.1. Left-fast, constant letters, etc.

We divide the letters $x \in \{a_1^{\pm 1}, \dots, a_m^{\pm 1}\}$ into classes according to the growth of the words $\phi^j(x), j = 1, 2, \dots$, and divide the edges of Δ into classes correspondingly.

- If $\phi(x) = x$ then x is called a *constant letter.*
- If x is a *non*-constant letter, then the function $n \mapsto |\phi^n(y)|$ grows like a polynomial of degree $d \in \{1, \dots, m-1\}$ or else as an exponential function of n.
- Let x be a non-constant letter. If the distance between the preferred future of x and the beginning of the word $\phi^n(x)$ grows at least quadratically as a function of n, we say that x is *left-fast*; if this is not the case, we say that x is *left-slow. Right-fast* and *right-slow* are defined similarly. Note that x is left-fast (resp. slow) if and only if x^{-1} is right-fast (resp. slow).
- Let x be a non-constant letter. If $\phi(x) = uxv$ (the shown occurrence of x need not be the preferred future), where u consists only of constant letters, then we say that x is *left para-linear* . (We place no restriction on v; in particular it may contain occurrences of x.) *Right para-linear* is defined similarly.

Definition 5.2. For left para-linear letters, we define the *(left) para-preferred future* (pp-future) to be the left-most occurrence of x in $\phi(x)$. The (right) pp-future of a right para-linear letter is defined similarly, and edges in Δ inherit these designations from their labels.

(It is possible that a letter might be both left para-linear and right para-linear, and in such cases the left and right pp-futures need not agree. But when we discuss pp-futures, it will always be clear from the context whether we are favouring the left or the right.)

The following lemma indicates the origin of the terminology 'left-fast' (cf. [3, Lemma 4.2.2]). (A slight irritation arises from the fact that there may exist letters x such that x is not left-fast but $\phi(x)$ contains left-fast letters; this difficulty accounts for a certain clumsiness in the statement of the lemma.)

Lemma 5.3. *There exists a constant C_0 with the following property: if $x \in \{a_1, \ldots, a_n\}$ is such that $\phi(x)$ contains a left-fast letter x' and if $UVx \in F$ is a reduced word with V positive*[3] *and $|V| \geq C_0$, then for all $j \geq 1$, the preferred future of x' is not cancelled when one freely reduces $\phi^j(UVx)$. Moreover, $|\phi^j(UVx)| \to \infty$ as $j \to \infty$.*

Proof. We factorize the reduced word $\phi^j(x)$ as $Y_{x,j}x'Z_{x,j}$ to emphasise the placement of the preferred future of a fixed left-fast letter x' from $\phi(x)$. The fact that x' is left-fast implies that $j \mapsto |Y_{x,j}|$ grows at least quadratically.

Fix C_0 sufficiently large to ensure that for each of the finitely many possible $x \in \{a_1, \ldots, a_n\}$, the integer $|Y_{x,j}|$ is greater than Bj whenever $j \geq C_0/B$, where B is the bounded cancellation constant.

The Bounded Cancellation Lemma assures us that during the free reduction of the naive expansion of $\phi(UVx)$, at most B letters of the positive word $\phi(Vx)$ will be cancelled. At most B further letters will be cancelled when the naive expansion of $\phi^2(UVx)$ is freely reduced, and so on. Since V and ϕ are positive and $|V| \geq C_0$, it follows that $\phi^j(V)$ will not be completely cancelled during the free reduction of $\phi^j(UVx)$ if $j \leq C_0/B$. When j reaches $j_0 := \lceil C_0/B \rceil$ the distance from the preferred future of x' to the left end of the uncancelled segment of $\phi^j(Vx)$ is at least $|Y_{x,j_0}|$, which is greater than Bj_0 and hence C_0. Repeating the argument with Y_{x,j_0} in place of V, we conclude that the length of the uncancelled segment of $\phi^j(Vx)$ in $\phi^j(UVx)$ remains positive and goes to infinity with j. □

Significant elaborations of the previous argument will be developed in Section 8.

Definition 5.4 (New edges, cancellation and consumption)**.** Fix a 2-cell in Δ. One edge in the top of the cell is the preferred future of the bottom edge; this will be called *old* and the remaining edges will be called *new*. (These concepts

[3] i.e. no inverses a_j^{-1} appear in V

are unambiguous relative to a fixed 2-cell or (folded) corridor, but 'old edge' would be ambiguous if applied simply to a 1-cell of Δ.)

Two (undirected) edges $\varepsilon_1, \varepsilon_2$ in the naive top of a corridor are said to *cancel* each other if their images in the folded corridor coincide. If ε_1 lies to the left[4] of ε_2, we say that ε_2 has been cancelled *from the left* and ε_1 has been cancelled *from the right*. If ε_1 is the preferred future of an edge ε in the bottom of the corridor and ε_2 is a new edge in the 2-cell whose bottom is ε', then we say that ε' has *(immediately) consumed* ε *from the right*. 'Consumed from the left' is defined similarly.

Let e and e' be edges in $\bot(S)$ for some corridor S, with e to the left (resp. right) of e'. If an edge in the future of e cancels a preferred future of e', then we say that e *eventually consumes* e' *from the left (resp. right).*

Lemma 5.5. *A pair of old edges cannot cancel each other.*

Proof. Suppose that two old edges in the naive top of a corridor S are labelled x and cancel each other. These edges are the preferred futures of edges on $\bot(S)$ that bound an arc α labelled by a reduced word $x^{-1}wx$. Consider the freely-reduced factorisation $\phi(x) = uxv$ where the visible x is the preferred future. The arc in the naive top of S corresponding to α is labelled $v^{-1}x^{-1}u^{-1}Wuxv$, where W is the naive expansion of $\phi(w)$. The old edges that we are considering are labelled by the visible occurrences of x in this word and our assumption that these edges cancel means that the subarc labelled $x^{-1}u^{-1}Wux$ becomes a loop (enclosing a zero-area sub-diagram) in the diagram Δ.

But this is impossible, because $x^{-1}wx$ is freely reduced, which means that W is not freely equal to the empty word, and hence neither is $x^{-1}u^{-1}Wux$. $\square$

Corollary 5.6. *An edge labelled by a parabolic letter x can only be consumed by an edge labelled y with* Supp(x) *strictly contained in* Supp(y).

Remark 5.7. A non-constant letter can only be (eventually) consumed from the left (resp. right) by a right-fast (resp. left-fast) letter.

Remark 5.8. The number of old letters in the naive top of a corridor S is $|S|$, so the length of corridors in the future of S will grow relentlessly unless old letters are cancelled by new letters or the corridor hits a boundary or a singularity.

An obvious separation argument provides us with another useful observation concerning cancellation:

Lemma 5.9. *Let ε_1, ε_2 and ε_3 be three (not necessarily adjacent) edges that appear in order of increasing subscripts as one reads from left to right along the bottom of a corridor. If the future of ε_2 contains an edge of $\partial\Delta$ or of a singularity, then no edge in the future of ε_1 can cancel with any edge in the future of ε_3.*

[4]Recall that corridors have a left-right orientation.

6. Counting Non-constant Letters

In this section we fix a corridor S_0 in Δ and bound the contribution of non-constant letters to the length of $\perp(S_0)$.

6.1. **The first decomposition of S_0.** Choose an edge ε on the bottom of S_0. As we follow the preferred future of ε forward one of the following (disjoint) events must occur:

1. The last preferred future of ε lies on the boundary of Δ.
2. The last preferred future of ε lies in a singularity.
3. The last preferred future of ε dies in a corridor S (i.e. cancels with another edge from the naive top of S).

We shall bound the length of S_0 by finding a bound on the number of edges in each of these three cases.

We divide Case (3) into two sub-cases:

3a. The preferred future of ε dies when it is cancelled by an edge that is not in the future of S_0.

3b. The preferred future of ε dies when it is cancelled by an edge that is in the future of S_0.

6.2. **Bounding the easy bits.** Label the sets of edges in S_0 which fall into the above classes $S_0(1), S_0(2), S_0(3a)$ and $S_0(3b)$ respectively. We shall see that $S_0(3b)$ is by far the most troublesome of these sets.

The first of the bounds in the following lemma is obvious, and the second follows immediately from Proposition 2.3.

Lemma 6.1. $|S_0(1)| \leq |\partial\Delta|$ and $|S_0(2)| \leq 2B\,|\partial\Delta|$.

Lemma 6.2. $|S_0(3a)| \leq B\,|\partial\Delta|$.

Proof. The preferred future of each $\varepsilon \in S_0(3a)$ dies in some corridor in the future of S_0. Since there are less than $|\partial\Delta|/2$ corridors, we will be done if we can argue that the preferred future of at most $2B$ such edges can die in each corridor S.

Lemma 3.3 tells us that the future of S_0 intersects S in a connected region, the bottom of which is an interval I. The Bounded Cancellation Lemma assures us that only the edges within a distance B of the ends of I can be consumed in S by an edge from outside the interval. And by definition, if a preferred future of an edge from $S_0(3a)$ is to die in S, then it must be consumed by an edge from outside I. □

We have now reduced Theorem 4.1 to the problem of bounding $S_0(3b)$, i.e. of understanding cancellation *within* the future of S_0. This will require a great deal of work. As a first step, we further decompose S_0, mingling the above decomposition based on the fates of preferred futures of edges with the natural decomposition of S_0 into colours, as defined in Definition 3.1.

6.3. **The chromatic decomposition of S_0.** We fix a colour μ and write $\mu(S_0)$ to denote the interval of $\perp(S_0)$ consisting of edges coloured μ. We shall abuse terminology to the extent of referring to $\mu(S_0)$ as *a colour*, evoking the mental picture of the 2-cells in S_0 being painted with their respective colours. (Recall that the 2-cells of S_0 are in 1-1 correspondence with the edges of $\perp(S_0)$.)

We shall subdivide $\mu(S_0)$ into five subintervals according to the fates of the preferred futures of edges. To this end, we define $l_\mu(S_0)$ to be the rightmost edge in $\mu(S_0)$ whose immediate future contains a left-fast edge that is ultimately consumed from the left by an edge of S_0, and we define $A_1(S_0,\mu)$ to be the set of edges in $\perp(S_0)$ from the left end of $\mu(S_0)$ to $l_\mu(S_0)$, inclusive. We define $A_2(\mu, S_0) \subset \mu(S_0)$ to consist of the remaining edges in $\mu(S_0)$ whose preferred futures are ultimately consumed from the left by an edge of S_0.

Similarly, we define $r_\mu(S_0)$ to be the leftmost edge $\mu(S_0)$ that has a right-fast edge in its immediate future that is ultimately consumed from the right by an edge of S_0, and we define $A_5(S_0,\mu)$ to be the set of edges in $\perp(S_0)$ from the right end of $\mu(S_0)$ to $r_\mu(S_0)$, inclusive. We define $A_4(\mu, S_0) \subset \mu(S_0)$ to consist of the remaining edges in $\mu(S_0)$ whose preferred futures are ultimately consumed from the right by an edge of S_0.

Finally, we define $A_3(S_0,\mu)$ to be the remainder of the edges in $\mu(S_0)$.

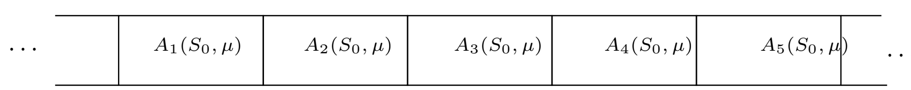

FIGURE 11. The second decomposition of S_0

Modulo the fact that any of the $A_i(S_0,\mu)$ might be empty, Figure 10 is an accurate portrayal of μ: the $A_i(S_0,\mu)$ are connected and they occur in ascending order of suffix from left to right.

The chromatic decomposition of S_0 is connected to the decomposition of Subsection 6.1 by the equality in the following lemma, which is a tautology. The inequality in this lemma is a restatement of Lemmas 6.1 and 6.2.

Lemma 6.3.

$$\bigcup_\mu A_3(S_0,\mu) = S_0 \smallsetminus S_0(3b) \quad \text{and} \quad \sum_\mu |A_3(S_0,\mu)| < (3B+1)\ |\partial\Delta|.$$

Thus the following lemma is a step towards bounding the size of $S_0(3b)$.

Lemma 6.4.

$$|A_1(S_0,\mu)| \le C_0 \quad \text{and} \quad |A_5(S_0,\mu)| \le C_0.$$

Proof. We prove the result only for $A_1(S_0,\mu)$; the proof for $A_5(S_0,\mu)$ is entirely similar.

As in Lemma 5.9, we know that the entire future of the edges of $A_1(S_0, \mu)$ to the left of $l_\mu(S_0)$ must eventually be consumed from the left by edges of S_0. This means that we are essentially in the setting of Lemma 5.3, with $l_\mu(S_0)$ in the role of x and $A_1(S_0, \mu)$ in the role of Vx.

Thus if the length of $A_1(S_0, \mu)$ were greater than C_0, then we would conclude that no left-fast edge in the immediate future of $l_\mu(S_0)$ would be cancelled from the left by an edge of $\perp(S_0)$, contradicting the definition of $l_\mu(S_0)$. □

Corollary 6.5.

$$\sum_\mu |A_1(S_0, \mu)| \leq C_0 \, |\partial\Delta| \quad \text{and} \quad \sum_\mu |A_5(S_0, \mu)| \leq C_0 \, |\partial\Delta|.$$

6.4. **A further decomposition of $A_2(S_0, \mu)$ and $A_4(S_0, \mu)$.** It remains to bound $A_2(S_0, \mu)$ and $A_4(S_0, \mu)$. We deal only with $A_4(S_0, \mu)$, the argument for $A_2(S_0, \mu)$ being entirely similar.

First partition $A_4(S_0, \mu)$ into subintervals $C_{(\mu,\mu')}$ that consist of edges that are eventually consumed by edges of a specified colour μ'. Then partition $C_{(\mu,\mu')}$ into two subintervals: $C_{(\mu,\mu')}(1)$ begins at the right of $C_{(\mu,\mu')}$ and ends with the last non-constant edge; $C_{(\mu,\mu')}(2)$ consists of the remaining (constant) edges. See Figure 12.

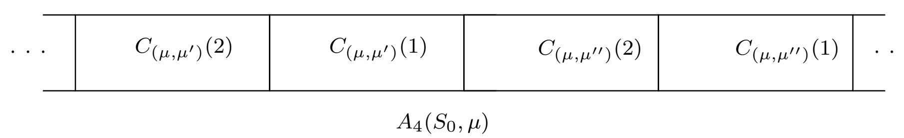

FIGURE 12. $C_{(\mu,\mu')}(1)$ and $C_{(\mu,\mu')}(2)$.

In the course of this section we will bound the size of the intervals $C_{(\mu,\mu')}(1)$ and during the following four sections we bound the sum over all pairs (μ, μ') of the sizes of the intervals $C_{(\mu,\mu')}(2)$ to get the desired bound on $|S_0(3b)|$. In order to control this sum, we have to address the question of which colours can be adjacent.

6.5. **Adjacent Colours.** In Corollary 3.4 we saw that in any corridor S, the edges in $\perp(S)$ of a fixed colour form an interval. We say that two distinct colours μ and μ' are *adjacent* in S if the closed intervals $\mu(S)$ and $\mu(S')$ have a common endpoint in $\perp(S)$. (Equivalently, there is a pair of 2-cells in S, one coloured μ and the other μ', that share an edge labelled t.) We write $\mathcal{Z}$ to denote the set of ordered pairs (μ, μ') such that μ and μ' are adjacent in some corridor S with $\mu(S)$ to the left of $\mu'(S)$ in $\perp(S)$, and we write $\mathcal{Z}$ to denote the set of unordered pairs.

Lemma 6.6.

$$|\mathcal{Z}| < 2\,|\partial\Delta| - 3.$$

Proof. We shall express this proof in the language of the forest $\mathcal{F}$ introduced in Remark 3.2. Suppose that μ and μ' are adjacent in S. In S we can connect the centre of some 2-cell coloured μ to the centre of some 2-cell coloured μ' by an arc contained in the union of the pair of 2-cells. The union of this arc and the trees in $\mathcal{F}$ corresponding to the colours μ and μ' disconnects the disc Δ; each of the other trees in $\mathcal{F}$ is entirely contained in a component of the complement, and the colours with trees in different components can never be adjacent in any corridor.

We can encode adjacencies of colours by a chord diagram: draw a round circle with marked points representing the colours of Δ in the cyclic order that they appear in $\partial\Delta$, then connect two points by a straight line if the corresponding colours are adjacent in some corridor. The final phrase of the preceding paragraph tells us that the lines in this chord diagram do not intersect in the interior of the disc. A simple count shows that since there are less than $|\partial\Delta|$ colours, there are less than $2\,|\partial\Delta| - 3$ lines in this diagram. □

6.6. **Non-constant letters in $C_{(\mu,\mu')}$ that are not left-fast.** We stated in the introduction that a careful analysis of van Kampen diagrams would allow us to reduce the Main Theorem to the study of blocks of constant letters. In this section we achieve the last step of this reduction.

Lemma 6.7. *There is a constant C_1 depending only on ϕ with the following property:*

Let S be a corridor and let μ_1 and μ_2 be colours that occur in S with μ_1 to the left of μ_2 (but do not assume that $\mu_1(S)$ is adjacent to $\mu_2(S)$). Let $I \subset A_4(S,\mu_1)$ be a sub-interval that satisfies the following conditions

1. *the left-most edge of I is non-constant and*
2. *the preferred future of each edge in I is eventually consumed by an edge of $\mu_2(S)$.*

Then $|I| \leq C_1$. In particular, $|C_{(\mu,\mu')}(1)| \leq C_1$ for all $(\mu,\mu') \in \mathcal{Z}$.

It suffices to take $C_1 = 2mB^2$, where m is the rank of F, and B is the constant from the Bounded Cancellation Lemma.

Proof. The region I being considered contains no edge with a right-fast letter in the ϕ-image of its label. Since all exponential letters are both left-fast and right-fast, all non-constant edges in the future of I are parabolic.

We begin the argument at the stage in time where μ_2 starts cancelling I. For notational convenience we assume that this time is in fact time(S). (If it is not, then the fact that the length of I may have increased in passing from time(S) to this time adds greater strength to the bound we obtain.)

We focus on the leftmost edge ε_0 of I that is labelled by a non-constant letter x for which $\mathrm{Supp}(x)$ is maximal among the supports of all edge-labels from I (with respect to inclusion). Let y be the label on the edge ε_0' of $\mu_2(S)$ that eventually consumes ε_0 (oriented as shown in Figure 13). Note that Supp(x) is strictly contained in Supp(y), by Corollary 5.6. If ε_0' consumes ε_0 immediately, then the bounded cancellation lemma tells us that ε_0 is a distance less than B from the righthand end of I. If not, then we proceed one step into the future[5] and appeal to the conditioning done in Proposition 4.5(5) to assume that for all $j \geq 1$, the rightmost letter in $\phi^j(y)$ whose support includes x is y. We shall call the edge in the future of ε_0' carrying the rightmost y the *highlighted* future of ε_0' (perhaps it is not the preferred future).

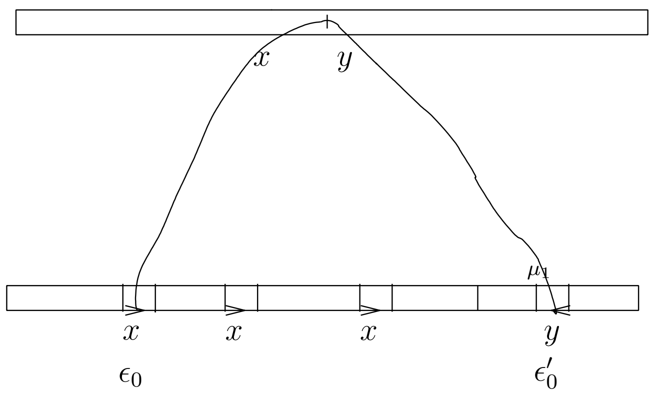

FIGURE 13. The edge labelled ϵ_0' will eventually consume ϵ_0.

The first important point to observe is that the maximality of Supp(x) ensures that there will never be any new edges labelled x in the future of I ('new' in the sense of 5.4).

The second important point to note is that the edges labelled x in the future of ε_0' that are to cancel with the futures of the edges labelled x in I must all lie to the left of the highlighted future of ε_0'. The point here is that the highlighted future of ε_0' cannot be cancelled by an edge of I (by the maximality of x), and in order for it to be cancelled from the other side, all the edges to its right labelled x would have to be cancelled first, which would mean that they too were cancelling with something not in the future of I.

We now come to the key observation of the proof: at each stage j steps into the future of S, the leftmost[6] edge ε_j' in the future of ε_0' that is labelled x must be cancelled by an edge from the future of I *immediately*, i.e. in the corridor where it appears at time$(S)+j$. Indeed if this were not the case, then ε_j' would develop a preferred future which, being an old edge (in the sense of Definition 5.4), could only cancel with a new edge (Lemma 5.5) in the future of I. And since we have arranged that there be no new edges labelled x, the

[5]proceeding one step into the future also allows us to assume that there are no letters coloured μ_1 to the right of I

[6]we have already noted that this is to the left of the highlighted future of ε_0'

preferred future of ε'_j would never cancel with an edge in the future of I. But this cannot be, because the continuing existence of a preferred future for ε'_j would prevent anything to its *right* consuming an edge in the future of I, and the penultimate sentence in the third paragraph of this proof implies that no new edges labelled x will ever appear to its *left* in the future of ε'_0. Thus if ε'_j is not cancelled immediately then we have a contradiction to the fact that ε'_0 must eventually consume ε_0.

We have just proved that at time$(S) + j$ the edge ε'_j must cancel with the preferred future of an edge ε_j in I that is labelled x. According to the Bounded Cancellation Lemma, the preferred future of ε_j at (time$(S) + j - 1$) must lie within a distance B of the right end of the future of I. Since there is no cancellation within the future I, an iteration of this argument shows that for as long as there exist edges labelled x in the future of I, each successive pair of these edges is separated by less than $B + |\phi(y)| \leq 2B$ edges at each moment in time, and the rightmost must be within a distance B of the right end of the future of I.

But since $\phi(x)$ contains at least one letter other than the preferred future of x, it follows that there cannot be a pair of edges of I labelled x that remain unconsumed at time$(S) + 2B$, for otherwise they would have grown a distance more than $2B$ apart, contradicting the conclusion of the previous paragraph. And proceeding one more step into the future, the last edge labelled x must be consumed.

Since at most B letters of I are cancelled at the right at each stage in its future, all of the edges of I labelled x are within a distance less than $2B^2$ of the right end of I, and they are all consumed when I has flowed $2B$ steps into the future. If no non-constant edges remain in the future of I at this stage, then we know that $|I| \leq 4B^2$.

If there do remain non-constant edges, we take the maximal interval of the future of I at time$(S)+2B$ whose leftmost edge is non-constant, and we repeat the argument. (This interval is obtained from the complete future of I by removing a possibly-empty collection of constant edges at its left extremity.)

We proceed in this manner. The interval that we begin with at each iteration has strictly fewer strata than the previous one and therefore the procedure stops before $m = \text{rank}(F)$ iterations. At the time when it stops (at most time$(S) + 2mB$), the future of I has been cancelled entirely, except possibly for a block of constant edges at its left extremity. With one final appeal to the bounded cancellation lemma, we deduce that $|I| \leq 2mB^2$. □

Corollary 6.8.

$$\sum_{(\mu,\mu')\in\mathcal{Z}} |C_{(\mu,\mu')}(1)| < 2C_1\,|\partial\Delta|.$$

Proof. This follows immediately from Lemmas 6.6 and 6.7. □

7. The Bound on $\sum_{\mu\in S_0} |A_4(S_0,\mu)|$ and $\sum_{\mu\in S_0} |A_2(S_0,\mu)|$

The sum of our previous arguments has reduced us to the nub of the difficulties that one faces in trying to prove the Main Theorem, namely the possible existence of large blocks of constant letters in the words labelling the bottoms of corridors. Now we must obtain a bound on

$$\sum_{(\mu,\mu')\in\mathcal{Z}} |C_{(\mu,\mu')}(2)|$$

that will enable us to bound $\sum_{\mu\in S_0} |A_4(S_0,\mu)|$ and[7] $\sum_{\mu\in S_0} |A_2(S_0,\mu)|$ by a linear function of $|\partial\Delta|$. These are the final estimates required to complete the proof of the Main Theorem — see Section 11 for a résumé of the proof.

The regions $C_{(\mu,\mu')}(2)$ are static, in the sense that they do not change under iteration by ϕ, so the considerations of future growth that helped us so much in previous sections cannot be brought to bear directly. Rather, we must analyse the complete history of blocks of constant letters, understand how large blocks come into existence, and use global considerations to limit the sum of the sizes of all such blocks.

Because of the global nature of the arguments, we shall not obtain bounds on the sizes of the individual sets $C_{(\mu,\mu')}(2)$. Instead, we shall identify an associated block of constant letters elsewhere in the diagram (a "team") that is amenable to a delicate string of balancing arguments that facilitates a bound on a union of associated regions $C_{(\mu,\mu')}(2)$.

Our strategy is motivated by the following considerations. Believing Theorem 4.1 to be true, we seek payment from the global geometry of Δ to compensate us for having to handle the troublesome blocks of constant edges $C_{(\mu,\mu')}(2)$; the currencies of payment are *consumed colours* and dedicated subsets of edges on $\partial\Delta$ — since Δ can have at most $|\partial\Delta|$ of each, if we prove that adequate payment is available then our troubles will be bounded and the Main Theorem will follow. The chosen currencies are apposite because, as we shall see in Section 8, a large block of edges labelled by constant letters can only come into existence if a colour (or colours) associated to a component of this block in the past was consumed completely, or else the boundary of Δ intruded into the past of the block (or else something nearby) causing smaller regions of constant edges to elide.

In the remainder of this section we shall explain how various estimates on the behaviour of blocks of constant letters in Δ can be combined to obtain the bounds that we require on $\sum_{\mu\in S_0} |A_4(S_0,\mu)|$ and $\sum_{\mu\in S_0} |A_2(S_0,\mu)|$. We hope that this explanation will provide the diligent reader with a useful road map and

[7] In practice we only need concern ourselves with A_4, the arguments for A_2 being entirely similar

sufficient motivation to sustain them through the many technicalities needed to establish the estimates in subsequent sections.

In the following proposition, M is the maximum length of the images $\phi(x)$ of the basis elements of F, while T_1 is the constant from the Pincer Lemma 8.26, and C_1 is the upper bound on the lengths of the intervals $C_{(\mu,\mu')}(1)$ from Lemma 6.7, T_0 comes from the Two Colour Lemma 8.4 and C_4 comes from Lemma 9.4. The constant λ_0 is defined above Definition 8.22, and B is the bounded cancellation constant from Lemma 2.4.

The Constant K_1 is defined to be

$$2C_1+6\lambda_0+2B(5T_0+6T_1+2)+2MC_4(6T_1+8T_0+3)+(B+3)(3T_1+2T_0)M+5M+2.$$

Proposition 7.1.

$$\sum_{\mu\in S_0} |A_4(S_0,\mu)| \leq K_1\,|\partial\Delta|.$$

7.1. **Dramatis Personae.** The "proof" that we are about to present is essentially a scheme for reducing the proposition to a series of technical lemmas that will be proved in Sections 9 and 10. These lemmas are phrased in the language associated to *teams*, the precise definition of which will also be given in Section 9. Many of the proofs involve global cancellation arguments based on the *Pincer Lemma*, which will be proved in the next section. Intuitively speaking, a *team* (typically denoted $\mathcal{T}$) is a contiguous region of $\|\,\mathcal{T}\,\|$ constant letters all of which are to be consumed by a fixed left para-linear edge (the *reaper*). Notwithstanding this intuition, it is preferable for technical reasons to define a team to be a set of pairs of colours $(\mu,\mu')\in\mathcal{Z}$, where μ' is fixed and the different *members* of the team correspond to different values of μ. We write $(\mu,\mu')\in\mathcal{T}$ to denote membership. Teams also have *virtual members*, denoted $(\mu,\mu')\in_v\mathcal{T}$ (see Definition 9.8). There are less than $2\,|\partial\Delta|$ teams (Lemma 9.10).

Each pair (μ,μ') with $C_{(\mu,\mu')}(2)$ non-empty is either a member or a virtual member of a team (Lemma 9.10). There are *short* teams (Definition 9.6) and long teams, of which some are *distinguished* (Lemma 9.29). There are four types of *genesis* of a team, (G1), (G2), (G3) and (G4) (see Subsection 9.2). Teams of genesis (G3) have associated to them a pincer $\Pi_{\mathcal{T}}$ (Definition 9.12) yielding an auxiliary set of colours $\chi(\Pi_{\mathcal{T}})$. There is also a set of colours $\chi_P(\mathcal{T})$ associated to the time before the pincer $\Pi_{\mathcal{T}}$ comes into play. For long, undistinguished teams, we also need to consider certain sets $\chi_c(\mathcal{T})$ and $\chi_\delta(\mathcal{T})$ of colours consumed in the past of $\mathcal{T}$ (see the proof of Lemma 9.29). Such teams may also have three sets of edges in $\partial\Delta$ associated to them: $\partial^{\mathcal{T}}$, $\text{down}_1(\mathcal{T})$ and $\text{down}_2(\mathcal{T})$. An important feature of the definitions of $\partial^{\mathcal{T}}$ and $\text{down}_1(\mathcal{T})$ is that the sets associated to different teams are disjoint. This disjointness is crucial in the following proof, where we use the fact that the sum of their cardinalities is at most $|\partial\Delta|$. Similarly, the disjointness of the sets $\chi_c(\mathcal{T})$ is used to estimate the sum of their cardinalities by $|\partial\Delta|$ and likewise for $\chi_\delta(\mathcal{T})$ and $\chi_P(\mathcal{T})$.

It is not necessarily true that the sets $\mathrm{down}_2(\mathcal{T})$ are disjoint for different teams, but we shall explain how to account for the amount of 'double-counting' that can occur (see Lemma 9.29).

Associated to every team one has the time $t_1(\mathcal{T})$ at which the reaper starts consuming the team (see Subsection 9.1). Teams genesis (G3) also have two earlier times $t_2(\mathcal{T})$ and $t_3(\mathcal{T})$ associated to them as well as an auxiliary set of edges $Q(\mathcal{T})$, the definitions of which are somewhat technical (see Definition 9.13 *et seq.*).

In Section 10 we describe a *bonus scheme* that assigns a set of extra edges, $\mathrm{bonus}(\mathcal{T})$ to each team. These bonuses are assigned so as to ensure that $|\mathrm{bonus}(\mathcal{T})| + \parallel \mathcal{T} \parallel$ dominates the sum of the cardinalities of the sets $C_{(\mu,\mu')}(2)$ associated to the members and virtual members of $\mathcal{T}$.

Proof of Proposition 7.1.

Recall that $A_4(S_0,\mu)$ is partitioned into disjoint regions $C_{(\mu,\mu')}$ which in turn are partitioned into $C_{(\mu,\mu')}(1)$ and $C_{(\mu,\mu')}(2)$.

Given any μ_1 and μ_2, at most one ordering of $\{\mu_1,\mu_2\}$ can arise in S_0. Thus Lemma 6.6 implies that there are less than $2\,|\partial\Delta|$ pairs $(\mu,\mu')\in\mathcal{Z}$ with $C_{(\mu,\mu')}\subset\perp(S_0)$ non-empty. It follows immediately from this observation and Lemma 6.7 that

$$\sum_{(\mu,\mu')\in\mathcal{Z}} |C_{(\mu,\mu')}(1)| \leq 2C_1\,|\partial\Delta|.$$

Lemma 9.29 accounts for the set of distinguished long teams $\mathrm{D}\Lambda$:

$$\sum_{\mathcal{T}\in\mathrm{D}\Lambda}\ \sum_{(\mu,\mu')\in\mathcal{T}} |C_{(\mu,\mu')}(2)| \leq 6B\,|\partial\Delta|(T_1+T_0).$$

For all other teams $\mathcal{T}$ we rely on Lemma 10.2 which states

$$\sum_{(\mu,\mu')\in\mathcal{T}\ \text{or}\ (\mu,\mu')\in_v\mathcal{T}} |C_{(\mu,\mu')}(2)| \leq\ \parallel\mathcal{T}\parallel + |\mathrm{bonus}(\mathcal{T})| + B. \tag{7.1}$$

We next consider the *genesis* of teams. All teams of genesis (G4) are short (Lemma 9.7). And by Definition 9.6 for the short teams $\mathcal{T}\in\Sigma$ we have

$$\sum_{\mathcal{T}\in\Sigma}\ \sum_{(\mu,\mu')\in\mathcal{T}} |C_{(\mu,\mu')}(2)| \leq 2\lambda_0\,|\partial\Delta| + \sum_{\mathcal{T}\in\Sigma}\big(|\mathrm{bonus}(\mathcal{T})| + B\big).$$

Lemma 9.20 tells us that for teams of genesis (G1) and (G2) we have

$$\parallel\mathcal{T}\parallel\ \leq 2MC_4|\mathrm{down}_1(\mathcal{T})| + |\partial^{\mathcal{T}}|,$$

whilst for teams of genesis (G3) we have

$$\parallel\mathcal{T}\parallel\ \leq 2MC_4\big(|\mathrm{down}_1(\mathcal{T})| + |Q(\mathcal{T})|\big) + T_0\big(|\chi_P(\mathcal{T})|+1\big) + |\partial^{\mathcal{T}}| + \lambda_0.$$

Let $\mathcal{G}_3$ denote the set of teams of genesis (G3) with $Q(\mathcal{T})$ non-empty. In Definition 9.25 we break $Q(\mathcal{T})$ into pieces so that

$$|Q(\mathcal{T})| \;= t_3(\mathcal{T}) - t_2(\mathcal{T}) + |\mathrm{down}_2(\mathcal{T})|.$$

Making crucial use of the Pincer Lemma, in Corollary 9.24 we prove that

$$\sum_{\mathcal{T}\in\mathcal{G}_3} t_3(\mathcal{T}) - t_2(\mathcal{T}) \;\leq\; 3T_1\,|\partial\Delta|,$$

and in Corollary 9.31 we prove that

$$\sum_{\mathcal{T}\in\mathcal{G}_3} |\mathrm{down}_2(\mathcal{T})| \leq (2+3T_1+5T_0)\,|\partial\Delta|.$$

This completes the estimate on $|Q(\mathcal{T})|$ and hence $\|\,\mathcal{T}\,\|$.

Section 10 is dedicated to the proof of Proposition 10.13, which states

$$\sum_{\text{teams}} |\mathrm{bonus}(\mathcal{T})| \leq \big((B{+}3)(3T_1{+}2T_0)M{+}6BT_1{+}4BT_0{+}2\lambda_0{+}2B{+}5M{+}1\big)\,|\partial\Delta|.$$

Adding all of these estimates and recalling that there are less than $2\,|\partial\Delta|$ teams, we deduce:

$$\sum_{\mu\in S_0} |A_4(S_0,\mu)| \leq K_1\,|\partial\Delta|,$$

where K_1 is

$$2C_1{+}6\lambda_0{+}2B(5T_0{+}6T_1{+}2){+}2MC_4(6T_1{+}8T_0{+}3){+}(B{+}3)(3T_1{+}2T_0)M{+}5M{+}2.$$

Thus the proposition is proved. □

Remark 7.2. The stated value of the constant K_1 is an artifact of our proof: we have simplified the estimates at each stage for the sake of clarity rather than trying to optimise the constants involved. Nevertheless, we have made some effort to make the arguments constructive so as to prove that there exists an algorithm to calculate the Dehn function of $F\rtimes_\phi\mathbb{Z}$ directly from ϕ. This is explained in some detail in [12].

By a precisely analogous argument, we also have

Proposition 7.3.

$$\sum_{\mu\in S_0} |A_2(S_0,\mu)| \leq K_1\,|\partial\Delta|,$$

where K_1 is the constant defined prior to Proposition 7.1.

8. The pleasingly rapid consumption of colours

This section contains the cancellation lemmas that we need to control the manner in which colours are consumed. The key result in this direction is the *Pincer Lemma* (Theorem 8.26).

8.1. The Buffer Lemma.

Lemma 8.1. *Let $I \subset \bot(S)$ be an interval of edges labelled by constant letters, and suppose that the colours $\mu_1(S)$ and $\mu_2(S)$ lie either side of I, adjacent to it. Provided that the whole of I does not die in S, no non-constant edge coloured μ_1 will ever cancel with a non-constant edge coloured μ_2.*

Proof. Suppose that the future of I in $\top(S)$ is a non-empty interval labelled w_0. If $\mu_1(S)$ is to the left of I, then reading from the left beginning with the last non-constant edge coloured μ_1, on the naive top of S we have an interval labelled xw_1y, where y is a non-constant letter coloured μ_2 and w_1 contains w_0 and perhaps some constant letters from μ_1 and μ_2.

Our conditioning of ϕ (Proposition 4.5) ensures that, for all non-constant letters z, the rightmost non-constant letter in $\phi^j(z)$ is the same for all $j \geq 1$. Therefore, in order for there to ever be cancellation between non-constant letters coloured μ_1 and μ_2, we must have $x = y^{-1}$. Thus on $\top(S)$ there is an interval labelled xwx^{-1}, where w is the (non-empty) free-reduction of w_1.

At times greater than time(S), the future of the interval that we are considering will continue to have a core subarc labelled xw_jx^{-1}, where w_j is a conjugate of w by a (possibly-empty) word in constant letters (unless the interval hits a singularity or the boundary). In particular, no non-constant letters from μ_1 and μ_2 can ever cancel each other. □

In the light of the Bounded Cancellation Lemma we deduce:

Corollary 8.2. *Let $I \subset \bot(S)$ be an interval of edges labelled by constant letters, and suppose that the colours $\mu_1(S)$ and $\mu_2(S)$ lie either side of I, adjacent to it. If $|I| \geq B$ then there is never any cancellation between non-constant letters in μ_1 and μ_2.*

8.2. The Two Colour Lemma.

Definition 8.3. Suppose that U and V are positive words[8] and that for some $k > 0$ the only negative exponents occurring in $\phi^k(UV^{-1})$ are on constant letters. Then we say that U *ϕ-neuters V^{-1} in at most k steps.*

We shall also apply the term ϕ-neuters to describe the cancellation between colours $\mu(S), \mu'(S) \subseteq \bot(S)$ that are adjacent in corridors of van Kampen diagrams, and the following lemma remains valid in that context.

Proposition 8.4 (Two Colour Lemma)**.** *There exists a constant T_0 depending only on ϕ so that for all positive words U and V, if U ϕ-neuters V^{-1} then it does so in at most T_0 steps.*

Proof. We express V^{-1} as a product of three subwords: reading from the left of V^{-1}, the first subword ends with the last letter y such that $\phi(y)$ contains

[8] i.e. none of their letters are inverses a_j^{-1}

a left-fast letter; the second subword follows the first and ends with the last non-constant letter in V^{-1}; the remainder of V^{-1} consists entirely of constant letters.

Lemma 5.3 tells us that the length of the first subword is less than C_0, and the proof of Lemma 6.7 provides a bound of C_1 on the length of the second subword.

Now consider the freely reduced form of $\phi^k(UV^{-1})$, and let v_k denote its subword that begins with the first letter of negative exponent and ends with the final non-constant letter. The argument just applied to V^{-1} shows that v_k has length less than $C_0 + C_1$ for all $k \geq 0$.

Suppose that U ϕ-neuters V^{-1} in exactly N steps, let α_{N-1} be the letter of $\phi^{N-1}(UV^{-1})$ that consumes the last letter of v_{N-1}, and let α_k be the ancestor of α_{N-1} in $\phi^k(UV^{-1})$. Write $\phi^k(UV^{-1}) = w_k\alpha_k u_k v_k w_k'$.

Lemma 5.3 shows that $|u_k| < C_0$ for all $k < N$, and we have just argued that $|v_k| < C_0 + C_1$. Thus we obtain a bound (independent of U and V) on the number of words $\alpha_k u_k v_k$ that arise as k varies — call this number T_0. If N were greater than T_0, then some configuration $\alpha_k u_k v_k$ with v_k non-empty would recur. But this is nonsense, because once there is this repetition, the words v_k will continue to repeat, and thus V^{-1} will never be ϕ-neutered, contrary to assumption. □

Corollary 8.5. *There exists a constant $T_0{}'$, depending only on ϕ, with the following property: if U and V are positive words, V begins with a non-constant letter and $\phi^k(UV^{-1})$ is positive for some $k > 0$, then the least such k is less than $T_0{}'$.*

Proof. The preceding lemma provides an upper bound on the least integer N such that $\phi^N(UV^{-1})$ contains no non-constant letters with negative exponent. Up to this point, the rightmost non-constant letter in $\phi^k(UV^{-1})$ may have been spawning constant letters to its right, and thus $\phi^k(UV^{-1})$ may have a terminal segment consisting of constant letters. Since the rightmost non-constant letter of $\phi^k(V^{-1})$ does not vary with k when $k < N$ (by Proposition 4.5), the length of this segment grows at a constant rate ($< M$) during each application of ϕ. Similarly, its length changes at a constant rate after time N, decreasing until it is eventually cancelled.

Since $N \leq T_0$, this segment of constant letters has length less than MT_0 at time N, and hence is cancelled entirely before time $T_0(M + 1)$. □

8.3. **The disappearance of colours: Pincers and implosions.** In this subsection we turn our attention to the detailed study of how non-adjacent colours along a corridor in Δ can come together solely as a result of the mutual annihilation of the intervening colours. Such an event determines a *pincer* (Figure 14), which is defined as follows.

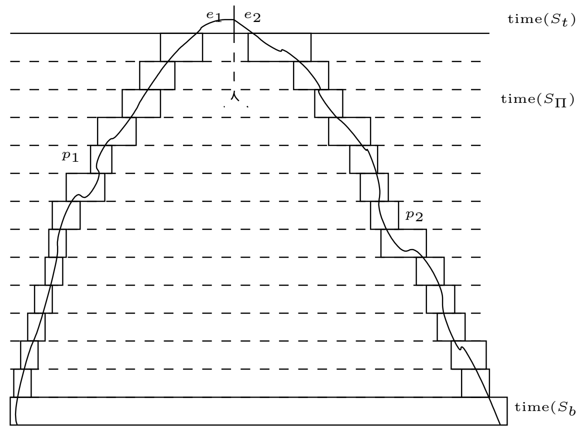

FIGURE 14. A pincer.

Definition 8.6. Consider a pair of paths p_1, p_2 in $\mathcal{F} \subseteq \Delta$ tracing the histories of 2 non-constant edges e_1, e_2 that cancel in a corridor S_t. Let μ_i denote the colour of the 2-cells along p_i. Suppose that at time τ_0 these paths lie in a common corridor S_b. Under these circumstances, we define the *pincer* $\Pi = \Pi(p_1, p_2, \tau_0)$ to be the subdiagram of Δ enclosed by the chains of 2-cells along p_1 and p_2, and the chain of 2-cells connecting them in S_b.

When it creates a desirable emphasis, we shall write $S_b(\Pi)$ and $S_t(\Pi)$ in place of S_b and S_t.

We define S_Π to be the earliest corridor of the pincer in which $\mu_1(S_\Pi)$ and $\mu_2(S_\Pi)$ are adjacent. We define $\tilde{\chi}(\Pi)$ to be the set of colours $\mu \notin \{\mu_1, \mu_2\}$ such that there is a 2-cell in Π coloured μ. And we define

$$\text{Life}(\Pi) = \text{time}(S_\Pi) - \text{time}(S_b).$$

Proposition 8.7 (Unnested Pincer Lemma)**.** *There exists a constant $\hat{T}_1$, depending only on ϕ, such that for any pincer Π*

$$\text{Life}(\Pi) \leq \hat{T}_1(1 + |\tilde{\chi}(\Pi)|).$$

Fix a pincer Π and assume $\text{Life}(\Pi) \neq 0$. The idea of the proof of Proposition 8.7 is as follows: we shall identify a constant $\hat{T}_1$ and argue that if none of the colours $\mu \in \tilde{\chi}(\Pi)$ were consumed entirely by $\text{time}(S_b) + \hat{T}_1$, the situation reached would be so stable that no colours could be consumed in Π at subsequent times, contradicting the fact that all but μ_1 and μ_2 must be consumed by $\text{time}(S_\Pi)$.

With this approach in mind, we make the following definition:

Definition 8.8. Let p be a positive integer. A *p-implosive array* of colours in a corridor S is an ordered tuple $A(S) = [\nu_0(S), \ldots, \nu_r(S)]$, with $r > 1$, such that:

(1) each pair of colours $\{\nu_j, \nu_{j+1}\}$ is *essentially adjacent* in S, meaning that there are no non-constant edges of any other colour separating $\nu_j(S)$ from $\nu_{j+1}(S)$;
(2) in each of the corridors $S = S^1, S^2, \dots, S^p$ in the future of S, every $\nu_j(S^i)$ contains a non-constant edge;
(3) in S^p, *either* a non-constant edge coloured ν_0 cancels a non-constant edge coloured ν_r (and hence the colours ν_j with $j = 1, \dots, r-1$ are consumed entirely), *or else* all of the non-constant letters in $\nu_j(S^p)$, for $j = 1, \dots, r-1$, are cancelled in S^p by edges from one of the colours of the array, while $\nu_0(S^p)$ and $\nu_r(S^p)$ contain non-constant letters that survive in the free-reduction of the naive future of the interval $\nu_0(S^p) \dots \nu_r(S^p) \subset \bot(S^p)$ (but may nevertheless be cancelled in S^p by edges from colours external to the array).

Arrays satisfying the first of the conditions in (3) are said to be of Type I, and those satisfying the second condition are said to be of Type II. (These types are not mutually exclusive.)

The *residual block* of an array of Type II is the interval of constant edges between the rightmost non-constant letter of ν_0 and the leftmost non-constant letter of ν_r in the free reduction of the naive future of $\nu_0(S^p) \dots \nu_r(S^p)$. The *enduring block* of the array is the set of constant edges in $\bot(S)$ that have a future in the residual block.

Note that there may exist *unnamed colours* between $\nu_j(S)$ and $\nu_{j+1}(S)$ consisting entirely of constant edges.

Remarks 8.9. Let $[\nu_0(S), \dots, \nu_r(S)]$ be a p-implosive array.

(1) Any implosive subarray of $[\nu_0(S), \dots, \nu_r(S)]$ is p-implosive (same p).

(2) If an edge of ν_i cancels with an edge of ν_j and $j - i > 1$, then this cancellation can only take place in S^p. If the edges cancelling are non-constant, then the subarray $[\nu_i(S), \dots, \nu_j(S)]$ is p-implosive of Type I.

(3) Given $x, y, w \in F$, if the freely reduced words representing x, y and $\phi(xwy)$ consist only of constant letters, then so does the reduced form of w, since the subgroup generated by the constant letters is invariant under $\phi^{\pm 1}$. It follows that the residual block of any array of Type II contains edges from at most two of the colours ν_j, and if there are two colours they must be essentially adjacent, i.e. $\nu_j(S^p), \nu_{j+1}(S^p)$.

(4) For the same reason, the enduring block of an implosive array of Type II is an interval involving at most two of the ν_j, and if there are two such colours then they must be essentially adjacent.

Lemma 8.10. *The ordered list of colours along each corridor before* $\mathrm{time}(S_\Pi)$ *in a pincer* Π *must contain an implosive array.*

Proof. At the top of the pincer there is cancellation between non-constant edges. Lemma 8.1 tells us that before $\mathrm{time}(S_\Pi)$ the colours of these edges must have been separated by a non-constant letter of a different colour, hence the list of non-constant colours along the bottom of S_Π is a 1-implosive array. This same list of colours defines an implosive array at each earlier time in the pincer until, going backwards in time, further non-constant colours appear. Suppose μ has non-constant letters in Π at time t but not time $t+1$. Let ν_0 be the first colour to the left of μ that contains non-constant letters at time $t+1$, and let ν_r be the first such colour to the right. If S_t is the corridor at time t, then the list of essentially-adjacent non-constant colours $[\nu_0(S_t), \ldots, \mu(S_t), \ldots, \nu_r(S_t)]$ is a 1-implosive array. And $[\nu_0(S_{t'}), \ldots, \mu(S_{t'}), \ldots, \nu_r(S_{t'})]$ is a $(t'-t+1)$-implosive array for each earlier time t' until (going backwards in time) either further non-constant colours appear or else we reach the bottom of the pincer. □

If, further to the above lemma, we can argue that there is a constant $\hat{T}_1$ such that each corridor before time(S_Π) contains a p-implosive array with $p \le \hat{T}_1$, then we will know that at least one of the colours from $\tilde{\chi}(\mathcal{P})$ is *essentially consumed* (i.e. comes to consist of constant edges only) during each interval of $\hat{T}_1$ units in time during the lifetime of the pincer. Thus Proposition 8.7 is an immediate consequence of the following result, which will be proved in (8.18).

Proposition 8.11 (Regular Implosions)**.** *There is a constant $\hat{T}_1$ depending only on ϕ such that every implosive array in any minimal area diagram Δ is p-implosive for some $p \le \hat{T}_1$.*

The first restriction to note concerning implosive arrays is this:

Lemma 8.12. *If $[\nu_0(S), \ldots, \nu_r(S)]$ is implosive of Type I, then $r \le B$. If it is implosive of Type II, then $r < 2B$.*

Proof. In Type I arrays, the interval $\nu_1(S^p) \ldots \nu_{r-1}(S^p) \subset \perp(S^p)$ is to die in S^p, so $r-1 < B$ by the Bounded Cancellation Lemma. For Type II arrays, one applies the same argument to the intervals joining $\nu_0(S^p)$ and $\nu_r(S^p)$ to the residual block of constant letters. □

Remark 8.13. In the light of Lemma 8.12, an obvious finiteness argument would provide the bound required for Lemma 8.11 if we were willing to restrict ourselves to implosive arrays with a uniform bound on their length. Motivated by this observation, we seek to prove that every implosive array contains an implosive sub-array that is uniformly *short*.

In order to identify a suitable notion of *short*, we need to consider a further decomposition of the colours $\nu_j(S_b)$ in a p-implosive array $[\nu_0(S_b), \ldots, \nu_r(S_b)]$.

Previously (Subsection 6.3) we partitioned each colour $\nu_j(S_b)$ into five intervals $A_1(S_b, \nu_j), \ldots, A_5(S_b, \nu_j)$ and then further decomposed A_4 into subintervals $C_{(\nu_j,\nu')}(1)$ and $C_{(\nu_j,\nu')}(2)$ according to the colours of the edges that were going to consume these subintervals in the future. There is a corresponding decomposition of A_2 into intervals which we denote $C^2_{(\nu_j,\nu')}(1)$ and $C^2_{(\nu_j,\nu')}(2)$ (where ν' is now to the left of ν_j in S_b).

Adapting to our new focus, we now define $R_j(S_b) = A_5(\nu_1, S_b) \cup C_{(\nu_j,\nu_{j+1})}(1)$, and $L_j(S_b) = A_1(\nu_1, S_b) \cup C^2_{(\nu_j,\nu_{j+1})}(1)$. We also define $C^R_j(S_b)$ to be $C_{(\nu_j,\nu_{j-1})}(2)$ minus any edges from the excluded block, and $C^L_j(S_b)$ to be $C^2_{(\nu_j,\nu_{j-1})}(2)$ minus any edges from the excluded block. Thus we obtain a decomposition of $\nu_j(S_b)$ into five intervals (see Figure 15)

$$L_j(S_b),\ C^L_j(S_b),\ \text{Mess}(S_b, \nu_j),\ C^R_j(S_b),\ R_j(S_b)$$

where $\text{Mess}(S_b, \nu_j)$ contains the edges whose preferred future dies at the time of implosion together with edges from the excluded block[9].

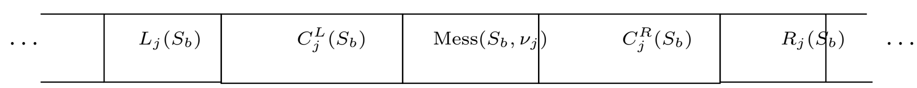

FIGURE 15. The decomposition of the colour ν_j

The terminal colours in our array, ν_0 and ν_r, play a special role. This is reflected in the fact that we shall only need to consider the segment of ν_0 from its right end up to and including the edge one to the left of $\text{Mess}(S_b, \nu_0)$. And in ν_r we shall only need to consider the segment from its left end up to and including the edge one to the right of $\text{Mess}(S_b, \nu_r)$. We write $\mathcal{L}(\nu_j, S_b)$ and $\mathcal{R}(\nu_j, S_b)$, respectively, to denote these sub-intervals of $\nu_j(S_b)$.

Definition 8.14. The length of $A(S) = [\nu_0(S), \dots, \nu_r(S)]$, written $\| A(S) \|$, is the number of edges in the interval $\mathcal{L}(\nu_0, S) \dots \mathcal{R}(\nu_r, S) \subset \bot(S)$. (Note that $\| A(S) \|$ takes account of the unnamed colours.)

In keeping with the notation in the definition of p-implosive, we shall write S^t for the corridor t steps into the future of S_b; in particular $S^0 = S_b$ and each ν_j with $j = 1, \dots, r-1$ essentially vanishes in S^p.

By definition, no preferred future of any edge in $\text{Mess}(\nu_j, S_b)$ is cancelled before S^p. Hence these intervals do not shrink in length before that time, and as in the proof of Lemma 8.12 we can use the Bounded Cancellation Lemma to bound the sum of their lengths:

Lemma 8.15. *After excluding the edges of the enduring block, the sum of the lengths of the intervals* $\text{Mess}(\nu_j, S_b)$ *is at most* $2B$.

Combining this estimate with the bounds from Lemmas 5.3 and 6.7, we deduce that for $j = 1, \dots, r-1$

[9] At this point the reader may find it helpful to recall that only arrays of Type II have excluded blocks, and such a block is either contained in a single colour, or in adjacent colours $\nu_j(S_b) \cup \nu_{j+1}(S_b)$ with the intervening intervals $R_j(S_b) \dots L_{j+1}(S_b)$ empty.

$$|\nu_j(S_b)| \leq |C_j^L(S_b)| + |C_j^R(S_b)| + 2C_0 + 2C_1 + 2B + \mathcal{E}_j,$$

where $\mathcal{E}_j$ is the number of edges from the excluded block coloured ν_j.

Similarly,

$$|\mathcal{L}(\nu_0, S_b)| \leq |C_0^R(S_b)| + C_0 + C_1 + B + \mathcal{E}_0$$

and

$$|\mathcal{R}(\nu_r, S_b)| \leq |C_r^L(S_b)| + C_0 + C_1 + B + \mathcal{E}_r.$$

This motivates us to define an array of colours $[\nu_0(S), \ldots, \nu_r(S)]$ to be *very short* if for $j = 1, \ldots, r-1$ we have

$$|\nu_j(S)| \leq 2C_0 + 2C_1 + 5B + 1,$$

and

$$|\mathcal{L}(\nu_0, S)| \leq C_0 + C_1 + 5B + 1,$$

and

$$|\mathcal{R}(\nu_r, S)| \leq C_0 + C_1 + 5B + 1,$$

and for $j = 0, \ldots, r-1$ the interval formed by the unnamed colours between $\nu_j(S)$ and $\nu_{j+1}(S)$ has total length at most B.

An implosive array is said to be *short* if it satisfies the weaker inequalities obtained by increasing each of these bounds by $2BT_0$.

Lemma 8.16. *Let $A = [\nu_0(S^0), \ldots, \nu_r(S^0)]$ be a p-implosive array with $p \geq T_0$.*

(1) If $[\nu_0(S^{T_0}), \ldots, \nu_r(S^{T_0})]$ is very short, then A is short.

(2) If A is short, then $\| A \| \leq 2B(2C_0+2C_1+5B+1+2BT_0)+2B^2(1+2T_0)$.

Proof. Item (1) is an immediate consequence of the Bounded Cancellation Lemma 2.4. The (crude) bound in (2) is an immediate consequence of Lemma 8.15 and the inequalities in the definition of *short*; the first summand is an estimate on the sum of the lengths of the named colours, and the second summand accounts for the unnamed colours. □

The following lemma is the key step in the proof of Proposition 8.7.

Lemma 8.17. *If $A(S^0) = [\nu_0(S^0), \ldots, \nu_r(S^0)]$ is a p-implosive array, then at least one of the following statements is true:*

(1) $p \leq 2T_0$;

(2) $A(S^0)$ is short;

(3) $p > 2T_0$ and $A(S^{T_0})$ contains an implosive sub-array $[\nu_k(S^{T_0}), \ldots, \nu_l(S^{T_0})]$ that is very short.

Proof. Assume $p > 2T_0$ and that $[\nu_0(S^0), \ldots, \nu_r(S^0)]$ is not short. We claim that there is a block of at least $B+1$ constant letters in the interval determined by the array $\mathcal{L}(\nu_0, S^{T_0}) \ldots \mathcal{L}(\nu_r, S^{T_0})$. Indeed, by definition, if an array is not short then either one of the $\mathcal{E}_j$ has length at least $B+1$, or one of the blocks of unnamed colours has length at least $B(2T_0+1)+1$, or else at least one of the intervals of constant letters $C_j^L(S^0)$ or $C_j^R(S^0)$ has length at least $B(T_0+1)+1$. In the first case, since $\mathcal{E}_j$ is in the excluded block, none of its edges are cancelled before the moment of implosion, and hence it contributes a block of at least $B+1$ constant letters to $A(S^{T_0})$; in the second case, the Bounded Cancellation Lemma assures us that the length of the appropriate block of unnamed colours can decrease by at most $2B$ at each step before the implosion of the array, and hence it still contributes a block of at least $B+1$ constant edges to $A(S^{T_0})$; and similarly, in the third case, $C_j^*(S^0)$ can decrease by at most B at each step before the implosion of the array.

Let β be a block of at least $B+1$ constant edges in $A(S^{T_0})$ with non-constant edges e_l and e_ρ immediately to its left and right, respectively. The Buffer Lemma 8.1 assures us that the non-constant edges in the future of e_l will never interact with the non-constant edges in the future of e_ρ. Thus at least one of e_l or e_ρ must be *stabbed in the back*, i.e. its entire non-constant future must be consumed by edges on its own side of β. Suppose, for ease of notation, that it is e_l and let ν_i be the colour of e_l. We claim that if ν_k is the colour of the letter that ultimately consumes e_l, then $k \le i-2$.

We shall derive a contradiction from the assumption that the edge which ultimately consumes e_l is coloured ν_{i-1}. There are two cases to consider according to whether e_ρ is also coloured ν_i. If it is, then we consider the word V labelling the arc of $\perp(S^0)$ from the left end of $\nu_i(S^0)$ to the past of e_l; the consumption of the non-constant future of e_l completes the ϕ-neutering of V by the word labelling $\nu_{i-1}(S^0)$, in particular this neutering will have taken more than T_0 steps in time, contradicting the Two Colour Lemma 8.4. If e_ρ is not coloured ν_i, then the consumption of the non-constant future of e_l results in a new essential adjacency of colours and hence can only be complete at the moment of implosion, i.e. time(S^p). But this consumption constitutes the neutering of $\nu_i(S^{T_0})$ by $\nu_{i-1}(S^{T_0})$, and according to the Two Colour Lemma this neutering must be accomplished in at most T_0 units of time. Thus $p \le 2T_0$, contrary to our hypothesis.

Thus we have proved that the edge which ultimately consumes e_l is coloured ν_k where $k \le i-2$. Under these circumstances (or the symmetric situation with e_ρ in place of e_l) we say that ν_k *neuters* ν_i *from behind* and write $\nu_k \searrow \nu_i$.

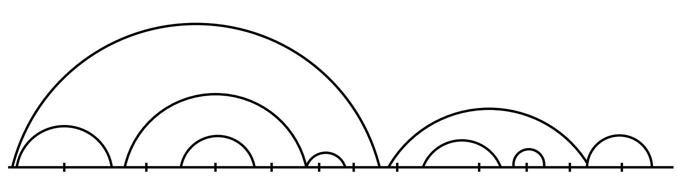

FIGURE 16. The nesting associated to $\searrow$

There is a natural *nesting* among the $\searrow$-related pairs of colours from the array: $(\nu_{k_1}, \nu_{j_1}) < (\nu_{k_2}, \nu_{j_2})$ if ν_{k_1} and ν_{j_1} both lie between ν_{k_2} and ν_{j_2} in S^0. See Figure 16.

We focus our attention on an innermost (i.e. minimal) pair with $\nu_k \searrow \nu_i$. By definition $|k-i| \geq 2$. If there were a block of at least $B+1$ constant letters between the closest non-constant letters of $\nu_k(S^{T_0})$ and $\nu_i(S^{T_0})$, then the preceding argument would yield a neutering from behind that contradicted the innermost nature of $\nu_k \searrow \nu_i$. Thus $[\nu_j(S^{T_0}), \ldots, \nu_k(S^{T_0})]$ is a very short array, and we are done. $\square$

8.18. *Proof of Regular Implosions (Prop.8.11):* Given the bound in Lemma 8.16(2), an obvious finiteness argument provides a constant τ such that every short implosive array is p-implosive with $p \leq \tau$. And the same bound applies to implosive arrays that contain a short sub-array (Remark 8.9(1)). So in the light of Lemmas 8.17 and 8.16(1), it suffices to let $\hat{T}_1 = \max\{2T_0, \tau\}$. $\square$

8.4. **Super-Buffers.** In this subsection we prove an important cancellation lemma based on Proposition 8.7, this lemma involves the following constant.

Definition 8.19. We fix an integer T_1' such that one gets repetitions in all T_1'-long subsequences of 5-tuples of reduced words

$$U_k := \Big(u_{k,1}, u_{k,2}, u_{k,3}, u_{k,4}, u_{k,5}\Big) \qquad k = 1, 2, \ldots$$

with $|u_{k,1}|$ and $|u_{k,1}|$ at most $C_0 + C_1 + 2B + 1$, while $|u_2^k|$ and $|u_4^k|$ are at most $C_0 + C_1$, and $|u_3^k| \leq 4B + 1$. That is, for some $t_1 \leq t_2 \leq T_1'$ and

$$\Big(u_{t_1,1}, u_{t_1,2}, u_{t_1,3}, u_{t_1,4}, u_{t_1,5}\Big) = \Big(u_{t_2,1}, u_{t_2,2}, u_{t_2,3}, u_{t_2,4}, u_{t_2,5}\Big).$$

Stipulation 8.20. Assume $T_1' \geq \hat{T}_1$.

The cancellation lemma we need is most easily phrased in terms of colours of subwords, which we define as follows, keeping firmly in mind the example of a stack of partial corridors excised from the interior of a van Kampen diagram, retaining their memory of the colours to which the edges belong.

We have a word W with a decomposition into preferred subwords $V = V_1 V_2 \cdots V_k$, where each V_i is either positive or negative; we think of these subwords as having colours $\mu_1, \ldots \mu_k$. Take the freely reduced words $\phi(V_i)$, concatenate them, then cancel to form a freely reduced word. There is some freedom in the choice of cancellation scheme, as in the folding of corridors, but we fix a choice, thus assigning to each letter of the freely reduced form of $\phi(V)$ the colour μ_i of its ancestor. We repeat this process, thus assigning colours to the letters in the reduced form of $\phi^k(V)$ for each integer $k > 0$.

The process that we have just described is an algebraic description of a choice of minimal area van Kampen diagram for $t^{-k} V t^k \phi^k(V)^{-1}$. Thus the following lemma is a comment on the form of such diagrams.

Proposition 8.21. *Let $V = V_1V_2V_3$ be a concatentation of words (coloured ν_1, ν_2, ν_3) each of which is either positive or negative. If W is a subword of the reduced form of $\phi^{T_1'}(V)$ and W has a non-constant letter coloured ν_i for each $i \in \{1,2,3\}$, then for all $k \geq 0$ there are non-constant letters in $\phi^k(W)$ coloured ν_2.*

Proof. Let $\nu_i(W)$ denote the subword of W coloured ν_i, and let ν_i^j denote the maximal subword coloured ν_i in (the reduced word representing) $\phi^i(V_1V_2V_3)$. Note that $\nu_2(W) = \nu_2^{T_1'}$, and more generally $\nu_2^{T_1'+j}$ is the maximal word in $\phi^j(W)$ coloured ν_2.

Fix $k > T_1'$ and consider the diagram formed by the stack of corridors described prior to the proposition. The bottom of the first corridor is labelled V, and we regard it as being divided into three coloured intervals according to the decomposition $V_1V_2V_3$. Since $\nu_2(W)$ contains non-constant letters and $T_1' > \hat{T}_1$, the array formed by these colours is not implosive (Proposition 8.7), and hence $\nu_1(W)$ and $\nu_3(W)$ will never essentially consume $\nu_2(W)$. However, the proposition is not yet proved because there remains the possibility that ν_2 may essentially vanish because it neuters $\nu_1(W)$, say, and is then neutered by $\nu_3(W)$. We proceed under this assumption, seeking a contradiction. (The case where the roles of ν_1 and ν_3 are reversed is entirely similar.)

For each $1 \leq i \leq T_1'$, we have $\phi^i(V_1V_2V_3) = \nu_1^i, \nu_2^i$ and ν_3^i. Write $\nu_2^i \equiv V^i(1)V^i(2)V^i(3)$, where $V^i(1)$ ends with last non-constant letter in ν_2^i whose entire non-constant future is eventually consumed by letters coloured ν_1, and $V^i(3)$ begins with the leftmost non-constant letter whose entire non-constant future is eventually consumed by letters coloured ν_3. Lemmas 5.3 and 6.7 tell us that $V^i(1)$ and $V^i(3)$ have length at most $C_0 + C_1$.
Claim: $V^i(2)$ contains exactly one non-constant edge and has length no more than $4B+1$.

We are assuming that $\nu_2(W)$ neuters $\nu_1(W)$. Consider the (non-constant) edge ε_i in ν_2^i that will eventually consume the final non-constant edge in $\nu_1(W)$. Note that ε_i is the leftmost non-constant edge in $V^i(2)$. Moreover, we are assuming that $\nu_3(W)$ ultimately neuters $\nu_2(W)$, so in particular it consumes the entire future of any edge to the right of ε_i, which forces ε_i to be the rightmost non-constant edge in $V^i(2)$. The Buffer Lemma tells us that ε_i must lie within $2B$ of both ends of $V^i(2)$, and hence the claim is proved.

Looking to the left of $V^i(1)$, we now consider the subword L^i of ν_1^i that begins with the leftmost non-constant edge in the future of which there is a non-constant letter that cancels with a letter coloured ν_2. And looking to the right of $V^i(3)$, we consider the subword that ends with the rightmost non-constant letter in the future of which there is a non-constant letter that cancels with a letter coloured ν_2. any of whose non-constant future cancels with an edge painted ν_2. As in previous arguments, The Buffer Lemma and Lemmas 5.3, 6.7 tell is that $|R^i|, |L^i| \leq C_0 + C_1 + 2B + 1$, for all i.

We have already bounded the lengths of $V^i(1), V^i(2)$ and $V^i(3)$ by $C_0 + C_1, 4B+1$ and C_0+C_1, respectively. Thus we are now in a position to invoke the repetitive behaviour described in Definition 8.19: for some positive integers i and t with $i+t \le T_1'$, we get a repetition

$$\Big(R^i, V^i(1), V^i(2), V^i(3), L^i\Big) = \Big(R^{i+t}, V^{i+t}(1), V^{i+t}(2), V^{i+t}(3), L^{i+t}\Big).$$

For as long as we are assured of the continuing presence of ν_1^{i+s} and ν_3^{i+s}, the fate of $\nu_2^i = V^i(1)V^i(2)V^i(3)$ under s iterations of ϕ depends only on $(R^i, V^i(1), V^i(2), V^i(3), L^i)$. Thus

$$\Big(V^j(1), V^j(2), V^j(3)\Big) = \Big(V^{j+t}(1), V^{j+t}(2), V^{j+t}(3)\Big)$$

for all $j \ge i$ within the time scale of this assurance. However this leads us to an absurd conclusion, because once ν_1 has become constant, at all subsequent time, the surviving word coloured ν_2 contains as a proper subword, the ν_2 word that existed at the corresponding times in the cycles (of period t) before T_1', and in particular they can never essentially vanish, contrary to our assumption that ν_3 eventually neuters ν_2. □

8.5. **Nesting and the Pincer Lemma.** In subsequent sections we would like to bound the life of pincers by arguing that during the lifetime of a pincer, colours must be consumed at a predictable rate (appealing to Proposition 8.7), noting that there are only a limited number of colours. However, the bounds we need will require us to ascribe each consumed colour to a *unique* pincer. Thus we encounter problems whenever one pincer is contained in another. For reasons that will become apparent in subsequent sections, in situations where we must confront this problem, the inner of the two pincers will have a long block of constant edges along the corridor immediately above its peak. More precisely, we will find ourselves in the situation described in the following definition. The appearance of the constant $\lambda_0 := 2B(T_0+1)+1$ in the following definition is explained by the role that this constant played in the course of Lemma 8.17.

Definition 8.22. Consider one pincer Π_1 contained in another Π_0. Suppose that in the corridor $S \subseteq \Pi_0$ at the top of Π_1 (where its boundary paths $p_1(\Pi_1)$ and $p_2(\Pi_1)$ come together) the future in $\top(S)$ of at least one of the edges containing $p_1(\Pi_1) \cap \bot(S)$ or $p_2(\Pi_1) \cap \bot(S)$ contains no non-constant edges, and this future[10] lies in an interval of at least λ_0 constant edges contained in Π_0. Then we say that Π_1 is *nested in* Π_0. (in Figure 17, the λ_0-long block

[10]We allow this future to be empty, in which case "contained in" means that the immediate past of the long block of constant edges is not separated from Π_1 by any edge that has a future in $\top(S)$.

of constant edges are shown in black.) We say that Π_1 is *left-loaded* or *right-loaded* according to the direction in which the λ_0-long block of constant edges extends from the peak of Π_1.

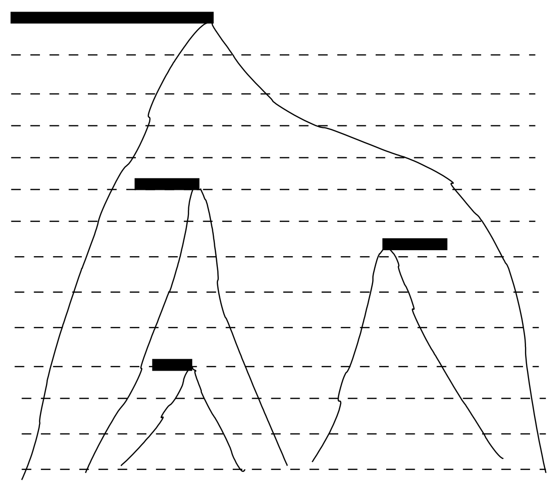

FIGURE 17. A depiction of nesting

Remark 8.23. A nested pincer cannot be both left-loaded and right-loaded (cf. Remark 8.9(3)).

If Π_1 is left-loaded, then the future of $p_1(\Pi_1)\cap\bot(S)$ contains no non-constant edges. It may happen that the future of $p_2(\Pi_1)$ also contains no non-constant edges; in this case the colour μ of $p_2(\Pi_1)$ essentially vanishes in S due to cancellation between non-constant edges of μ and some colour to its right. Symmetric considerations apply to right-loaded pincers.

Definition 8.24. For a pincer Π_0, let $\{\Pi_i\}_{i\in I}$ be the set of all pincers nested in Π_0. Then define

$$\chi(\Pi_0) = \tilde{\chi}(\Pi_0) \smallsetminus \bigcup_{i\in I} \tilde{\chi}(\Pi_i).$$

Lemma 8.25. *If the pincer* Π_1 *is nested in* Π_0 *then* $\mathrm{time}(S_t(\Pi_1)) < \mathrm{time}(S_{\Pi_0})$.

Proof. The presence of the hypothesised block of constant letters in $\top(S_t(\Pi_1))$ makes this an immediate consequence of the Buffer Lemma 8.1. □

Define $T_1 := T_1' + 2T_0$. The following theorem is the main result of this section.

Theorem 8.26 (Pincer Lemma)**.** *For any pincer* Π

$$\mathrm{Life}(\Pi) \leq T_1(1 + |\chi(\Pi)|).$$

Proof. The heart of our proof of Proposition 8.7 was that in each block of $\hat{T}_1$ steps in time between time(S_b) and time(S_Π) at least one colour essentially disappears. Our proof of the present theorem is an elaboration of that argument: we must argue for the essential disappearance of a colour that is not contained in any of pincers nested in Π. Thus we concentrate on that region of the pincer Π that is exterior to the set of *co-level*[11] *1* pincers nested in it; let $\{\Pi_j\}$, $j = 1, \dots, J$ be the set of such, indexed in order of appearance from left to right.

For $j = 1, \dots, J-1$, let Σ_j denote the set of colours along the bottom of Π that have a non-constant edge strictly between Π_j and Π_{j+1}; if Π_j is left-loaded, then we include the colour of $p_2(\Pi_j)$ in Σ_j, and if Π_j is right-loaded, then we include the colour of $p_1(\Pi_j)$ in Σ_{j-1}. Likewise, we define Σ_0 to be the set of non-constant colours that lie to the left of Π_1 together with the colour of $p_1(\Pi)$, and we define Σ_J to be the set of non-constant colours that lie to the right of Π_J together with the colour of $p_2(\Pi)$.

In order to prove the theorem, we derive a contradiction from the assumption that in the first T_1 units of time in the life of Π no colours in the union of the Σ_j essentially vanish. (There is no loss of generality in starting at the bottom of the pincer, since given any other starting time, one can discard the pincer below that level.) We label the corridors, beginning at the bottom of Π and proceeding in time as $S^0, S^1, \dots$

We focus on a single Σ_j, and write its colours in order as $\nu_1, \dots, \nu_r$. We analyse how the colours in Σ_j come to vanish. The first important observation is that $2 \le i \le r-1$, it is not possible for the colour ν_i to essentially vanish (at any time) due to cancellation merely between the colours in Σ_j. For if this happened, there would be an implosive array in S^0 containing $\nu_i(S^0)$ and so, by Proposition 8.7, ν_i would vanish before S^{T_1}, contrary to our assumption.

There remains the possibility that ν_2 may neuter ν_1 (after S^{T_1}). This can happen in two ways. The first is that Π_{j-1} is left-loaded: in this case the neutering happens within time T_0 of the top of Π_{j-1} (by Two Colour Lemma), and we are then in a stable situation in the sense that ν_3 cannot subsequently neuter ν_2, by Proposition 8.21. Now suppose that Π_{j-1} is right-loaded. Consider the earliest time t_0 at which there is a block of at least $B+1$ constant edges in the past of the λ_0-long block associated to Π_{j-1}. If ν_2 is to neuter ν_1, then it must do so within T_0 steps of this time. Indeed, within T_0 steps, if the non-constant edges of ν_1 to the right of the block have not been consumed by ν_2, then they will never be consumed by a colour from Σ_j.

There is a further event that we must account for, which is closely related to neutering: it may happen that ν_1 is the colour of $p_2(\Pi_{j-1})$ and that ν_2 consumes all of the non-constant edges to the right of the block of constant edges discussed above; this is not a neutering but nevertheless the Two Colour

[11] i.e. those that are maximal with respect to inclusion among the pincers nested in Π

Lemma applies. We would like to apply Proposition 8.21 in this situation to conclude that ν_3 cannot subsequently neuter ν_2. This is legitimate provided $t_0 \geq \text{time}(S^{T_1'})$. If $t_0 < \text{time}(S^{T_1'})$, then we still know that ν_3 cannot neuter ν_2 before S^{T_1}, because by hypothesis no colour from Σ_j essentially vanishes before this time. On the other hand, the Two Colour Lemma tells us that if ν_3 is to neuter ν_2, then it must do so within T_0 steps from t_0, and $t_0 + T_0 \leq \text{time}(S^{T_1})$. Thus, once again, we conclude that ν_3 can never neuter ν_2.

Entirely similar arguments show that it cannot happen that ν_r is neutered by ν_{r-1} and that subsequently ν_{r-2} neuters ν_{r-1}.

We have established the existence of a stable situation: proceeding past the point where the restricted amount of possible neutering within Σ_j has occurred, we may assume that the next essential disappearance of a colour from Σ_j can only occur as a result of cancellation with a colour from some Σ_i with $i \neq j$. Such further cancellation must occur, of course, because all but two[12] of the colours in $\bigcup_j \Sigma_j$ must be consumed within Π.

Passing to innermost pair of interacting Σ_k we may assume $i = j - 1$ (cf. proof of Lemma 8.17). Thus our proof will be complete if we can argue that cancellation between non-constant edges from Σ_{j-1} and Σ_j is impossible. We have argued that the colours which are to cancel will be essentially adjacent within time T_0 of the top of Π_{j-1}. On the other hand, there is a block of λ_0 constant edges separating Σ_{j-1}-nonconstant edges and Σ_j-nonconstant edges at the top of Π_{j-1}. Since $\lambda_0 > 2B(T_0 + 1)$ at least $B + 1$ of these constant edges remain T_0 steps later. The Buffer Lemma now obstructs the supposed cancellation between non-constant edges in Σ_{j-1} and Σ_j. $\square$

9. Teams and their Associates

We begin the process of grouping pairs of colours (μ, μ') into teams.

9.1. **Pre-teams.** The whole of $C_{(\mu,\mu')}(2)$ will ultimately be consumed by a single edge $\varepsilon_0 \in \mu'(S_0)$. We consider the time t_0 at which the future of ε_0 starts consuming the future of $C_{(\mu,\mu')}(2)$. If $|C_{(\mu,\mu')}(2)| > 2B$, then this consumption will not be completed in three steps of time (Lemma 2.4). We claim that in this circumstance, the leftmost μ'-coloured edge after the first two steps of the cancellation must be left para-linear. Indeed it is not left-constant since it must consume edges in the future of $C_{(\mu,\mu')}(2)$, and since no non-constant μ'-edges are cancelled by μ in passing from the first to the second stage of cancellation, the leftmost non-constant μ'-label must remain the same (Proposition 4.5). We denote this left para-linear edge at time $t_0 + 2$ by ε^μ.

Let ε_μ be the rightmost edge in the future of $C_{(\mu,\mu')}(2)$ at time t_0. We trace the ancestry of ε_μ and ε^μ in the trees of $\mathcal{F} \subset \Delta$ corresponding to the colours

[12]Degenerate cases with few colours are covered by the Two Colour Lemma and the Buffer Lemma.

μ and μ' (as defined in 3.2). We go back to the last point in time $\hat{t}_1(\mu,\mu')$ at which both ancestors lay in a common corridor *and* the interval on the bottom of this corridor between the pasts of ε_μ and ε^μ is comprised entirely of constant edges whose future is eventually consumed by the ancestor of ε^μ at this time. We denote this corridor $S_\uparrow$.

Definition 9.1. The ancestor of ε^μ at time $\hat{t}_1(\mu,\mu')$ is called the *reaper* and is denoted $\hat{\rho}(\mu,\mu')$. The set of edges in $\perp(S_\uparrow)$ which are eventually consumed by $\hat{\rho}(\mu,\mu')$ is denoted $\hat{\mathfrak{T}}(\mu,\mu')$. This is a contiguous set of edges. The *pre-team* $\hat{\mathcal{T}}(\mu,\mu')$ is defined to be the set of pairs (μ_1,μ') such that $\hat{\mathfrak{T}}(\mu,\mu')$ contains edges coloured μ_1. The number of edges in $\hat{\mathfrak{T}}(\mu,\mu')$ is denoted $\parallel \hat{\mathcal{T}} \parallel$.

In a little while we shall define *teams* to be pre-teams satisfying a certain maximality condition (see Definition 9.6).

Remark 9.2. If $\hat{t}_1(\mu,\mu') < \text{time}(S_0)$ then near the right-hand end of $\hat{\mathfrak{T}}(\mu,\mu')$ one may have an interval of colours ν such that $\nu(S_0)$ is empty.

In the proof of Proposition 7.1 we saw that it would be desirable if (whatever our final definition of *team* and bonus may be) the following inequality (7.1) should hold for all teams:

$$\sum_{(\mu,\mu')\in\mathcal{T} \text{ or } (\mu,\mu')\in_v\mathcal{T}} |C_{(\mu,\mu')}(2)| \leq \parallel \mathcal{T} \parallel + |\text{bonus}(\mathcal{T})| + B. \tag{9.1}$$

The following lemma shows that, even without introducing a bonus scheme or virtual members, the desired inequality is straightforward for pre-teams with $\hat{t}_1(\mu,\mu') \geq \text{time}(S_0)$.

Lemma 9.3. *If* $\hat{t}_1(\mu,\mu') \geq \text{time}(S_0)$ *then* $\hat{\mathcal{T}}(\mu,\mu')$ *satisfies*

$$\sum_{(\mu,\mu')\in\hat{\mathcal{T}}(\mu,\mu')} |C_{(\mu,\mu')}(2)| \leq \parallel \hat{\mathcal{T}}(\mu,\mu') \parallel + B.$$

Proof. By definition $\mu'(S_0)$ does not start consuming any of the $C_{(\mu_1,\mu')}(2)$ with $(\mu_1,\mu') \in \hat{\mathcal{T}}$ before $\hat{t}_1(\mu,\mu')$ (apart from a possible nibbling of length $< B$ from the rightmost team member at time $\hat{t}_1(\mu,\mu') - 1$). Since each $C_{(\mu_1,\mu')}(2)$ consists only of edges consumed by $\mu'(S_0)$, the future of each $C_{(\mu_1,\mu')}(2)$ at time $\hat{t}_1(\mu,\mu')$ will have the same length as $C_{(\mu,\mu')}(2)$ (except that the rightmost may have lost these $< B$ edges). And these futures are contained in $\hat{\mathfrak{T}}(\mu,\mu')$. $\square$

The case where $\hat{t}_1(\mu,\mu') < \text{time}(S_0)$ is more troublesome. As $\hat{\mathfrak{T}}(\mu,\mu')$ flows forwards in time, the number of constant letters in the future of $\hat{\mathfrak{T}}(\mu,\mu')$ that are consumed by $\hat{\rho}(\mu,\mu')$ between $\hat{t}_1(\mu,\mu')$ and time(S_0) may be outweighed by the number of constant letters generated to the left of the future of $\hat{\mathfrak{T}}(\mu,\mu')$ that will ultimately be consumed by $\hat{\rho}(\mu,\mu')$.

It is to circumvent the failure of inequality (9.1) in this setting that we are obliged to instigate the bonus scheme described in Section 10.

9.2. **The Genesis of pre-teams.** We fix $\hat{\mathcal{T}}(\mu, \mu')$ with $\hat{t}_1(\mu, \mu') < \text{time}(S_0)$ and consider the various events that occur at $\hat{t}_1(\mu, \mu')$ to prevent us pushing the pre-team back one step in time. We write S_ω to denote the corridor at time $\hat{t}_1(\mu, \mu')$ containing $\hat{\mathfrak{T}}(\mu, \mu')$.

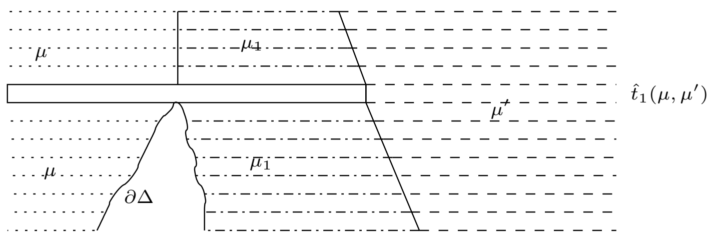

FIGURE 18. A team of genesis (G1)

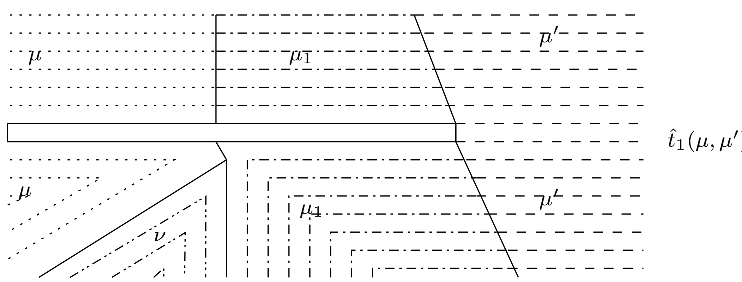

FIGURE 19. A team of genesis (G2)

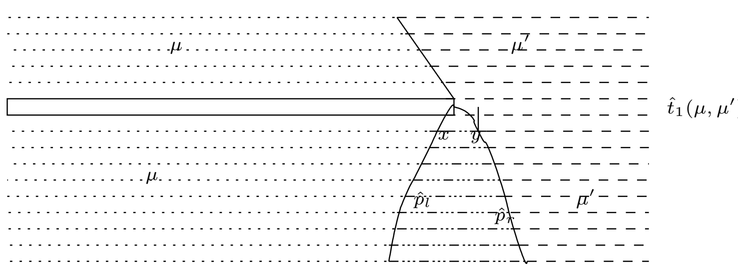

FIGURE 20. A team of genesis (G3)

There are four types of events:

(G1) The immediate past of $C_{(\mu,\mu')}(S_\omega)$ is separated from the past of $\hat{\rho}(\mu, \mu')$ by an intrusion of $\partial\Delta$ (Figure 18).

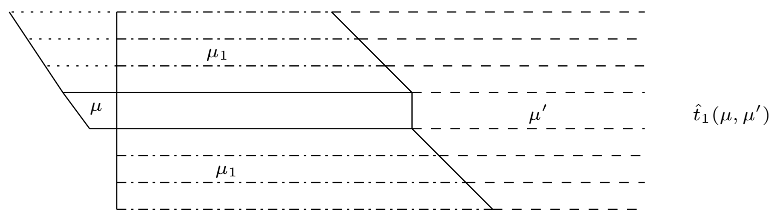

FIGURE 21. A team of genesis (G4)

(G2) We are not in case (G1), but the immediate past of $C_{(\mu,\mu')}(S_\omega)$ is separated from the past of $\hat{\rho}(\mu, \mu')$ because of a singularity (Figure 19).

(G3) The immediate past of $C_{(\mu,\mu')}(S_\omega)$ is still in the same corridor as the past of $\hat{\rho}(\mu, \mu')$, but it is separated from it by a non-constant letter (Figure 20).

(G4) We are not in any of the above cases, but the immediate past of the rightmost letter in $C_{(\mu,\mu')}(S_\omega)$ is not constant (Figure 21).

The following lemma explains why Figures 20 and 21 are an accurate portrayal of cases (G3) and (G4).

Let M_{inv} be the maximum length of $\phi^{-1}(x)$ over generators x of F, and $C_4 = M_{inv}.M$.

Lemma 9.4. *If I is an interval on $\top(S)$ labelled by a word w in constant letters then the reduced word labelling the past of I in $\bot(S)$ is of the form $u\alpha v$, where α is a word in constant letters and $|u|$ and $|v|$ are less than C_4. Moreover, if the past of the leftmost (resp. rightmost) letter in w is constant, then u (resp. v) is empty.*

In particular, $|I| \leq |\alpha| + 2MC_4$.

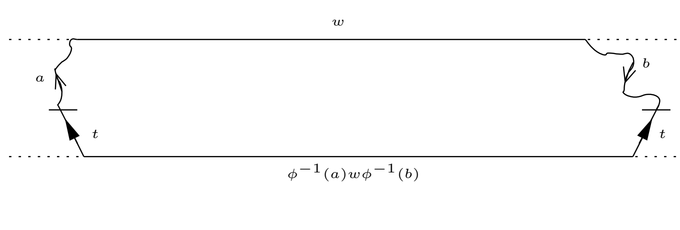

FIGURE 22. The proof of Lemma 9.4

Proof. See Figure 22. Follow the path from the left end of I to $\bot(S)$. This passes through a (possibly empty) path a^{-1}, followed by an edge labelled t^{-1}, where the length of a is less than M (since it can be chosen to be on the top of a 2-cell which has an edge in I). Similarly, at the right end of I we have a path labelled bt^{-1}, where the length of b is less than M. The path along $\bot(S)$

joining the two endpoints of these paths is labelled by the reduced word freely equal in F to $\phi^{-1}(awb) = \phi^{-1}(a)w\phi^{-1}(b)$. The only non-constant edges in this word come from $\phi^{-1}(a)$ and $\phi^{-1}(b)$, which have lengths at most $M.M_{inv}$. This proves the assertion in the first sentence.

The assertion in the second sentence follows from the observation that if x, y and $\phi(x\beta y)$ consist only of constant letters, then so does the reduced form of β, and the assertion in the final sentence follows immediately from the first. $\square$

Remark 9.5. It is convenient to assume that $MC_4 < \lambda_0$. (In the unlikely event that this is not the case, we simply increase λ_0.)

We are finally in a position to make an appropriate definition of a team.

Definition 9.6. All pre-teams $\hat{\mathcal{T}}(\mu, \mu')$ with $\hat{t}_1(\mu_1, \mu') \geq \text{time}(S_0)$ are defined to be teams, but the qualification criteria for pre-teams with $\hat{t}_1(\mu_1, \mu') < \text{time}(S_0)$ are more selective.

If the genesis of $\hat{\mathcal{T}}(\mu, \mu')$ is of type (G1) or (G2), then the rightmost component of the pre-team may form a pre-team at times before $\hat{t}_1(\mu, \mu')$. In particular, it may happen that $(\mu_1, \mu') \in \hat{\mathcal{T}}(\mu, \mu')$ but $\hat{t}_1(\mu, \mu') > \hat{t}_1(\mu_1, \mu')$ and hence $(\mu, \mu') \notin \hat{\mathcal{T}}(\mu_1, \mu')$. To avoid double counting in our estimates on $\| \mathcal{T} \|$ we disqualify the (intuitively smaller) pre-team $\hat{\mathcal{T}}(\mu_1, \mu')$ in these settings.

If the genesis of $\hat{\mathcal{T}}(\mu, \mu')$ is of type (G4), then again it may happen that what remains to the right of $\hat{\mathcal{T}}(\mu, \mu')$ at some time before $\hat{t}_1(\mu, \mu')$ is a pre-team. In this case, we disqualify the (intuitively larger) pre-team $\hat{\mathcal{T}}(\mu, \mu')$.

The pre-teams that remain after these disqualifications are now defined to be *teams*.

A typical team will be denoted $\mathcal{T}$ and all hats will be dropped from the notation for their associated objects (e.g. we write $\mathfrak{T}(\mu, \mu')$ instead of $\hat{\mathfrak{T}}(\mu, \mu')$).

A team is said to be *short* if $\| \mathcal{T} \| \leq \lambda_0$ or $\sum_{(\mu,\mu') \in \mathcal{T}} |C_{(\mu,\mu')}(2)| \leq \lambda_0$. Let Σ denote the set of short teams.

Lemma 9.7. *Teams of genesis (G4) are short.*

Proof. Lemma 9.4 implies that $\mathfrak{T}$ is in the immediate future of an interval of length at most C_4. And we have decreed (Remark 9.5) that $MC_4 < \lambda_0$. $\square$

We wish our ultimate definition of a team to be such that every pair (μ, μ') with $C_{(\mu,\mu')}(2)$ non-empty is assigned to a team. The above definition fails to achieve this because of two phenomena: first, a pre-team $\hat{\mathcal{T}}(\mu, \mu')$ with genesis of type (G4) may have been disqualified, leaving (μ, μ') teamless; second, in our initial discussion of pre-teams (the first paragraph of Section 9.1) we excluded pairs (μ, μ') with $|C_{(\mu,\mu')}(2)| \leq 2B$. The following definitions remove these difficulties.

Definition 9.8 (Virtual team members)**.** If a pre-team $\hat{\mathcal{T}}(\mu,\mu')$ of type (G4) is disqualified under the terms of Definition 9.6 and the smaller team necessitating disqualification is $\hat{\mathcal{T}}(\mu_1,\mu')$, then we define $(\mu,\mu') \in_v \hat{\mathcal{T}}(\mu_1,\mu')$ and $\hat{\mathcal{T}}(\mu,\mu') \subset_v \hat{\mathcal{T}}(\mu_1,\mu')$. We extend the relation $\subset_v$ to be transitive and extend $\in_v$ correspondingly. If $(\mu,\mu') \in_v \mathcal{T}$ then (μ_2,μ') is said to be a *virtual member* of the team $\mathcal{T}$.

Definition 9.9. If (μ,μ') is such that $1 \le |C_{(\mu,\mu')}(2)| \le 2B$ and (μ,μ') is neither a member nor a virtual member of any previously defined team, then we define $\mathcal{T}_{(\mu,\mu')} := \{(\mu,\mu')\}$ to be a (short) team with $\| \mathcal{T}_{(\mu,\mu')} \| = |C_{(\mu,\mu')}(2)|$.

Lemma 9.10. *Every $(\mu,\mu') \in \mathcal{Z}$ with $C_{(\mu,\mu')}(2)$ non-empty is a member or a virtual member of exactly one team, and there are less than $2\,|\partial\Delta|$ teams.*

Proof. The first assertion is an immediate consequence of the preceding three definitions, and the second follows from the fact that $|\mathcal{Z}| < 2\,|\partial\Delta|$. □

9.3. **Pincers associated to teams of Genesis (G3).** In this subsection we describe the pincer $\Pi_{\mathcal{T}}$ canonically associated to each team of genesis $(G3)$. The definition of $\Pi_{\mathcal{T}}$ involves the following concept which will prove important also for teams of other genesis.

Definition 9.11. We define the *narrow past* of a team $\mathcal{T}$ to be the set of constant edges that have a future in $\mathfrak{T}$. The narrow past may have several components at each time, the set of which are ordered left to right according to the ordering in $\mathfrak{T}$ of their futures. We call these components *sections*.

For the remainder of this subsection we consider only long teams of genesis (G3).

Definition 9.12 (The Pincer $\tilde{\Pi}_{\mathcal{T}}$)**.** The paths labelled $\hat{p}_l$ and $\hat{p}_r$ in Figure 20 determine a pincer and are defined as follows. Let $x(\mathcal{T})$ be the leftmost non-constant edge to the right of μ in the immediate past of $\mathcal{T}$, and let $x_1(\mathcal{T})$ be the edge that consumes it. Define $\tilde{p}_l(\mathcal{T})$ to be the path in $\mathcal{F}$ that traces the history of $x(\mathcal{T})$ to the boundary, and let $\tilde{p}_r(\mathcal{T})$ be the path that traces the history of $x_1(\mathcal{T})$. (Note that $x_1(\mathcal{T})$ is left-fast.)

Define $\tilde{t}_2(\mathcal{T})$ to be the earliest time at which the paths $\tilde{p}_l(\mathcal{T})$ and $\tilde{p}_r(\mathcal{T})$ lie in the same corridor. The segments of the paths $\tilde{p}_l(\mathcal{T})$ and $\tilde{p}_r(\mathcal{T})$ after this time, together with the path joining them along the bottom of the corridor at time $\tilde{t}_2(\mathcal{T})$ form a pincer. We denote this pincer $\tilde{\Pi}_{\mathcal{T}}$.

The Pincer Lemma argues for the regular disappearance of colours within a pincer during those times when more than two colours continue to survive along the corridors of $\tilde{\Pi}_{\mathcal{T}}$. However, when there are only two colours the situation is more complicated.

We claim that the following situation cannot arise: $\text{time}(S_{\hat{\Pi}_{\mathcal{T}}}) \le t_1(\mathcal{T}) - T_0$, the path $\tilde{p}_l(\mathcal{T})$ and the entire narrow past of $\mathcal{T}$ are in the same corridor at time $t_1(\mathcal{T}) - T_0$, and at this time they are separated only by constant edges. For if this were the case, then the colour of $\tilde{p}_r(\mathcal{T})$ would ϕ-neuter the colour of $\tilde{p}_l(\mathcal{T})$ but would take more than T_0 steps to do so, contradicting the Two Colour Lemma. Thus at least one of the three hypotheses in the first sentence of this paragraph is false; we consider the three possibilities. The troublesome case (3) leads to a cascade of pincers as depicted in Figure 23.

Definition 9.13 (The Pincer $\Pi_{\mathcal{T}}$ and times $t_2(\mathcal{T})$ and $t_3(\mathcal{T})$)**.**

(1) *Some section of the narrow past of $\mathcal{T}$ is not in the same corridor as $\tilde{p}_l(\mathcal{T})$ at time $t_1(\mathcal{T}) - T_0$:* In this case[13] we define $t_2(\mathcal{T}) = t_3(\mathcal{T})$ to be the earliest time at which the entire narrow past of $\mathcal{T}$ lies in the same corridor as $\tilde{p}_l(\mathcal{T})$ and has length at least λ_0.

(2) *Not case (1), there are no non-constant edges between $\tilde{p}_l(\mathcal{T})$ and the narrow past of $\mathcal{T}$ at time $t_1(\mathcal{T}) - T_0$:* In this case $\mathrm{time}(S_{\tilde{\Pi}_{\mathcal{T}}}) > t_1(\mathcal{T}) - T_0$. We define $\Pi_{\mathcal{T}} = \tilde{\Pi}_{\mathcal{T}}$ and $t_3(\mathcal{T}) = \mathrm{time}(S_{\Pi_{\mathcal{T}}})$. If the narrow past of $\mathcal{T}$ at time $t_1(\mathcal{T}) - T_0$ has length less than λ_0, we define $t_2(\mathcal{T}) = t_3(\mathcal{T})$, and otherwise $t_2(\mathcal{T}) = \tilde{t}_2(\mathcal{T})$.

(3) *Not in case (1) or case (2):* In this case there is at least one non-constant edge between the narrow past of $\mathcal{T}$ and $\tilde{p}_l(\mathcal{T})$ at $t_1(\mathcal{T}) - T_0$. We pass to the latest time at which there is such an intervening non-constant edge and consider the path $\tilde{p}_l'(\mathcal{T})$ that traces the history of the leftmost intervening non-constant edge $x'(\mathcal{T})$ and the path $\tilde{p}_r'(\mathcal{T})$ that traces the history of the edge $x_1'(\mathcal{T})$ that cancels with $x'(\mathcal{T})$. We define $\tilde{t}_2'(\mathcal{T})$ to be the earliest time at which the paths $\tilde{p}_l'(\mathcal{T})$ and $\tilde{p}_r'(\mathcal{T})$ lie in the same corridor and consider the pincer formed by the segments of the paths $\tilde{p}_l'(\mathcal{T})$ and $\tilde{p}_r'(\mathcal{T})$ after time $\tilde{t}_2'(\mathcal{T})$ together with the path joining them along the bottom of the corridor at time $\tilde{t}_2'(\mathcal{T})$.

We now repeat our previous analysis with the primed objects $\tilde{p}_l'(\mathcal{T}), \tilde{t}_2'(\mathcal{T})$ *etc.* in place of $\tilde{p}_l(\mathcal{T}), \tilde{t}_2(\mathcal{T})$ *etc.*, checking whether we now fall into case (1) or (2); if we do not then we pass to $\tilde{p}_l''(\mathcal{T}), \tilde{t}_2''(\mathcal{T})$ *etc.*, and iterate the analysis until we do indeed fall into case (1) or (2), at which point we acquire the desired definitions of $\Pi_{\mathcal{T}}$, $t_2(\mathcal{T})$, $t_3(\mathcal{T})$.

Define $p_l(\mathcal{T})$ (resp. $p_r(\mathcal{T})$) to be the left (resp. right) boundary path of the pincer $\Pi_{\mathcal{T}}$ extended backwards in time through $\mathcal{F}$ to $\partial\Delta$. Define $p_l^+(\mathcal{T})$ to be the sequence of non-constant edges (one at each time) lying immediately to the right of the narrow past of $\mathcal{T}$ from the top of $\Pi_{\mathcal{T}}$ to time $t_1(\mathcal{T})$. (These are edges of the leftmost of the primed $\tilde{p}_l(\mathcal{T})$ considered in case (3).)

Definition 9.14. Let $\mathcal{T}$ be a long team of genesis (G3). Let $\chi_P(\mathcal{T})$ be the set of colours containing the paths $\tilde{p}_l(\mathcal{T}), \tilde{p}_l'(\mathcal{T}), \tilde{p}_l''(\mathcal{T}), \dots$ that arise in (iterated applications of) case (3) of Definition 9.13 but do not become $p_l(\mathcal{T})$.

[13] this includes the possibility that $\tilde{p}_l(\mathcal{T})$ does not exist at time $t_1(\mathcal{T}) - T_0$

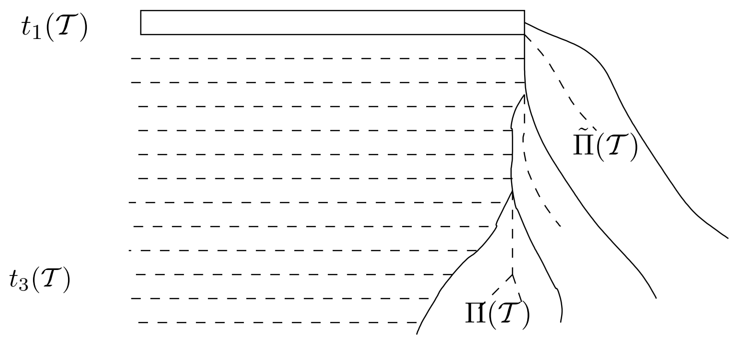

FIGURE 23. The cascade of pincers.

The preceding definitions are framed so as to make the following important facts self-evident.

Lemma 9.15.

(1) If $\mathcal{T}$ is a long team of genesis (G3),

$$t_1(\mathcal{T}) - t_3(\mathcal{T}) \le T_0(|\chi_P(\mathcal{T})| + 1).$$

(2) If $\mathcal{T}_1$ and $\mathcal{T}_2$ are disjoint then $\chi_P(\mathcal{T}_1) \cap \chi_P(\mathcal{T}_2) = \emptyset$.

9.4. **The length of teams.**

Definition 9.16. Define $\text{down}_1(\mathcal{T}) \subset \partial\Delta$ to consist of those edges e that are labelled t and satisfy one of the following conditions:

1. e is at the left end of a corridor containing a section of the narrow past of $\mathcal{T}$ that is not leftmost at that time;
2. e is at the right end of a corridor containing a section of the narrow past of $\mathcal{T}$ that is not rightmost at that time;
3. e is at the right end of a corridor which contains the rightmost section of the narrow past of $\mathcal{T}$ at that time but which does not intersect $p_l(\mathcal{T})$.

All of the edges shown on the boundary in Figure 24 are contained in $\text{down}_1(\mathcal{T})$.

Definition 9.17. Define $\partial^{\mathcal{T}} \subset \partial\Delta$ to be the set of (necessarily constant) edges that have a preferred future in $\mathfrak{T}$.

We record an obvious disjointness property of the sets defined above.

Lemma 9.18.

(1) For distinct teams $\mathcal{T}_1$ and $\mathcal{T}_2$, $\partial^{\mathcal{T}_1}$ and $\partial_{\mathcal{T}_2}$ are disjoint.
(2) For distinct teams $\mathcal{T}_1$ and $\mathcal{T}_2$, $\text{down}_1(\mathcal{T}_1)$ and $\text{down}_1(\mathcal{T}_2)$ are disjoint.

Definition 9.19. Suppose that $\mathcal{T}$ is a team of genesis (G3). We define $Q(\mathcal{T})$ be the set of edges ε with the following properties: $p_l(\mathcal{T})$ passes through ε before time $t_3(\mathcal{T})$, and the corridor S with $\varepsilon \in \perp(S)$ contains the entire narrow past of $\mathcal{T}$ and this narrow past has length at least λ_0.

The following lemma gives us a bound on $|\mathfrak{T}|$, which will reduce our task to that of bounding $|Q(\mathcal{T})|$ for teams of genesis (G3).

Lemma 9.20.

1. If the genesis of $\mathcal{T}$ is of type (G1) or (G2), then

$$\| \mathcal{T} \| \leq 2MC_4 \, |\mathrm{down}_1(\mathcal{T})| + |\partial^{\mathcal{T}}|.$$

2. If the genesis of $\mathcal{T}$ is of type (G3), then

$$\| \mathcal{T} \| \leq 2MC_4 \, |\mathrm{down}_1(\mathcal{T})| + |\partial^{\mathcal{T}}| + 2MC_4 \, |Q(\mathcal{T})| + 2MC_4 T_0\big(|\chi_P(\mathcal{T})| + 1\big) + \lambda_0.$$

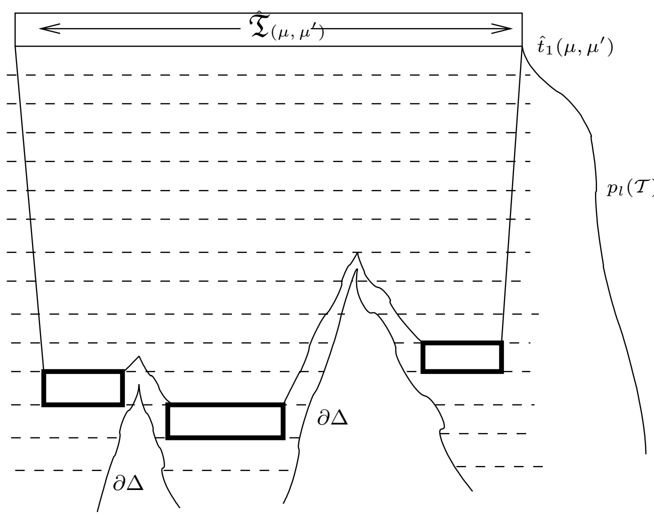

FIGURE 24. Bounding the size of a team in terms of $|\mathrm{down}_1|$ and $|p_l|$

Proof. The first thing to observe is that at any stage in the past of $\mathfrak{T}$ the set of letters lying in a single corridor form a connected region. As in Lemma 9.4, this is simply a matter of noting that if $\phi(aub) = w$ where w, a and b consist only of constant letters, then u must equal a word in constant letters.

Consider the past of $\mathfrak{T}$ at a time t. Write k_t for the number of corridors that contain a non-trivial component of this past. The total increase in length of these components when one goes forward to time $t+1$ is bounded by $2MC_4k_t$, since the connectedness of the past implies that the only growth that can happen for existing components occurs at their extremities, where a block of at most MC_4 constant letters may be added. This follows from Lemma 9.4. Also at time $t+1$, constant letters from $\partial\Delta$ may join the past of $\mathfrak{T}$, and there may be new components of constant letters (each of length less than $2MC_4$) whose ancestors at time t were non-constant letters. Thus we have three possible causes of increase. The first and third account for growth of at most $2MC_4k_{t+1}$ and the second (boundary) contribution is the number of elements of $\partial^{\mathcal{T}}$ that occur at time $t+1$. If the genesis of $\mathcal{T}$ is of type (G1) or (G2), then at least k_{t+1} edges of $\mathrm{down}_1(\mathcal{T})$ occur at time t, compensating us for the growth summand $2MC_4k_{t+1}$. If the genesis of $\mathcal{T}$ is of type (G3) then we still have the above compensation *except* at those times where no edges of $\mathrm{down}_1(\mathcal{T})$ occur. At these latter times the whole of the narrow past of $\mathcal{T}$ lies in a single corridor through which $p_l(\mathcal{T})$ passes. Since the narrow past lies in a single corridor, it is connected and grows at most $2MC_4$ when moving forward one unit of time (unless added to by $\partial^{\mathcal{T}}$).

The summands $2MC_4\,|Q(\mathcal{T})|$ and $2MC_4T_0\big(|\chi_P(\mathcal{T})|+1\big)$ in item (2) of the lemma account for the growth of the narrow past in the intervals of time below $t_3(\mathcal{T})$, and from $t_3(\mathcal{T})$ to $t_1(\mathcal{T})$, respectively. The additional summand λ_0 allows us to desist from our estimating if the narrow past of $\mathcal{T}$ ever shrinks to have length less than λ_0. ☐

9.5. **Bounding the size of $Q(\mathcal{T})$.** For the remainder of this section we concentrate exclusively on long teams of genesis (G3) with $Q(\mathcal{T})$ non-empty. We denote the set of such teams by $\mathcal{G}_3$. Our goal is to bound $|Q(\mathcal{T})|$. (In the light of our previous results, this will complete the required analysis of the length of teams.)

Recall from Definition 9.13 that for teams of genesis (G3), the paths $p_l(\mathcal{T})$ and $p_r(\mathcal{T})$ and the chain of 2-cells joining them in the corridor at time $t_2(\mathcal{T})$ form a pincer denoted $\Pi_{\mathcal{T}}$. The set $\chi(\Pi_{\mathcal{T}})$ was defined in Definition 8.24.

An important feature of teams in $\mathcal{G}_3$ is:

Lemma 9.21. *If $\mathcal{T}\in\mathcal{G}_3$ then there exists a block of at least λ_0 constant edges immediately adjacent to $\Pi_{\mathcal{T}}$ at each time from $t_3(\mathcal{T})$ to the top of $\Pi_{\mathcal{T}}$, and adjacent to $p_l^+(\mathcal{T})$ from then until $t_1(\mathcal{T})$. (At time $t_1(\mathcal{T})$ this block contains $\mathfrak{T}$.)*

Proof. The hypothesis that $Q(\mathcal{T})$ is non-empty means that the narrow past of $\mathcal{T}$ at some time before $t_3(\mathcal{T})$ has length at least λ_0 and is contained in the same corridor as $p_l(\mathcal{T})$ (see Definition 9.19). The definition of $t_3(\mathcal{T})$ implies that the narrow past of $\mathcal{T}$ is contained in a block of constant letters immediately adjacent to $p_l(\mathcal{T})$ or $p_l^+(\mathcal{T})$ from time $t_3(\mathcal{T})$ until $t_1(\mathcal{T})$. Since the length of the narrow past of $\mathcal{T}$ does not decrease before $t_1(\mathcal{T})$, these blocks of constant letters must have length at least λ_0. ☐

The following is an immediate consequence of the Pincer Lemma.

Lemma 9.22. *For all $\mathcal{T}\in\mathcal{G}_3$,*

$$t_3(\mathcal{T})-t_2(\mathcal{T})=\mathrm{Life}(\Pi_{\mathcal{T}})\leq T_1(|\chi(\Pi_{\mathcal{T}})|+1).$$

Lemma 9.23. *If $\mathcal{T}_1, \mathcal{T}_2 \in \mathcal{G}_3$ are distinct teams then $\chi(\Pi_{\mathcal{T}_1}) \cap \chi(\Pi_{\mathcal{T}_2}) = \emptyset$.*

Proof. The pincers $\Pi_{\mathcal{T}_i}$ are either disjoint or else one is contained in the other. In the latter case, say $\Pi_{\mathcal{T}_1} \subset \Pi_{\mathcal{T}_2}$, the existence of the block of λ_0 constant edges established in Lemma 9.21 means that $\Pi_{\mathcal{T}_1}$ is actually nested in $\mathcal{T}_2$ in the sense of Definition 8.24. Thus $\chi(\Pi_{\mathcal{T}_1}) \cap \chi(\Pi_{\mathcal{T}_2}) = \emptyset$ (by Definition 8.24). □

Corollary 9.24. $\sum_{\mathcal{T} \in \mathcal{G}_3} t_3(\mathcal{T}) - t_2(\mathcal{T}) \leq 3T_1\, |\partial\Delta|$.

It remains to bound the number of edges in $Q(\mathcal{T})$ which occur before $t_2(\mathcal{T})$; this is cardinality of the following set.

Definition 9.25. For $\mathcal{T} \in \mathcal{G}_3$ we define $\mathrm{down}_2(\mathcal{T})$ to be the set of edges in $\partial\Delta$ that lie at the righthand end of a corridor containing an edge in $Q(\mathcal{T})$ before time $t_2(\mathcal{T})$.

The remainder of this section is dedicated to obtaining a bound on

$$\sum_{\mathcal{T} \in \mathcal{G}_3} |\mathrm{down}_2(\mathcal{T})|,$$

(see Corollary 9.31).

At this stage our task of bounding $\|\, \mathcal{T}\, \|$ would be complete if the the sets $\mathrm{down}_2(\mathcal{T})$ associated to distinct teams were disjoint — unfortunately they need not be, because of the possible nesting of teams as shown in Figures 17 and 25. Thus we shall be obliged to seek further pay-off for our troubles. To this end we shall identify two sets of consumed colours $\chi_c(\mathcal{T})$ and $\chi_\delta(\mathcal{T})$ that arise from the nesting of teams.

In order to analyse the effect of nesting we need the following vocabulary.

There is an obvious left-to-right ordering of those paths in the forest $\mathcal{F}$ which begin on the arc of $\partial\Delta \smallsetminus \partial S_0$ that commences at the initial vertex of the left end of S_0. (First one orders the trees, then the relative order between paths in a tree is determined by the manner in which they diverge; the only paths which are not ordered relative to each other are those where one is an initial segment of the other, and this ambiguity will not concern us.)

Notation: We write $\mathcal{G}_3'$ for the set of teams $\mathcal{T} \in \mathcal{G}_3$ such that $\mathrm{down}_2(\mathcal{T}) \neq \emptyset$.

We shall need the following obvious separation property.

Lemma 9.26. *Consider $\mathcal{T} \in \mathcal{G}_3'$. If a path p in $\mathcal{F}$ is to the left of $p_l(\mathcal{T})$ and a path q is the right of $p_r(\mathcal{T})$, then there is no corridor connecting p to q at any time $t < t_2(\mathcal{T})$.*

Proof. The hypothesis $\mathrm{down}_2(\mathcal{T}) \neq \emptyset$ implies that before $t_2(\mathcal{T})$ the paths $p_l(\mathcal{T})$ and $p_r(\mathcal{T})$ are not in the same corridor. □

Definition 9.27. $\mathcal{T}_1 \in \mathcal{G}_3'$ is said to be *below* $\mathcal{T}_2 \in \mathcal{G}_3'$ if $p_l(\mathcal{T}_2)$ and $p_r(\mathcal{T}_2)$ both lie between $p_l(\mathcal{T}_1)$ and $p_r(\mathcal{T}_1)$ in the left-right ordering described above.

$\mathcal{T}_1$ is said to be *to the left* of $\mathcal{T}_2$ if both $p_l(\mathcal{T}_2)$ and $p_r(\mathcal{T}_2)$ lie to the right of $p_r(\mathcal{T}_1)$.

We say that $\mathcal{T}$ is at *depth* 0 if there are no teams above it. Then, inductively, we say that a team is at depth $d+1$ if d is the maximum depth of those teams above $\mathcal{T}$.

A *final depth* team is one with no teams below it.

Note that there is a complete left-to-right ordering of teams $\mathcal{T} \in \mathcal{G}_3$ at any given depth.

Lemma 9.28. *If there is a team from $\mathcal{G}_3'$ below $\mathcal{T} \in \mathcal{G}_3'$, then $t_1(\mathcal{T}) \geq \text{time}(S_0) \geq t_2(\mathcal{T})$.*

Proof. The first thing to note is that if $\text{time}(S_0)$ were less than $t_2(\mathcal{T})$, then the narrow past of $\mathcal{T}$ at time $t_2(\mathcal{T})$ must contain at least λ_0 edges. This is because the length of the narrow past of $\mathcal{T}$ cannot decrease before $t_1(\mathcal{T})$, and at $\text{time}(S_0)$ the narrow past is the union of the intervals $C_{(\mu,\mu')}(2)$ with $(\mu,\mu') \in \mathcal{T}$, which has length at least λ_0 since $\mathcal{T}$ is assumed not to be short.

Thus if $\text{time}(S_0) < t_2(\mathcal{T})$ then we are in the non-degenerate situation of Definition 9.13 and the defining property of $t_2(\mathcal{T})$ means that before time $t_2(\mathcal{T})$ no edge to the right of $p_r(\mathcal{T})$ lies in the same corridor as all the colours of $\mathcal{T}$ (cf. Lemma 9.26). In particular this is true of the past of the reaper of $\mathcal{T}$ (assuming that it has a past at time $t_2(\mathcal{T})$). On the other hand, the reaper of $\mathcal{T}$ has a past in S_0 (by the very definition of a team), as do all of the colours of $\mathcal{T}$. And since they lie in a common corridor at $\text{time}(S_0)$, they must also do so at all times up to $t_1(\mathcal{T})$. This contradiction implies that in fact $\text{time}(S_0) \geq t_2(\mathcal{T})$.

Consider Figure 17. Suppose that $\mathcal{T}' \in \mathcal{G}_3'$ is below $\mathcal{T}$. The proof of Lemma 9.21 tells us that there is a block of constant edges extending from the top of $\Pi_{\mathcal{T}'}$ containing the narrow past of $\mathcal{T}'$, and there is a similarly long block extending from the path $p_l^+(\mathcal{T})$ at each subsequent time until $t_1(\mathcal{T}')$. Thereafter the future of the block is contained in the block of constant edges that evolves into the union of the $C_{(\mu,\mu')}(2) \subseteq \perp(S_0)$ with $(\mu,\mu') \in \mathcal{T}'$, which is long by hypothesis.

At no time can this evolving block extend across $p_l(\mathcal{T})$ because by definition the edges along $p_l(\mathcal{T})$ are labelled by non-constant letters. Thus the evolving block is trapped to the right of $p_l(\mathcal{T})$ and to the left of $p_r(\mathcal{T})$. In particular, it must vanish entirely before the time at the top of the pincer $\Pi_{\mathcal{T}}$, which is no later than $t_1(\mathcal{T})$ and therefore $t_1(\mathcal{T}) \geq \text{time}(S_0)$. □

The following is the main result of this section.

Lemma 9.29. *There exist sets of colours $\chi_c(\mathcal{T})$ and $\chi_\delta(\mathcal{T})$ associated to each team $\mathcal{T} \in \mathcal{G}_3'$ such that the sets associated to distinct teams are disjoint and the following inequalities hold.*

For each fixed team $\mathcal{T}_0 \in \mathcal{G}_3'$ (of depth d say), the teams of depth $d+1$ that lie below $\mathcal{T}_0$ may be described as follows:

- *There is at most one* distinguished team $\mathcal{T}_1$, *and*
$$\|\mathcal{T}_1\| \leq 2B\Big(T_1(1+|\chi(\Pi_{\mathcal{T}_0})|) + T_0(|\chi_P(\mathcal{T}_0)|+1)\Big).$$
- *There are some number of final-depth teams.*
- *For each of the remaining teams $\mathcal{T}$ we have*
$$|\mathrm{down}_2(\mathcal{T}_0) \cap \mathrm{down}_2(\mathcal{T})| \leq T_1\Big(1+|\chi_c(\mathcal{T})|\Big) + T_0\Big(|\chi_\delta(\mathcal{T})|+2\Big).$$

Proof. The first thing to note is that if two teams $\mathcal{T}, \mathcal{T}' \in \mathcal{G}_3'$ are at the same depth, then $\mathrm{down}_2(\mathcal{T})$ and $\mathrm{down}_2(\mathcal{T}')$ are disjoint. Indeed if $\mathcal{T}$ is to the left of $\mathcal{T}'$, then at times before $t_2(\mathcal{T})$ the paths $p_l(\mathcal{T})$ and $p_l(\mathcal{T}')$ never lie in the same corridor. Let $\mathcal{T} \in \mathcal{G}_3'$ be a team of level $d+1$ that is below $\mathcal{T}_0$ and consider the edge e at the right end of a corridor earlier than $t_2(\mathcal{T})$ that contains an edge in $Q(\mathcal{T})$. We are concerned with the fact that this edge may be in $\mathrm{down}_2(\mathcal{T}_0)$. In this situation we say that $\mathcal{T}_0$ and $\mathcal{T}$ *double count* e.

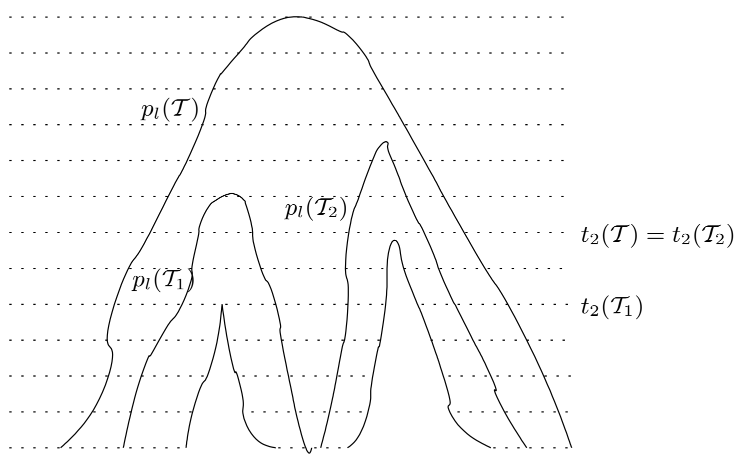

FIGURE 25. A depiction of double-counting

Let $\mathcal{T}_1, \ldots, \mathcal{T}_r$ be the teams in $\mathcal{G}_3'$ of depth $d+1$ which double-count with $\mathcal{T}_0$, ordered from left to right, with the final-depth teams deleted. We define $\chi_c(\mathcal{T})$ to be empty for teams not on this list. $\mathcal{T}_1$ will be the distinguished team.

Since there is no double-counting between teams of the same level, the sets of times at which $\mathcal{T}_1, \ldots, \mathcal{T}_r$ double-count with $\mathcal{T}_0$ must be disjoint. Indeed if $i < j$ then the set of times at which $\mathcal{T}_i$ double-counts with $\mathcal{T}_0$ is earlier than the set of times at which $\mathcal{T}_j$ double-counts with $\mathcal{T}_0$ (Lemma 9.26). Moreover, the times for each $\mathcal{T}_i$ form an interval, which we denote $\mathcal{I}_i$.

We assume $r \geq 2$ and describe the construction of the sets $\chi_c(\mathcal{T}_i)$ and $\chi_\delta(\mathcal{T}_i)$ that account for double-counting.

The first thing to note is that each $\mathcal{I}_i$ must be later than $t_2(\mathcal{T}_1)$, by Lemma 9.26. The second thing to note is that the entire interval of time $\mathcal{I}_i$ must also be earlier than $t_1(\mathcal{T}_1)$. Indeed if some double-counting by $\mathcal{T}_i$ and $\mathcal{T}_0$ were to occur after $t_1(\mathcal{T}_1)$, then we would have $t_2(\mathcal{T}_k) > t_1(\mathcal{T}_1)$. But then $\text{time}(S_0) > t_1(\mathcal{T}_1)$, so Lemma 9.28 would imply that there was no team below $\mathcal{T}_1$, contrary to hypothesis.

We separately consider the intervals $\mathcal{I}_i \cap [t_2(\mathcal{T}_1), t_3(\mathcal{T}_1)]$ and $\mathcal{I}_i \cap [t_3(\mathcal{T}_1), t_1(\mathcal{T}_1)]$, whose union is all of $\mathcal{I}_i$.

For that part of $\mathcal{I}_i$ before $t_3(\mathcal{T}_1)$, the proofs of the Pincer Lemma (Theorem 8.26) and Proposition 8.7 tell us that colours in $\chi(\Pi_{\mathcal{T}_1})$ will be consumed at the rate of at least one per T_1 units of time. Define $\chi_c(\mathcal{T}_i)$ to be this set of consumed colours. We have

$$\Big| \mathcal{I}_i \cap [t_2(\mathcal{T}_1), t_3(\mathcal{T}_1)] \Big| \leq T_1(1 + |\chi_c(\mathcal{T}_i)|).$$

Now consider $\mathcal{I}_i \cap [t_3(\mathcal{T}_1), t_1(\mathcal{T}_1)]$. Define $\chi_\delta(\mathcal{T}_i)$ as follows. The discussion in Definition 9.13 shows that in any period of time of length T_0 in the interval $[t_3(\mathcal{T}_1), t_1(\mathcal{T}_1)]$ at least one colour in $\chi_P(\mathcal{T}_1)$ disappears. Let $\chi_\delta(\mathcal{T}_i)$ be the set of colours in $\chi_P(\mathcal{T}_1)$ which disappear during $\mathcal{I}_i \cap [t_3(\mathcal{T}_1), t_1(\mathcal{T}_1)]$ (these disappearances correspond to the discontinuities in the 'path' $p_l^+(\mathcal{T}_1)$). By construction, we then have[14]

$$\Big| \mathcal{I}_i \cap [t_3(\mathcal{T}_1), t_1(\mathcal{T}_1)] \Big| \leq T_0(|\chi_\delta(\mathcal{T}_i)| + 2),$$

and combining these estimates we have

$$|\mathcal{I}_i| \leq T_1\Big(1 + |\chi_c(\mathcal{T}_i)|\Big) + T_0\Big(|\chi_\delta(\mathcal{T}_i)| + 2\Big),$$

as required. Since the intervals $\mathcal{I}_i$ are disjoint, the sets $\chi_c(\mathcal{T}_i)$, $i = 2, \ldots, r$ are mutually disjoint. And by construction, these sets are also disjoint from the sets associated to teams other than the $\mathcal{T}_i$ under consideration (i.e. those under other depth d teams, or those of different depths). The same considerations hold for the sets $\chi_\delta(\mathcal{T}_i)$, $i = 2, \ldots, r$.

In Figure 26, the shaded region is where we recorded the regular disappearance of the colours forming $\chi_c(\mathcal{T}_i)$, whilst in Figure 27, the shaded region is where we recorded the regular disappearance of the colours forming $\chi_\delta(\mathcal{T}_i)$.

It remains to establish the inequality

$$\| \mathcal{T}_1 \| \leq 2B\Big(T_1(|\chi(\Pi_{\mathcal{T}_0})| + 1) + (|\chi_P(\mathcal{T}_0)| + 1)\Big).$$

We first note (as in the proof of Lemma 9.28) that $\mathfrak{T}_1$ is trapped between $p_l(\mathcal{T})$ and $p_r(\mathcal{T})$, so it must be consumed entirely between the times $t_1(\mathcal{T}_1)$

[14]There is a 2 rather than the familiar 1 on the right to account for the colour containing $p_l(\mathcal{T}_1)$, which is not included in $\chi_P(\mathcal{T}_1)$; there might be up to T_0 corridors between $t_3(\mathcal{T}_1)$ and the top of $\Pi_{\mathcal{T}_1}$.

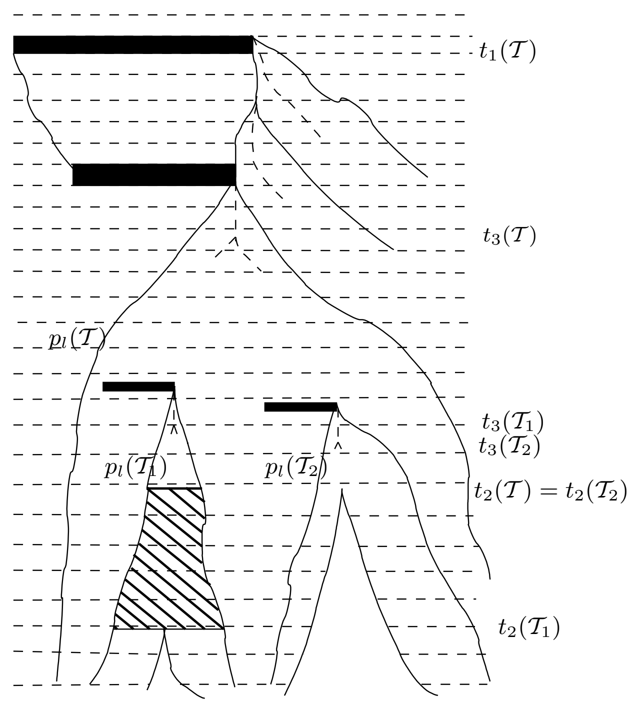

FIGURE 26. Finding the colours $\chi_c(\mathcal{T}_i)$

and $t_1(\mathcal{T}_0)$. But by the Bounded Cancellation Lemma, the length of the future of $\mathfrak{T}_1$ can decrease by at most $2B$ at each step in time. Therefore $\| \mathcal{T}_1 \| \leq 2B(t_1(\mathcal{T}_0) - t_1(\mathcal{T}_1))$.

$\mathcal{T}_1$ is assumed not be final-depth, so from Lemma 9.28 we have $t_2(\mathcal{T}_0) \leq \mathrm{time}(S_0) \leq t_1(\mathcal{T}_1)$. By combining these inequalities with Lemmas 9.22 and 9.15 we obtain:

$$\begin{aligned} \| \mathcal{T}_1 \| &\leq 2B\Big(t_1(\mathcal{T}_0) - t_1(\mathcal{T}_1)\Big) \\ &\leq 2B\Big(t_1(\mathcal{T}_0) - \mathrm{time}(S_0)\Big) \\ &\leq 2B\Big(t_1(\mathcal{T}_0) - t_2(\mathcal{T}_0)\Big) \\ &\leq 2B\Big[T_1\Big(1 + |\chi(\Pi_{\mathcal{T}_0})|\Big) + T_0\Big(|\chi_P(\mathcal{T}_0)| + 1\Big)\Big]. \end{aligned}$$

□

Corollary 9.30. *Summing over the set of teams $\mathcal{T} \in \mathcal{G}'_3$ that are not distinguished, we get*

$$\sum_{\mathcal{T}} \Big|\mathrm{down}_2(\mathcal{T})\Big| \leq 2\Big|\bigcup_{\mathcal{T}} \mathrm{down}_2(\mathcal{T})\Big| + \sum_{\mathcal{T}} T_1\Big(1+|\chi_c(\mathcal{T})|\Big) + \sum_{\mathcal{T}} T_0\Big(|\chi_\delta(\mathcal{T})|+2\Big).$$

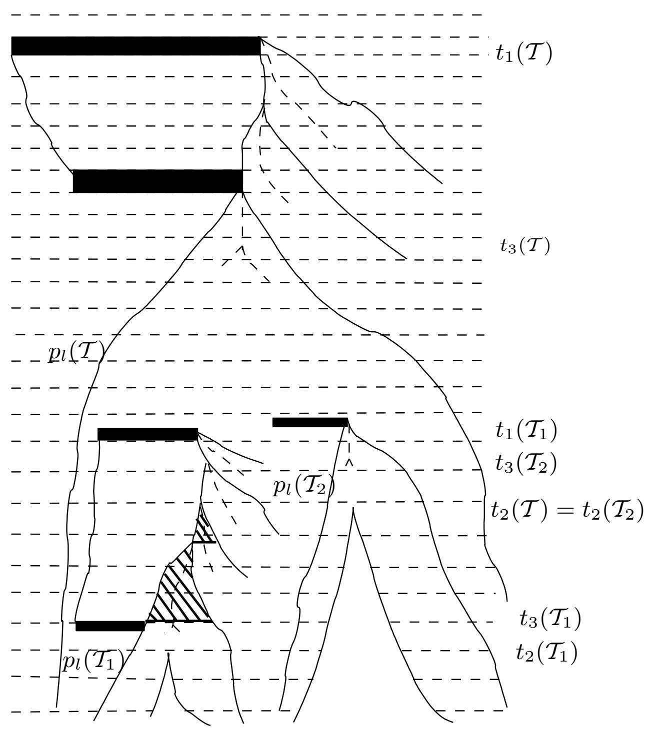

FIGURE 27. Finding the colours $\chi_\delta(\mathcal{T}_i)$

Proof. Suppose $\mathcal{T} \in \mathcal{G}_3'$ of depth $d+1$ is not final-depth and not distinguished, and that $\mathcal{T}$ double-counts with some $\mathcal{T}_0$ of depth d above it. Then, by Lemma 9.29, we have

$$\begin{aligned} |\mathrm{down}_2(\mathcal{T})| &= |\mathrm{down}_2(\mathcal{T}) \smallsetminus \mathrm{down}_2(\mathcal{T}_0)| + |\mathrm{down}_2(\mathcal{T}) \cap \mathrm{down}_2(\mathcal{T}_0)| \\ &\leq |\mathrm{down}_2(\mathcal{T}) \smallsetminus \mathrm{down}_2(\mathcal{T}_0)| + T_1(1 + |\chi_c(\mathcal{T})|) + T_0(2 + |\chi_\delta(\mathcal{T})|). \end{aligned}$$

Suppose that $\mathcal{T}' \in \mathcal{G}_3'$ is a team of depth $k < d$ and that $\mathcal{T}'$ is above $\mathcal{T}$. If $\mathcal{T}$ double-counts with $\mathcal{T}'$ at time t, then $\mathcal{T}$ double-counts with $\mathcal{T}_0$ at time t, by Lemma 9.26. Therefore, the set of edges that $\mathcal{T}$ double-counts with any team of lesser depth is exactly $\mathrm{down}_2(\mathcal{T}) \cap \mathrm{down}_2(\mathcal{T}_0)$.

Thus we have accounted for all double-counting other than than involving final depth teams. The factor 2 in the statement of the corollary accounts for this. □

And summing over the same set of teams again, we obtain:

Corollary 9.31.

$$\sum_{\mathcal{T}} |\mathrm{down}_2(\mathcal{T})| \;\leq\; |\partial\Delta|(2 + 3T_1 + 5T_0).$$

Proof. The sets of colours $\chi_c(\mathcal{T})$ and $\chi_\delta(\mathcal{T})$ are disjoint. And the union of the sets $\mathrm{down}_2(\mathcal{T})$ is a subset of $\partial\Delta$. The set of all colours and the set of edges in $\partial\Delta$ each have cardinality at most $|\partial\Delta|$. And the number of teams is less than $2\,|\partial\Delta|$ (Lemma 9.10). ☐

10. The Bonus Scheme

We have defined teams and obtained a global bound on $\sum \|\,\mathcal{T}\,\|$. If $C_{(\mu,\mu')}(2)$ is non-empty then (μ,μ') is a member or virtual member of a unique team. If this team is such that $t_1(\mathcal{T}) \geq \mathrm{time}(S_0)$, then no member of the team is virtual and we have the inequality

$$\|\,\mathcal{T}\,\| > \sum_{(\mu,\mu')\in\mathcal{T}} |C_{(\mu,\mu')}(2)| - B$$

established in Lemma 9.3. We indicated following this lemma how this inequality might fail in the case where $t_1(\mathcal{T}) < \mathrm{time}(S_0)$. In this section we take up this matter in detail and introduce a *bonus scheme* that assigns additional edges to teams in order to compensate for the possible failure of the above inequality when $t_1(\mathcal{T}) < \mathrm{time}(S_0)$.

By definition, at time $t_1(\mathcal{T})$ the reaper $\rho = \rho_{\mathcal{T}}$ lies immediately to the right of $\mathfrak{T}$. The edges of $\mathfrak{T}$ not consumed from the right by ρ by $\mathrm{time}(S_0)$ have a preferred future in S_0 that lies in $C_{(\mu,\mu')}(2)$ for some member $(\mu,\mu') \in \mathcal{T}$. However, not all of the edges of $C_{(\mu,\mu')}(2)$ need arise in this way: some may not have a constant ancestor at time $t_1(\mathcal{T})$. And if (μ,μ') is only a virtual member of $\mathcal{T}$, then no edge of $C_{(\mu,\mu')}(2)$ lies in the future of $\mathfrak{T}$. The *bonus* edges in $C_{(\mu,\mu')}(2)$ are a certain subset of those that do not have a constant ancestor at time $t_1(\mathcal{T})$. They are defined as follows.

Definition 10.1. Let $\mathcal{T}$ be a team with $t_1(\mathcal{T}) < \mathrm{time}(S_0)$ and consider a time t with $t_1(\mathcal{T}) < t < \mathrm{time}(S_0)$.

The *swollen future* of $\mathcal{T}$ at time t is the interval of constant edges beginning immediately to the left of the pp-future of $\rho_{\mathcal{T}}$.

Let e be a non-constant edge that lies immediately to the left of the swollen future of $\mathcal{T}$ but whose ancestor is not a right para-linear edge in this position. If e is a right para-linear and the (constant) rate at which e adds letters to the swollen future of $\mathfrak{T}$ is greater than the (constant) rate at which the future of the reaper cancels letters in the future of $\mathfrak{T}$, then we define e to be a *rascal*; if e is right-fast then we define it to be a *terror*. In both cases, we define the *bonus provided by* e to be the set of edges in the swollen future of $\mathcal{T}$ in S_0 that have e as their most recent non-constant ancestor, and are eventually consumed by $\rho_{\mathcal{T}}$.

The set $\mathrm{bonus}(\mathcal{T})$ is the union of the bonuses provided to $\mathcal{T}$ by all rascals and terrors.

Lemma 10.2. *For any team $\mathcal{T}$,*

$$\sum_{(\mu,\mu')\in\mathcal{T} \ \ or \ (\mu,\mu')\in_v T} |C_{(\mu,\mu')}(2)| \le \ \| \mathcal{T} \| \ + |\mathrm{bonus}(\mathcal{T})| + B.$$

Proof. If $t_1(\mathcal{T}) \geq \mathrm{time}(S_0)$, this follows immediately from Lemma 9.3. If $t_1(\mathcal{T}) < \mathrm{time}(S_0)$ then at each step in time between $t_1(\mathcal{T})$ and $\mathrm{time}(S_0)$ the only possible cause of growth in the length of the swollen future of the team is the possible action of a rascal or terror if such is present at that time. (There is no interaction of the swollen future with the boundary or singularities, because of the exclusions in the second paragraph of Definition 9.6.)

The swollen future has length $\| \mathcal{T} \|$ at time $t_1(\mathcal{T})$ and length at least $\sum |C_{(\mu,\mu')}(2)|$ at $\mathrm{time}(S_0)$. By definition, $|\mathrm{bonus}(\mathcal{T})|$ is a bound on the growth in length between these times. (The summand B is thus unnecessary in the case $t_1(\mathcal{T}) < \mathrm{time}(S_0)$.) □

The following lemma shows that our main task in this section will be to analyse the behaviour of rascals.

Lemma 10.3. *The sum of the lengths of the bonuses provided to all teams by terrors is less than $2M\,|\partial\Delta|$.*

Proof. Since it is right-fast, a terror will be separated from the team to which it is associated after one unit of time, and hence the bonus that it provides is less than M. There is at most one terror for each possible adjacency of colours and hence the total contributions of all terrors is less than $2M\,|\partial\Delta|$. □

The typical pattern of influence of rascals on a team is shown in Figure 28; there may be several times at which rascals appear at the left of $\mathcal{T}$ and provide a bonus for the team before being consumed from the left (or otherwise detached from the team).

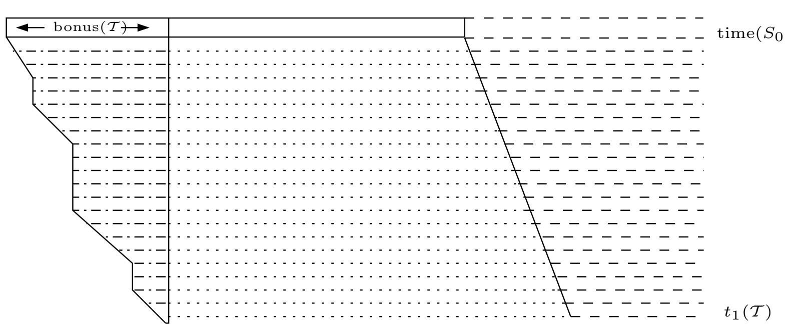

FIGURE 28. The generic situation below $\mathrm{time}(S_0)$.

Definition 10.4 (Rascals' Pincers)**.** We fix a team $\mathcal{T}$ with $t_1(\mathcal{T}) < \mathrm{time}(S_0)$ and consider the interval of time $[\tau_0(e), \tau_1(e)]$, where $\tau_0(e)$ is the time at which a rascal e appears at the left end of the swollen future of $\mathcal{T}$, and $\tau_1(e)$ is the time at which its future is no longer to the immediate left of the future of the swollen future of $\mathcal{T}$.

In the case where the pp-future $\hat{e}$ of e at time $\tau_1(e)$ is cancelled from the left by an edge e', we define $\tau_2(e)$ to be the earliest time when the pasts of $\hat{e}$ and e' are in the same corridor. The path in $\mathcal{F}$ that traces the pp-future of e up to $\tau_1(e)$ is denoted p_e and the path following through the ancestors of e' from $\tau_2(e)$ to $\tau_1(e)$ is denoted p'_e. The pincer[15] formed by p_e and p'_e with base at time $\tau_2(e)$ is denoted Π_e.

Lemma 10.5. *The total of all bonuses provided to all teams by rascals e with $\tau_1(e) \leq \mathrm{time}(S_0)$ is less than $(3T_1 + 2T_0 + 1)M\,|\partial\Delta|$.*

Proof. Consider a rascal e. We defer the case where e hits a singularity or the boundary. If this does not happen, the pp-future $\hat{e}$ of e at time $\tau_1(e)$ is cancelled from the left by an edge e' (which is right-fast since e is not constant). We consider the pincer Π_e defined above. The presence of the swollen future of $\mathcal{T}$ at the top of the pincer allows us to apply the Two Colour Lemma to conclude that $\tau_1(e) - T_0 \geq \mathrm{time}(S_{\Pi_e})$ (in the degenerate case discussed in the footnote, $\mathrm{time}(S_{\Pi_e})$ is replaced by $\tau_2(e)$). And the Pincer Lemma tells us that

$$\tau_1(e) - \tau_2(e) \leq T_1\Big(1 + |\chi(\Pi_e)|\Big) + T_0.$$

In fact, we could use $\tilde{\chi}(\Pi_e)$ instead of $\chi(\Pi_e)$ in this estimate because there cannot be any nesting amongst the pincers Π_e with $\tau_1(e) \leq \mathrm{time}(S_0)$, because nesting would imply that the swollen future of $\mathcal{T}$, which is immediately to the right of the lower rascal, would be trapped beneath the upper pincer, contradicting the fact that the team has a non-empty future in S_0.

In the case where e hits the boundary or is separated from the team by a singularity (at time $\tau_1(e)$) we define $\tau_2(e) = \tau_1(e)$. No matter what the fate of e, we define ∂^e to be the set of edges in $\partial\Delta$ at the left ends of corridors containing the future of e between $\tau_0(e)$ and $\tau_2(e)$. The sets ∂^e assigned to different rascals are disjoint, so summing over all rascals with $\tau_1(e) \leq \mathrm{time}(S_0)$ we have

$$\begin{aligned}\sum_e \Big(\tau_1(e) - \tau_0(e)\Big) &= \sum_e (\tau_1(e) - \tau_2(e)) + (\tau_2(e) - \tau_0(e)) \\ &\leq \sum_e T_1\Big(1 + |\chi(\Pi_e)|\Big) + T_0 + |\partial^e|.\end{aligned}$$

Since the sets $\chi(\Pi_e)$ and ∂^e are disjoint, the terms $T_1|\chi(\Pi_e)|$ and $|\partial^e|$ contribute less than $(T_1 + 1)\,|\partial\Delta|$ to this sum. And since the number of rascals is

[15]to lighten the terminology, here we allow the degenerate case where the "pincer" has no colours other than those of e and e'

bounded by the number of possible adjacencies of colours, the remaining terms contribute at most $(T_1+T_0)2\,|\partial\Delta|$. Thus

$$\sum_e \Big(\tau_1(e) - \tau_0(e)\Big) \;\le\; (3T_1 + 2T_0 + 1)\,|\partial\Delta|.$$

The bonus produced by each rascal in each unit of time is less than M, so the lemma is proved. □

It remains to consider the size of the bonuses provided by rascals e with $\tau_1(e) > \text{time}(S_0)$.

The bonuses that are not accounted for in Lemma 10.5 reside in blocks of constant edges along $\perp(S_0)$ each of which is the swollen future of some team, with a right para-linear letter at its left-hand end (the pp-future of a rascal) and a left para-linear letter at its left-hand end (the pp-future of the team's reaper).

Definition 10.6. A *left-biased* rascal e is one with $\tau_1(e) > \text{time}(S_0)$ that satisfies the following properties:

1. the pp-future of the rascal is (ultimately) consumed from the left by an edge of S_0,
2. the swollen future of $\mathcal{T}$ at time $\tau_1(e)$ has length at least λ_0 and the pp-future of the reaper $\rho_{\mathcal{T}}$ is still immediately to its right.

Definition 10.7. Let $\mathfrak{B} \subset \perp(S_0)$ be an interval of constant edges with a right para-linear letter at its left-hand end and a left-linear letter ρ at its right-hand end. We say that $\mathfrak{B}$ is *right biased* if ρ is ultimately consumed by an edge (to its right) in S_0. We define life($\mathfrak{B}$) to be the difference between time(S_0) and the time at which the left para-linear letter ρ is consumed. And we define the *effective volume* of $\mathfrak{B}$ to be the number of edges in $\mathfrak{B}$ that are ultimately consumed by ρ.

We have the following tautologous tetrad of possibilities covering the swollen teams whose bonuses are not entirely accounted for by Lemma 10.5.

Lemma 10.8. *Let $\mathfrak{B} \subset \perp(S_0)$ be an interval of constant edges that is the swollen future of a team with a rascal at its left-hand end and a left para-linear letter ρ at its right-hand end. Then at least one of the following holds:*

(i) *the length of $\mathfrak{B}$ is at most λ_0;*
(ii) *$\mathfrak{B}$ is the swollen future of a team with a left-biased rascal;*
(iii) *$\mathfrak{B}$ is right-biased;*
(iv) *neither of the non-constant letters at the ends of $\mathfrak{B}$ is ultimately consumed by an edge of S_0.*

We note here that when the length of $\mathfrak{B}$ is at most λ_0 then we have a short team, and we have already accounted for short teams. The following three lemmas correspond to eventualities (ii) to (iv).

Lemma 10.9. *The sum of the bonuses provided to all teams by left-biased rascals is less than* $(2M + 6MT_1 + 4MT_0 + 2\lambda_0 + 6BT_1 + 4BT_0)\,|\partial\Delta|$.

Proof. The proof of this result is similar to the work done in the previous section. We have a pincer Π_e associated to the rascal e. Since we are only concerned with the times when the rascal is immediately adjacent to a block of constant letters, it must be that at time $\tau_1(e) - T_0$ either we are below $\tau_0(e)$ or $\mathrm{time}(S_{\Pi_e})$ (cf. Definition 9.13). Therefore the following is an immediate consequence of the Pincer Lemma.

$$\tau_1(e) - \tau_2(e) \le T_1(1 + |\chi(\Pi_e)|) + T_0.$$

It now suffices to bound the amount of time for which e is adjacent to the narrow past of $\mathfrak{B}$ before $\tau_2(e)$. We define $\tau_0'(e)$ to be the latest time when the rascal e has contributed less than λ_0 edges to $\mathrm{bonus}(\mathcal{T})$. Then the bonus provided by e is at most $M(\tau_1(e) - \tau_0'(e)) + \lambda_0$. As in the previous section, we define $\mathrm{down}_2(e)$ to be those edges on the left end of corridors containing e at times before $\tau_2(e)$ but after $\tau_0'(e)$. Just as in Lemma 9.29 and the corollaries immediately following it, we then have a notion of *depth* of rascals describing the nesting of the pincers Π_e[16]. We also have *distinguished* rascals (corresponding to the distinguished teams in Lemma 9.29), and proceeding as in the proof of Lemma 9.29 we get the following estimates:

if e_1 is a distinguished rascal of depth $d+1$ and e_0 is the rascal of depth d above it, then the bonus provided by e_1 is at most $2B\Big(T_1(1+|\chi(\Pi_{e_0}))|) + T_0\Big)$, since all of the bonus provided by e_1 must disappear before $\tau_1(e_0)$;

for other rascals e of depth $d+1$ which are below e_0 we have a set of colours $\chi_c(e)$, disjoint for distinct teams such that

$$|\mathrm{down}_2(e) \cap \mathrm{down}_2(e_0)| \le T_1(1 + |\chi_c(e)|) + T_0.$$

Therefore, summing over the set of rascals which are not distinguished we get (cf Corollary 9.30)

$$\sum_e |\mathrm{down}_2(e)| \le 2\Big|\bigcup_e \mathrm{down}_2(e)\Big| + \sum_e \Big(T_1(1 + |\chi_c(e)|) + T_0\Big).$$

And summing over the same set of rascals, we get

$$\sum_e |\mathrm{down}_2(e)| \le (2 + 3T_1 + 2T_0)\,|\partial\Delta|.$$

Therefore, for undistinguished rascals, we have

$$\begin{aligned}\sum_e \tau_1(e) - \tau_0'(e) &= \sum(\tau_1(e) - \tau_2(e)) + \sum(\tau_2(e) - \tau_0'(e)) \\ &\le (3T_1 + 2T_0)\,|\partial\Delta| + (2 + 3T_1 + 2T_0)\,|\partial\Delta|,\end{aligned}$$

[16]One extends the paths p_e and p_e' of Definition 10.4 back in time to $\partial\Delta$ so as to define the order definining depth

and so the contribution of all left-biased rascals is at most

$$\Big((2+6T_1+4T_0)M+2\lambda_0+6BT_1+4BT_0\Big)\,|\partial\Delta|,$$

as required. □

Lemma 10.10. *The sum* $\sum \mathrm{life}(\mathfrak{B})$ *over those* $\mathfrak{B}$ *that are right-biased but do not satisfy conditions (i) or (ii) of Lemma 10.8 is at most* $(3T_1B+2T_0B)\,|\partial\Delta|$.

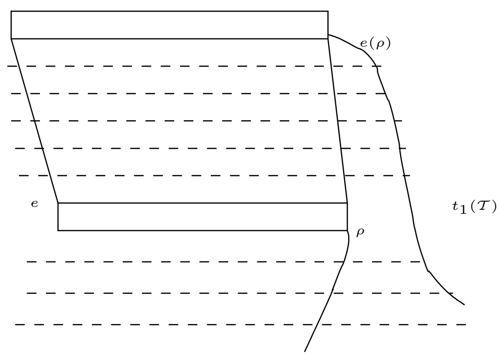

FIGURE 29. A depiction of a right-biased team.

Proof. Once again, as in Lemmas 10.5 and 10.9, we obtain compensation for the continuing existence of a non-constant letter by using the Pincer Lemma to see that colours must be consumed at a constant rate in order to facilitate the life of ρ. Thus we consider the left-fast edge that consumes the pp (i.e. left-most non-constant) future of ρ; this edge is denoted $e(\rho)$ in Figure 29. The Pincer Lemma and the 2 Colour Lemma tell us that if $\Pi_{e(\rho)}$ is the pincer associated to these paths (with S_0 at the bottom) then

$$\mathrm{life}(\mathfrak{B}) \le T_1(1+|\chi(\Pi_{e(\rho)})|)+T_0.$$

Suppose that $\mathfrak{B}$ and $\mathfrak{B}'$ are two right-biased blocks with associated edges $e(\rho)$ and $e(\rho')$ consuming their reapers. We claim that the sets $\chi(\Pi_{e(\rho)})$ and $\chi(\Pi_{e(\rho')})$ are disjoint. The key point to observe is that since we are not in case (ii) of Lemma 10.8 the length of the swollen future of $\mathfrak{B}$ increases from time(S_0) to the top of $\Pi_{e(\rho)}$; since $\mathfrak{B}$ had length at least λ_0, we therefore have a block of more than λ_0 of more than λ_0 constant edges at the top of $\Pi_{e(\rho)}$. Thus the pincers associated to $\mathfrak{B}$ and $\mathfrak{B}'$ are either disjoint or nested. Hence $\chi(\Pi_{e(\rho)})$ and $\chi(\Pi_{e(\rho')})$ are disjoint. Thus summing over all right-biased blocks $\mathfrak{B}$ we obtain

$$\sum_{\mathfrak{B}\text{ right-biased}} \mathrm{life}(\mathfrak{B}) \le (3T_1B+2T_0B)\,|\partial\Delta|,$$

as required. □

Since any letter consumes less than M constant letters in any unit of time, we conclude:

Corollary 10.11. *The sum of the effective volumes of all blocks that are right-biased but do not satisfy conditions (i) and (ii) of Lemma 10.8 is at most* $(3MT_1B + 2MT_0B)\,|\partial\Delta|$.

Lemma 10.12. *The sum of all blocks that satisfy condition (iv) of Lemma 10.8 is at most* $(2B + 1)\,|\partial\Delta|$.

Proof. Possibility (iv) involves several subcases: the key event which halts the growth of the swollen future of $\mathfrak{B}$ may be a collision with $\partial\Delta$ or a singularity; it may also be that the key event is that the future of the rascal or reaper adjacent to $\mathfrak{B}$ is cancelled by an edge that is not in the future of S_0.

But no matter what these key events may be, since we are in not in cases (ii) or (iii), associated to the blocks in case (iv) we have the following set of paths partitioning that part of the diagram Δ bounded by S_0 and the arc of $\partial\Delta$ connecting the termini of the edges at the ends of S_0:

The path π_l begins at time(S_0) and follows the pp-future of the rascal at the right-end of the future of $\mathfrak{B}$ until it hits the boundary, a singularity, or else is cancelled by an edge ε_l not in the future of S_0; if it hits the boundary, it ends; if it hits a singularity, π_l crosses to the bottom of the corridor S on the other side of the singularity, and turns left to follow $\perp(S)$ to the boundary (see Figure 30); if ε_l cancels with the pp-future of the rascal, then π_l follows the past of ε_l backwards in time to the boundary (see Figure 31).

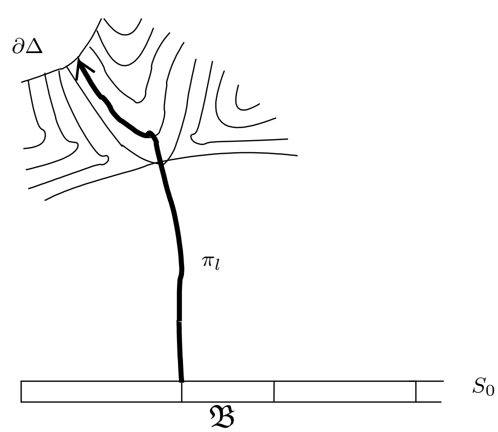

FIGURE 30. The path π_l hits a singularity.

The path π_r describing the fate of ρ is defined similarly (except that it turns right if it hits a singularity).

It is clear from the construction that no two of these paths can cross, thus we have the partition represented schematically in Figure 32.

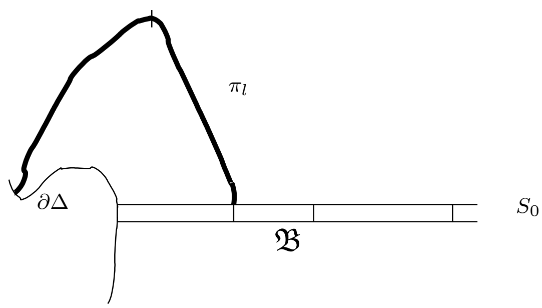

FIGURE 31. The path π_l in cancelled from outside of the future of S_0

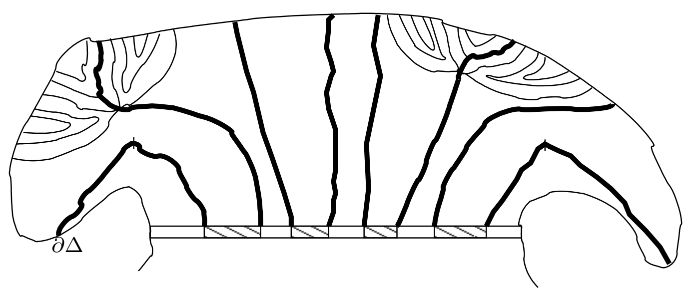

FIGURE 32. The schematic partition of Δ by the paths π_l and π_r.

Given a swollen team $\mathfrak{B}$ of type (iv), we follow the swollen future of $\mathfrak{B}$ until its flow is interrupted (at time $\iota(\mathfrak{B})$, say) by meeting a singularity, the boundary of Δ, or else its rascal or reaper is cancelled. Consider the set of corridors that contain some component of the swollen future of $\mathfrak{B}$ after $\iota(\mathfrak{B})$. Consider also the set of edges $\text{bdy}(\mathfrak{B}) \subseteq \partial\Delta$ that lie in the swollen future of $\mathfrak{B}$. We keep account of the set of corridors by recording the set of their ends on $\partial\Delta$, except that we ignore an end if we have to cross a path π_l or π_r to reach it. Note that at least one end of each corridor is recorded. Let $\text{up}(\mathfrak{B}) \subset \partial\Delta$ denote the set of ends recorded.

Since the sets $\text{bdy}(\mathfrak{B})$ and $\text{up}(\mathfrak{B})$ are contained in the portion of $\partial\Delta$ accorded to $\mathfrak{B}$ by the partition formed by the paths π_l and π_r, the sets associated to different $\mathfrak{B}$ are disjoint. In each unit of time beyond $\iota(\mathfrak{B})$ each component of the swollen future of $\mathfrak{B}$ can shrink by at most $2B$ (by Lemma 2.4). The set $\text{up}(\mathfrak{B})$ measures the sum of the number of components over all such times, and $|\text{bdy}(\mathfrak{B})|$ is the number of uncancelled edges. Thus we see that the length of the swollen future of $\mathfrak{B}$ at time $\iota(\mathfrak{B})$ is at most $2B|\text{up}(\mathfrak{B})|+|\text{bdy}(\mathfrak{B})|$. Finally, the continued presence of the rascal ensures that the swollen future of $\mathfrak{B}$ grows in each interval of time from $\text{time}(S_0)$ to $\iota(\mathfrak{B})$. Thus it follows that the length of $\mathfrak{B}$ is also bounded by this number. So summing over all $\mathfrak{B}$ of type (iv) we have:

$$\sum |\mathfrak{B}| \le \sum \Big(2B|\mathrm{up}(\mathfrak{B})| + |\mathrm{bdy}(\mathfrak{B})| \Big) \le (2B+1)\,|\partial\Delta|,$$

as required. □

Summarising the results of this section we have

Lemma 10.13. *Summing over all teams that are not short, we have*

$$\sum_{\mathcal{T}} |\mathrm{bonus}(\mathcal{T})| \le \Big((B+3)(3T_1+2T_0)M + 6BT_1 + 4BT_0 + 2\lambda_0 + 2B + 5M + 1 \Big)\,|\partial\Delta|.$$

11. The proof of the Main Theorem

Pulling all of the previous results together, define

$$K_1 = 2C_1+6\lambda_0+2B(5T_0+6T_1+2)+2MC_4(6T_1+8T_0+3)+(B+3)(3T_1+2T_0)M+5M+2,$$

and

$$K = 2C_0 + 2K_1 + 2B + 1.$$

Theorem 11.1. $|S_0| \le K\,|\partial\Delta|$.

Proof. The corridor S_0 can be subdivided into distinct colours which form connected regions. Each colour μ can be partitioned into connected (possibly empty) regions $A_1(S_0,\mu)$, $A_2(S_0,\mu)$, $A_3(S_0,\mu)$, $A_4(S_0,\mu)$ and $A_5(S_0,\mu)$. By Lemma 6.4, Proposition 7.1, Lemma 6.3, Proposition 7.3 and Lemma 6.4, respectively,

$$\begin{aligned}
\sum_{\mu\in S_0} |A_1(S_0,\mu)| &\le C_0\,|\partial\Delta|,\\
\sum_{\mu\in S_0} |A_2(S_0,\mu)| &\le K_1\,|\partial\Delta|,\\
\sum_{\mu\in S_0} |A_3(S_0,\mu)| &\le (2B+1)\,|\partial\Delta|,\\
\sum_{\mu\in S_0} |A_4(S_0,\mu)| &\le K_1\,|\partial\Delta|, \text{ and}\\
\sum_{\mu\in S_0} |A_5(S_0,\mu)| &\le C_0\,|\partial\Delta|.
\end{aligned}$$

Summing completes the proof of Theorem 11.1. □

Since there are at most $\frac{|\partial\Delta|}{2}$ corridors in Δ,

$$\mathrm{Area}(\Delta) \le \frac{K}{2}\,|\partial\Delta|^2,$$

which proves the Main Theorem.

12. Glossary of Constants

B – the Bounded Cancellation constant (Lemmas 2.4 and 2.3).

C_0 – maximum distance a left-fast (right-fast) letter can be from the left (right) edge of its colour if it is to be cancelled from the left (right) within the future of the corridor. See Lemma 6.4.

C_1 – an upper bound on the lengths of the subintervals $C_{(\mu,\mu')}(1)$ of $A_4(S_0,\mu)$. By definition, $C_{(\mu,\mu')}(1)$ is consumed by $\mu'(S_0)$; it begins at the right end of $A_4(S_0,\mu)$ and ends at the last non-constant letter. See Lemma 6.7. Note that one can take $C_1 = 2mB^2$.

M – the maximum of the lengths of the images $\phi(a_i)$ of the basis elements a_i, i.e. the maximum length of $u_1,\dots,u_m$ in the presentation $\mathcal{P}$ (see equation 1.1).

M_{inv} – the maximum of the lengths of $\phi^{-1}(a_i)$.

T_0 – the constant from the 2-Colour Lemma (Lemma 8.4). For all positive words U and V, if U neuters V^{-1} then it does so in at most T_0 steps.

$\hat{T}_1$ – the constant from the Unnested Pincer Lemma, Theorem 8.7.

T_1' – the constant from Definition 8.19. Recall that we stipulate that $T_1' \geq \hat{T}_1$.

$T_1 := T_1' + 2T_0 - T_1$ is the constant from the Pincer Lemma, Theorem 8.26.

$C_4 := MM_{inv}$

$\lambda_0 := \max\{2B(T_0+1)+1, MC_4\}$

Finally, K_1 is defined to be

$$2C_1+6\lambda_0+2B(5T_0+6T_1+2)+2MC_4(6T_1+8T_0+3)+(B+3)(3T_1+2T_0)M+5M+2,$$

and $K = 2C_0 + 2K_1 + 2B + 1$.

References

[1] M. Bestvina, The topology of $\mathrm{Out}(F_n)$, in *Proceedings of ICM, Bejing 2002, Vol.II*, Higher Education Press, Bejing, 2002. pp. 373–384.

[2] M. Bestvina and M. Feighn, A combination theorem for negatively curved groups, *J. Diff. Geom.*, **35** (1992), 85–101.

[3] M. Bestvina, M. Feighn and M. Handel, The Tits alternative for $\mathrm{Out}(F_n)$ I: Dynamics of exponentially growing automorphisms, *Ann. of Math. (2)*, **151** (2000), 517–623.

[4] M. Betvina, M. Feighn and M. Handel, The Tits alternative for $\mathrm{Out}(F_n)$ II: A Kolchin type theorem, preprint.

[5] M. Bestvina and M. Handel, Train tracks and automorphisms of free groups, *Ann. of Math. (2)*, **135**, 1–51.

[6] N. Brady and M.R. Bridson, On the absence of biautomaticity for graphs of abelian groups, preprint.

[7] M.R. Bridson, Polynomial Dehn functions and the length of asynchronously automatic structures, *Proc. London Math. Soc.(3)*, **85** (2002), 441–466.

[8] M.R. Bridson, On the subgroups of semihyperbolic groups, *Monog. L'Enseign. Math.*, **38** (2001), 85–111.

[9] M.R. Bridson, The geometry of the word problem, in *"Invitations to geometry and topology"* (M.R. Bridson and S.M. Salamon, eds.), Oxford University Press, 2002.

[10] M.R. Bridson and S.M. Gersten, The optimal isoperimetric inequality for torus bundles over the circle, *Quart. J. Math. Oxford Ser. (2)*, **47** (1996), 1–23.
[11] M. R. Bridson and D. P. Groves, The quadratic isoperimetric inequality for mapping tori of free group automorphisms II: The general case, preprint.
[12] M. R. Bridson and D.P. Groves, Alogrithms and constants for isoperimetric functions of free-by-cyclic groups, in preparation.
[13] M.R. Bridson and A. Haefliger, *Metric spaces of non-positive curvature*, Springer-Verlag, Berlin, 1999.
[14] M.R. Bridson and L. Reeves, On the absence of automaticity in certain free-by-cyclic groups, in preparation.
[15] P. Brinkmann, Hyperbolic automorphisms of free groups, *GAFA*, **10** (2000), 1071–1089.
[16] D. Cooper, Automorphisms of free groups have finitely generated fixed point sets, *J. Algebra*, **111** (1987), 453–456.
[17] D.P.A. Epstein, J.W. Cannon, D.F. Holt, S.V.F. Levy, M.S. Paterson and W.P. Thurston, *Word processing in groups*, Jones and Bartlett, Boston, 1992.
[18] M. Feighn and M. Handel, Mapping tori of free group automorphisms are coherent, *Ann. of Math. (2)*, **149** (1999), 1061–1077.
[19] S.M. Gersten, The automorphism group of a free group is not a CAT(0) group, *Proc. Amer. Math. Soc.*, **121** (1994), 999–1002.
[20] M. Gromov, Hyperbolic groups, in *Essays in group theory* (S.M. Gersten, ed.), Springer Verlag, MSRI Publ. **8** (1987), 75-263.
[21] E.R. van Kampen, On some lemmas in the theory of groups, *Amer. J. Math.*, **55** (1933), 268–273.
[22] B. Leeb, 3-Manifolds with(out) metrics of nonpositive curvature, *Invent. Math.* **122** (1995), 277–289.
[23] M. Lustig, Structure and conjugacy for automorphisms of free groups, preprint.
[24] R.C. Lyndon and P.E. Schupp, Combinatorial group theory, Springer-Verlag, Berlin, 1977.
[25] N. Macura, Quadratic isoperimetric inequality for mapping tori of polynomially growing automorphisms of free groups, *GAFA*, **10** (2000), 874–901.
[26] Z. Sela, The Nielsen-Thurston classification and automorphisms of a free group I, *Duke Math. J.*, **84** (1996), 379–397.

Martin R. Bridson, Mathematics Department, Imperial College of Science, Technology and Medicine, 180 Queen's Gate, London SW7 2BZ, UK
E-mail address: `m.bridson@ic.ac.uk`

Daniel Groves, Merton College, Oxford OX1 4JD, UK
E-mail address: `grovesd@maths.ox.ac.uk`